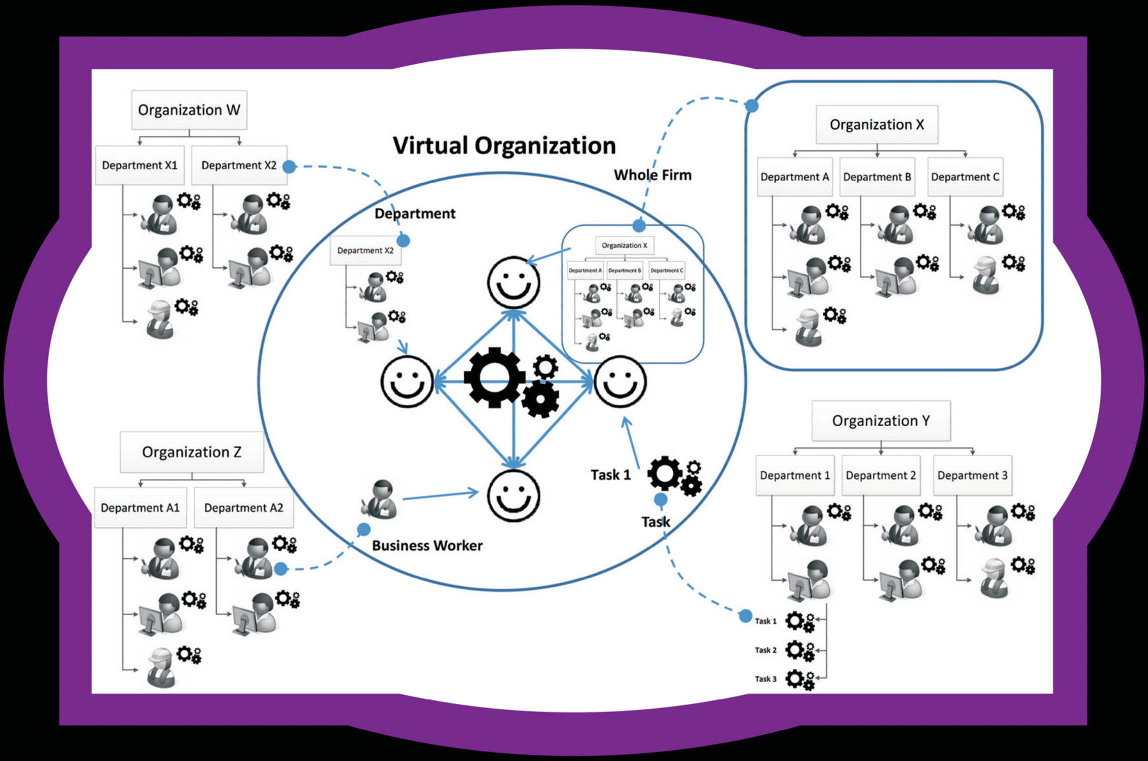
Computational Models for Business and Engineering Domains



ITHEA®

2014

Galina Setlak, Krassimir Markov
(editors)

COMPUTATIONAL MODELS FOR BUSINESS AND ENGINEERING DOMAINS

I T H E A®

Rzeszow - Sofia

2014

Galina Setlak, Krassimir Markov (ed.)

Computational Models for Business and Engineering Domains

ITHEA®

2014, Rzeszow, Poland; Sofia, Bulgaria,

ISBN: 978-954-16-0066-5 (printed)
ISBN: 978-954-16-0067-2 (online)

ITHEA IBS ISC No.: 30

First edition

Printed in Poland

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

This issue contains a monograph that concern actual problems of research and application of computational models for business and engineering domains, especially the new approaches, models, algorithms and methods for computational 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.

ISBN: 978-954-16-0066-5 (printed)
ISBN: 978-954-16-0067-2 (online)

C\o Jusautor, Sofia, 2014

PREFACE

In general, "Computational modeling" uses computer science methods, techniques and tools to study the behavior of different types and categories of artificial as we as natural systems – socio-technical (e.g. business and engineering), biological, physical and chemical systems. A computational model is a computational representation of the specific object, process or phenomenon finally developed in a form of computer program. It means that the model can be run on specific hardware/software architecture and advanced analysis of the structure or behavior of such artificial system may be conducted. Nowadays, thanks to advances in computer science, computational modeling is used in many application domains. This monograph is focused on computational models related to business and engineering systems where there is a need to understand how the complex system will behave under specific conditions. In such cases intuitive analytical solutions are not always available (sometimes even possible) or do not provide a solutions in a reasonable time. The results of a computational model analysis can help researchers to make predictions about what will happen in the real systems that are being studied in response to changing conditions. What is more, operation theories can be derived/deduced and verified on the basis of computational experiments. Rather than deriving a mathematical analytical solution to the problem, experimentation with the model is done by adjusting the parameters of the system in the computer program (computational representation of the model), and studying the differences in the outcome of the experiments. A computational model may contain numerous variables that characterize the system under study and computational analysis is done by adjusting these variables and observing how the changes affect the outcomes predicted by the model. Modeling can expedite research by allowing scientists to conduct thousands of experiments at a relatively low costs.

This issue of a monograph concerns the most recent problems solutions and new approaches in the form of models, algorithms, techniques and methods for computational modeling and analysis used in applications of intelligent and information systems to business and engineering domains. The topics weconsider as most important, which have been included in this issue are:

- Automatic control systems models
- Computational intelligence models
- Knowledge discovery and data mining models
- Natural language processing models
- Agent-oriented software engineering models
- Computational models and simulation
- Business intelligence models

We hope that this monograph constitutes a valuable source of knowledge for experts in the field of modern ICT solutions as well as for practical users.

We would like to express our special thanks to all authors of this monograph as well as to all who supported its publication.

Rzeszow – Sofia *G. Setlak, Kr. Markov*
September 2014

TABLE OF CONTENTS

Preface 3

Table of Contents 5

Index of Authors 7

Automatic Control Systems Models

Tadeusz Kaczorek

Polynomial approach to fractional descriptor electrical circuits 8

Computational Intelligence Models

Arkady Yuschenko

Fuzzy Techniques in Robotic Systems Control 22

Yevgeniy Bodyanskiy, Alina Shafronenko

Robust Adaptive Fuzzy Clustering for Data with Missing Values 34

Yevgeniy Bodyanskiy, Oleksii Tyshchenko, Daria Kopaliani

The Least Squares Support Vector Machine Based on a Neo-Fuzzy Neuron 44

Yevgeniy Bodyanskiy, Olena Vynokurova, Iryna Pliss, Peleshko Dmytro

Multilayer Neuro-Fuzzy System for Solving On-Line Diagnostics Tasks 52

Knowledge Discovery and Data Mining Models

Sergey Maruev, Eugene Levner, Dmitry Stefanovskyi, Alexander Troussov

Modeling Educational Processes in Modern Society by Navigating Multidimensional Networks 60

Sergey Maruev, Dmitry Stefanovskyi, Alexander Troussov, John Curry, Alexey Frolov

Multidimensional Networks for Heterogeneous Data Modeling 67

Olga Proncheva, Mikhail Alexandrov, Volodymyr Stepashko, Oleksiy Koshulko

Forecast of Forrester's Variables Using GMDH Technique 78

Anatoli Nachev

Application of Data Mining Techniques for Direct Marketing 86

Sergiy Chalyi, Olga Kalynychenko, Sergiy Shabanov-Kushnarenko, Vira Golyan

Discriminative Approach to Discovery Implicit Knowledge 96

Galina Setlak, Monika Piróg-Mazur, Łukasz Paśko

Intelligent Analysis of Manufacturing Data 109

Dmytro Terletskyi, Alexandr Provotar

Object-Oriented Dynamic Network 123

Natural Language Processing Models

Vera Danilova

Ontology building and Annotation of Destabilizing Events in News Feeds 137

Alexey Dobrov

Semantic and Ontological Relations in AIIRE Natural Language Processor 147

Svetlana Koshcheeva, Victor Zakharov
Comparing Methods of Automatic Verb-Noun Collocation Extraction 158

Krassimira Ivanova
Practical Aspects of Natural Language Addressing 172

Agent-Oriented Software Engineering Models

Jacek Jakieła, Paweł Litwin, Marcin Olech
Reducing Semantic Gap in Development Process of Management Information Systems for Virtual Organizations 187

Computational Simulation Models

Alexander Temruk, Mikhail Alexandrov
Tools for Analysis of Processes Measured on Sparse and Irregular Spatial-Temporal Grid with Application to Data of National Censuses 205

Alina Nasibullina, Mikhail Alexandrov, Alexander Kovaldji
Simple Free-Share Package for Visual Analysis of Multidimensional Data Sets 216

Roman Bazylevych, Marek Pałasiński, Roman Kutelmakh
Efficient Decomposition Algorithms for Solving Large Scale TSP 225

Sumeer Chakuu, Michał Nędza
Queuing Based Simulation Models for analyzing Runway Capacity and Managing SLOTS at the Airports 235

Sumeer Chakuu, Michał Nędza
Evaluation Of Runway Capacity and Slots at London Gatwick Airport Using Queuing Based Simulation 246

Damian Krzesimowski
The Fast Fourier Transform and Cepstrogram based Approach to the Assessment of Human Voice Stability 258

Business Intelligence Models

V. Stepashko, O. Samoilenko, R. Voloschuk
Informational Support of Managerial Decisions as a New Kind of Business Intelligent Systems 269

Vladimir Averkiev, Mikhail Alexandrov, Javier Tejada
Statistical Models for the Support of Overbooking in Transport Service 280

Justyna Stasieńko
How to Master Big data 287

INDEX OF AUTHORS

name		***pages***
Alexandrov	Mikhail	78, 205, 216, 280
Averkiev	Vladimir	280
Bazylevych	Roman	225
Bodyanskiy	Yevgeniy	34, 44, 52
Chakuu	Sumeer	235, 246
Curry	John	67
Danilova	Vera	137
Dmytro	Peleshko	52
Dobrov	Alexey	147
Frolov	Alexey	67
Golyan	Vira	96
Ivanova	Krassimira	172
Jakieła	Jacek	187
Kaczorek	Tadeusz	8
Kalynychenko	Olga	96
Kopaliani	Daria	44
Koshcheeva	Svetlana	158
Koshulko	Oleksiy	78
Kovaldji	Alexander	216
Krzesimowski	Damian	258
Kutelmakh	Roman	225
Levner	Eugene	60
Litwin	Paweł	187
Maruev	Sergey	60, 67
Nachev	Anatoli	86
Nasibullina	Alina	216
Nędza	Michał	235, 246
Olech	Marcin	187
Pałasiński	Marek	225
Paśko	Łukasz	109
Piróg-Mazur	Monika	109
Pliss	Iryna	52
Proncheva	Olga	78
Provotar	Alexandr	123
Samoilenko	Oleksandr	269
Setlak	Galina	109
Shabanov-Kushnarenko	Sergiy	96
Shafronenko	Alina	34
Stasieńko	Justyna	287
Stefanovskyi	Dmitry	60,67
Stepashko	Volodymyr	78, 269
Tejada	Javier	280
Temruk	Alexander	205
Terletskyi	Dmytro	123
Troussov	Alexander	60, 67
Tyshchenko	Oleksii	44
Voloschuk	Roman	269
Vynokurova	Olena	52
Yuschenko	Arkady	22
Zakharov	Victor	158

Automatic Control Systems Models

POLYNOMIAL APPROACH TO FRACTIONAL DESCRIPTOR ELECTRICAL CIRCUITS

Tadeusz Kaczorek

Abstract: *A new polynomial approach is proposed to analysis of the standard and positive descriptor electrical circuits composed of resistors, coils, capacitors and voltage (current) sources. It is shown that for given descriptor fractional electrical circuit the equivalent standard fractional electrical circuit can be found by premultiplication of the equation of the descriptor electrical circuit by suitable polynomial matrix of elementary row operations. The main result is demonstrated on simple positive fractional electrical circuit.*

Keywords: *Polynomial approach, descriptor, fractional, electrical circuits.*

ACM Classification Keywords: *I.2.8 Computing Methodologies – Control theory*

Introduction

Descriptor (singular) linear systems have been considered in many papers and books [Bru et. all, 2003a, 2003b; Campbell et. all, 1976; Dai, 1989; Guang-Ren, 2010; Kaczorek, 2004, 1992, 2011a, 2011e, 2011f, 2011g, 2013a, 2014a, 2014b; Virnik, 2008]. The eigenvalues and invariants assignment by state and output feedbacks have been investigated in [Kaczorek, 2004] and the minimum energy control of descriptor linear systems in [Kaczorek, 1992]. In positive systems inputs, state variables and outputs take only non-negative values [Farina et. all, 2000; Kaczorek, 2002]. Examples of positive systems are industrial processes involving chemical reactors, heat exchangers and distillation columns, storage systems, compartmental systems, water and atmospheric pollution models. A variety of models having positive linear behavior can be found in engineering, management science, economics, social sciences, biology and medicine, etc. The positive fractional linear systems and some of selected problems in theory of fractional systems have been addressed in monograph [Kaczorek, 2011f].

Descriptor standard positive linear systems by the use of Drazin inverse has been addressed in Bru et. all, 2003a, 2003b; Campbell et. all, 1976; Kaczorek, 2013a]. The shuffle algorithm has been applied to checking the positivity of descriptor linear systems in [Kaczorek, 2011a]. The stability of positive descriptor systems has been investigated in [Virnik, 2008]. Reduction and decomposition of descriptor fractional discrete-time linear systems have been considered in [Kaczorek, 2011e]. Standard and fractional systems and electrical linear circuits have been investigated in [Kaczorek, 2002, 2008, 2010, 2011c, 2011f]. Pointwise completeness and pointwise generacy of standard and positive 1D and 2D systems have been addressed in [Kaczorek, 2009, 2011b].

In this paper a new polynomial approach to analysis of fractional descriptor electrical circuit will be proposed.

The paper is organized as follows. In section 2 basic definitions and theorems concerning the descriptor fractional and positive electrical circuits are recalled. The main result is presented in section 3, where a procedure for reduction of the descriptor fractional electrical circuits to the standard fractional electrical circuits is proposed. Concluding remarks are given in section 4.

The following notation will be used: $\Re$ - the set of real numbers, $\Re^{n\times m}$ - the set of $n\times m$ real matrices and $\Re^n = \Re^{n\times 1}$, $\Re_+^{n\times m}$ - the set of $n\times m$ matrices with nonnegative entries and $\Re_+^n = \Re_+^{n\times 1}$, M_n- the set of $n\times n$ Metzler matrices (real matrices with nonnegative off-diagonal entries), I_n- the $n\times n$ identity matrix.

Preliminaries

The following Caputo definition of the fractional derivative will be used [Kaczorek, 2011f]

$$D_t^\alpha f(t) = \frac{d^\alpha}{dt^\alpha} f(t) = \frac{1}{\Gamma(p-\alpha)} \int_0^t \frac{f^{(p)}(\tau)}{(t-\tau)^{\alpha+1-p}} d\tau,\ p-1 \le \alpha < p \in N = \{1,2,...\}, \tag{1}$$

where $\alpha \in \Re$ is the order of fractional derivative and $f^{(p)}(\tau) = \frac{d^p f(\tau)}{d\tau^p}$ and $\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt$ is the gamma function.

Consider the continuous-time fractional linear system described by the state equations

$$D_t^\alpha x(t) = Ax(t) + Bu(t),\ \ 0 < \alpha < 1, \tag{2a}$$

$$y(t) = Cx(t) + Du(t), \tag{2b}$$

where $x(t) \in \Re^n$, $u(t) \in \Re^m$, $y(t) \in \Re^p$ are the state, input and output vectors and $A \in \Re^{n\times n}$, $B \in \Re^{n\times m}$, $C \in \Re^{p\times n}$, $D \in \Re^{p\times m}$.

Theorem 1. [Kaczorek, 2011f] The solution of equation (2a) is given by

$$x(t) = \Phi_0(t)x_0 + \int_0^t \Phi(t-\tau)Bu(\tau)d\tau,\quad x(0) = x_0\ , \tag{3}$$

where

$$\Phi_0(t) = E_\alpha(At^\alpha) = \sum_{k=0}^\infty \frac{A^k t^{k\alpha}}{\Gamma(k\alpha+1)}, \tag{4}$$

$$\Phi(t) = \sum_{k=0}^\infty \frac{A^k t^{(k+1)\alpha-1}}{\Gamma[(k+1)\alpha]} \tag{5}$$

and $E_\alpha(At^\alpha)$ is the Mittag-Leffler matrix function.

Definition 1. [Kaczorek, 2011f] The fractional system (2) is called the internally positive fractional system if and only if $x(t) \in \Re_+^n$ and $y(t) \in \Re_+^p$ for $t \ge 0$ for any initial conditions $x_0 \in \Re_+^n$ and all inputs $u(t) \in \Re_+^m,\ t \ge 0.$

Theorem 2. [Kaczorek, 2011f] The continuous-time fractional system (2) is internally positive if and only if the matrix A is a Metzler matrix and

$$A \in M_n,\ \ B \in \Re_+^{n\times m},\ \ C \in \Re_+^{p\times n},\ \ D \in \Re_+^{p\times m}. \tag{6}$$

Let the current $i_C(t)$ in a supercondensator (shortly condensator) with the capacity C be the α order derivative of its charge $q(t)$ [Kaczorek, 2011f]

$$i_C(t) = \frac{d^{\alpha}q(t)}{dt^{\alpha}}, \ 0 < \alpha < 1. \tag{7}$$

Using $q(t) = Cu_C(t)$ we obtain

$$i_C(t) = C\frac{d^{\alpha}u_C(t)}{dt^{\alpha}} \tag{8}$$

where $u_C(t)$ is the voltage on the condensator.

Similarly, let the voltage $u_L(t)$ on coil (inductor) with the inductance L be the β order derivative of its magnetic flux $\Psi(t)$ [Kaczorek, 2011f]

$$u_L(t) = \frac{d^{\beta}\Psi(t)}{dt^{\beta}}, \quad 0 < \beta < 1. \tag{9}$$

Taking into account that $\Psi(t) = Li_L(t)$ we obtain

$$u_L(t) = L\frac{d^{\beta}i_L(t)}{dt^{\beta}}, \tag{10}$$

where $i_L(t)$ is the current in the coil.

Consider an electrical circuit composed of resistors, n capacitors and m voltage sources. Using the equation (2.8) and the Kirchhoff's laws we may describe the transient states in the electrical circuit by the fractional differential equation

$$\frac{d^{\alpha}x(t)}{dt^{\alpha}} = Ax(t) + Bu(t), \quad 0 < \alpha < 1, \tag{11}$$

where $x(t) \in \Re^n$, $u(t) \in \Re^m$, $A \in \Re^{nxn}$, $B \in \Re^{n \times m}$. The components of the state vector $x(t)$ and input vector $u(t)$ are the voltages on the condensators and source voltages respectively. Similarly, using the equation (10) and the Kirchhoff's laws we may describe the transient states in the electrical circuit by the fractional differential equation

$$\frac{d^{\beta}x(t)}{dt^{\beta}} = Ax(t) + Bu(t), \quad 0 < \beta < 1, \tag{12}$$

where $x(t) \in \Re^n$, $u(t) \in \Re^m$, $A \in \Re^{n \times n}$, $B \in \Re^{n \times m}$. In this case the components of the state vector $x(t)$ are the currents in the coils.

Solution of the equation (11) (or (2.12)) satisfying the initial condition $x(0) = x_0$ is given by (3).

Now let us consider electrical circuit composed of resistors, capacitors, coils and voltage (current) source. As the state variables (the components of the state vector $x(t)$) we choose the voltages on the capacitors and the currents in the coils. Using the equations (8), (10) and Kirchhoff's laws we may write for the fractional linear circuits in the transient states the state equation

$$\begin{bmatrix} \frac{d^{\alpha}x_C}{dt^{\alpha}} \\ \frac{d^{\beta}x_L}{dt^{\beta}} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}\begin{bmatrix} x_C \\ x_L \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}u, \quad 0 < \alpha, \ \beta < 1, \tag{13a}$$

where the components $x_C \in \Re^{n_1}$ are voltages on the condensators, the components $x_L \in \Re^{n_2}$ are currents in the coils and the components of $u \in \Re^m$ are the source voltages and

$$A_{ij} \in \Re^{n_i \times n_j}, \; B_i \in \Re^{n_i \times m}, \quad i, j = 1,2. \tag{13b}$$

Theorem 3. The solution of the equation (13) for $0 < \alpha < 1; \; 0 < \beta < 1$ with initial conditions

$$x_C(0) = x_{10} \text{ and } x_L(0) = x_{20} \tag{14}$$

has the form

$$x(t) = \Phi_0(t)x_0 + \int_0^t [\Phi_1(t-\tau)B_{10} + \Phi_2(t-\tau)B_{01}]u(\tau)d\tau, \tag{15a}$$

where

$$x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}, \; x_0 = \begin{bmatrix} x_{10} \\ x_{20} \end{bmatrix}, \; B_{10} = \begin{bmatrix} B_1 \\ 0 \end{bmatrix}, \; B_{01} = \begin{bmatrix} 0 \\ B_2 \end{bmatrix},$$

$$T_{kl} = \begin{cases} I_n \text{ for } k = l = 0 \\ \begin{bmatrix} A_{11} & A_{12} \\ 0 & 0 \end{bmatrix} \text{ for } k = 1, l = 0 \\ \begin{bmatrix} 0 & 0 \\ A_{21} & A_{22} \end{bmatrix} \text{ for } k = 0, l = 1 \\ T_{10}T_{k-1,l} + T_{01}T_{k,l-1} \text{ for } k + l > 0 \end{cases} \tag{15b}$$

$$\Phi_0(t) = \sum_{k=0}^{\infty}\sum_{l=0}^{\infty} T_{kl} \frac{t^{k\alpha + l\beta}}{\Gamma(k\alpha + l\beta + 1)},$$

$$\Phi_1(t) = \sum_{k=0}^{\infty}\sum_{l=0}^{\infty} T_{kl} \frac{t^{(k+1)\alpha + l\beta - 1}}{\Gamma[(k+1)\alpha + l\beta]},$$

$$\Phi_2(t) = \sum_{k=0}^{\infty}\sum_{l=0}^{\infty} T_{kl} \frac{t^{k\alpha + (l+1)\beta - 1}}{\Gamma[k\alpha + (l+1)\beta]}.$$

Proof is given in [Kaczorek, 2010, 2011f].

The extension of Theorem 3 to systems consisting of *n* subsystems with different fractional orders is given in [Kaczorek, 2011d].

Reduction of descriptor linear electrical circuits to their standard equivalent forms

The following elementary row (column) operations will be used [Kaczorek, 1992]:

Multiplication of the *i*th row (column) by a real number c. This operation will be denoted by $L[i \times c]$ ($R[i \times c]$).

Addition to the *i*th row (column) of the *j*th row (column) multiplied by a real number c. This operation will be denoted by $L[i + j \times c]$ ($R[i + j \times c]$).

Interchange of the *i*th and *j*th rows (columns). This operation will be denoted by $L[i, j]$ ($R[i, j]$).

First the essence of the polynomial approach will be shown on the following simple example.

Example 1. Consider the fractional descriptor electrical circuit shown in Fig. 1 with given resistances R_1, R_2; inductances L_1, L_2 and source current i_z.

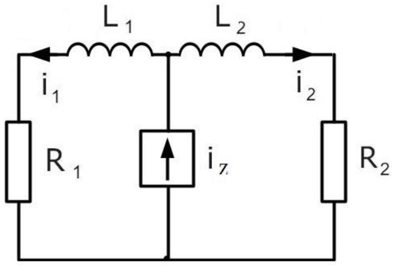


Fig. 1. Fractional electrical circuit

Using Kirchhoff's laws we can write the equations

$$L_1 \frac{d^\alpha i_1}{dt^\alpha} + R_1 i_1 = L_2 \frac{d^\alpha i_2}{dt^\alpha} + R_2 i_2 \tag{16a}$$

$$i_z = i_1 + i_2 \tag{16b}$$

The equations (16) can be written in the form

$$E \frac{d^\alpha}{dt^\alpha} \begin{bmatrix} i_1 \\ i_2 \end{bmatrix} = A \begin{bmatrix} i_1 \\ i_2 \end{bmatrix} + Bi_z \tag{17a}$$

where

$$E = \begin{bmatrix} L_1 & -L_2 \\ 0 & 0 \end{bmatrix} \; A = \begin{bmatrix} -R_1 & R_2 \\ -1 & -1 \end{bmatrix}, \; B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{17b}$$

Defining

$$E_1 = [L_1 \quad -L_2] \;\; A_1 = [-R_1 \quad R_2], \; A_2 = [-1 \quad -1], B_1 = [0], \; B_2 = [1] \tag{18}$$

we can write the equation (17) in the form

$$E_1 \frac{d^\alpha}{dt^\alpha} \begin{bmatrix} i_1 \\ i_2 \end{bmatrix} = A_1 \begin{bmatrix} i_1 \\ i_2 \end{bmatrix} + B_1 i_z \tag{19a}$$

and

$$0 = A_2 \begin{bmatrix} i_1 \\ i_2 \end{bmatrix} + B_2 i_z . \tag{19b}$$

The fractional differentiation of (19b) yields

$$0 = A_2 \frac{d^\alpha}{dt^\alpha} \begin{bmatrix} i_1 \\ i_2 \end{bmatrix} + B_2 \frac{d^\alpha i_z}{dt^\alpha} . \tag{20}$$

From (19a) and (20) we have

$$\begin{bmatrix} E_1 \\ -A_2 \end{bmatrix} \frac{d^\alpha}{dt^\alpha} \begin{bmatrix} i_1 \\ i_2 \end{bmatrix} = \begin{bmatrix} A_1 \\ 0 \end{bmatrix} \begin{bmatrix} i_1 \\ i_2 \end{bmatrix} + \begin{bmatrix} B_1 \\ 0 \end{bmatrix} i_z + \begin{bmatrix} 0 \\ B_2 \end{bmatrix} \frac{d^\alpha i_z}{dt^\alpha} . \tag{21}$$

Note that the matrix

$$\begin{bmatrix} E_1 \\ -A_2 \end{bmatrix} = \begin{bmatrix} L_1 & -L_2 \\ 1 & 1 \end{bmatrix} \tag{22}$$

is nonsingular and premultiplying (21) by its inverse we obtain

$$\frac{d^\alpha}{dt^\alpha}\begin{bmatrix} i_1 \\ i_2 \end{bmatrix} = \overline{A}\begin{bmatrix} i_1 \\ i_2 \end{bmatrix} + \overline{B}_0 i_z + \overline{B}_1 \frac{d^\alpha i_z}{dt^\alpha} \tag{23a}$$

where

$$\begin{aligned}
\overline{A} &= \begin{bmatrix} E_1 \\ -A_2 \end{bmatrix}^{-1}\begin{bmatrix} A_1 \\ 0 \end{bmatrix} = \begin{bmatrix} L_1 & -L_2 \\ 1 & 1 \end{bmatrix}^{-1}\begin{bmatrix} -R_1 & R_2 \\ 0 & 0 \end{bmatrix} = \frac{1}{L_1+L_2}\begin{bmatrix} -R_1 & R_2 \\ R_1 & -R_2 \end{bmatrix}, \\
\overline{B}_0 &= \begin{bmatrix} E_1 \\ -A_2 \end{bmatrix}^{-1}\begin{bmatrix} B_1 \\ 0 \end{bmatrix} = \begin{bmatrix} L_1 & -L_2 \\ 1 & 1 \end{bmatrix}^{-1}\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \\
\overline{B}_1 &= \begin{bmatrix} E_1 \\ -A_2 \end{bmatrix}^{-1}\begin{bmatrix} 0 \\ B_2 \end{bmatrix} = \begin{bmatrix} L_1 & -L_2 \\ 1 & 1 \end{bmatrix}^{-1}\begin{bmatrix} 0 \\ 1 \end{bmatrix} = \frac{1}{L_1+L_2}\begin{bmatrix} L_2 \\ L_1 \end{bmatrix}.
\end{aligned} \tag{23b}$$

Note that the electrical circuit with (23) is positive since $\overline{A} \in M_2$ and the matrices $\overline{B}_0$ and $\overline{B}_1$ have nonnegative entries.

The standard equation (23a) can be also obtained from the equation (21) by reducing the matrix (22) to the identity matrix I_2 using the elementary row operations

$$L[1+2\times L_2],\ L\left[1\times\frac{1}{L_1+L_2}\right],\ L[2+1\times(-1)]. \tag{24}$$

Performing the elementary row operations (324) on the matrix $\begin{bmatrix} 1 & 0 \\ 0 & s^\alpha \end{bmatrix}$ we obtain the polynomial matrix

$$L(s^\alpha) = \frac{1}{L_1+L_2}\begin{bmatrix} 1 & s^\alpha L_2 \\ -1 & s^\alpha L_1 \end{bmatrix} \tag{25}$$

satisfying the equalities

$$\begin{aligned}
L(s^\alpha)[Es^\alpha - A] &= \frac{1}{L_1+L_2}\begin{bmatrix} 1 & s^\alpha L_2 \\ -1 & s^\alpha L_1 \end{bmatrix}\begin{bmatrix} s^\alpha L_1 + R_1 & -s^\alpha L_2 - R_2 \\ 1 & 1 \end{bmatrix} \\
&= [I_n s^\alpha - \overline{A}] = \begin{bmatrix} s^\alpha + \dfrac{R_1}{L_1+L_2} & -\dfrac{R_2}{L_1+L_2} \\ -\dfrac{R_1}{L_1+L_2} & s^\alpha + \dfrac{R_2}{L_1+L_2} \end{bmatrix}, \\
L(s^\alpha)B &= \frac{1}{L_1+L_2}\begin{bmatrix} 1 & s^\alpha L_2 \\ -1 & s^\alpha L_1 \end{bmatrix}\begin{bmatrix} 0 \\ 1 \end{bmatrix} = [\overline{B}_0 + \overline{B}_1 s^\alpha] = \frac{s^\alpha}{L_1+L_2}\begin{bmatrix} L_2 \\ L_1 \end{bmatrix}.
\end{aligned} \tag{26}$$

Therefore, the reduction of the matrix (22) to identity matrix by the use of elementary row operations (24) is equivalent to premultiplication of the equation

$$[Es^\alpha - A]X = BU \tag{27}$$

by the polynomial matrix of elementary row operations (25), where $X = \mathcal{L}\begin{bmatrix} i_1 \\ i_2 \end{bmatrix}$, $U = \mathcal{L}[i_z]$ and $\mathcal{L}$ is the Laplace operator.

In general case let consider the descriptor electrical circuit described by the equation

$$E\frac{d^\alpha x}{dt^\alpha} = Ax + Bu \tag{28}$$

where $x(t) \in \Re^n$, $u(t) \in \Re^m$ are the state and input vectors and $E, A \in \Re^{n\times n}$, $B \in \Re^{n\times m}$. It is assumed that $\det E = 0$ and the pencil (E,A) is regular.

Applying to (28) the Laplace transform with zero initial conditions we obtain the equation

$$[Es^\alpha - A]X = BU \tag{29}$$

where $X = \mathcal{L}[x(t)]$, $U = \mathcal{L}[u(t)]$.

Theorem 4. There exists a nonsingular polynomial matrix

$$L(s^\alpha) = L_0 + L_1 s^\alpha + ... + L_\mu s^{\alpha\mu} \tag{30}$$

where μ is the nilpotent index of the pair (E,A), such that

$$L(s^\alpha)[Es^\alpha - A] = [I_n s^\alpha - \overline{A}] \tag{31}$$

if and only if the pencil (E,A) is regular, i.e.

$$\det[Es^\alpha - A] \neq 0 \tag{32}$$

for some $s^\alpha \in \mathrm{C}$ where **C** is the field of complex numbers.

Proof. The matrix $[I_n s^\alpha - \overline{A}]$ is nonsingular for every matrix $\overline{A} \in \Re^{n\times n}$.

From (31) and (32) it follows that the polynomial matrix (30) is nonsingular. Using elementary row operations the singular matrix E can be always reduced to the form $\begin{bmatrix} E_1 \\ 0 \end{bmatrix}$ where E_1 has the full row rank r_1 and L_1 is the matrix of elementary row operations.

Premultiplying (29) by L_1 we obtain

$$L_1[Es^\alpha - A]X = \begin{bmatrix} E_1 s^\alpha - A_1 \\ -A_2 \end{bmatrix} X = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} U \tag{33a}$$

where

$$L_1 E = \begin{bmatrix} E_1 \\ 0 \end{bmatrix},\ E_1 \in \Re^{r_1\times n},\ L_1 A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix},\ A_1 \in \Re^{r_1\times n},\ L_1 B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix},\ B_1 \in \Re^{r_1\times m}. \tag{33b}$$

Using (33) we can write the equation (28) in the form

$$E_1\frac{d^\alpha x}{dt^\alpha} = A_1 x + B_1 u, \tag{34a}$$

$$0 = A_2 x + B_2 u. \tag{34b}$$

The fractional differentiation of (34b) yields

$$0 = A_2 \frac{d^\alpha x}{dt^\alpha} + B_2 \frac{d^\alpha u}{dt^\alpha}. \tag{35}$$

From (34a) and (35) we have

$$\begin{bmatrix} E_1 \\ -A_2 \end{bmatrix} \frac{d^\alpha x}{dt^\alpha} = \begin{bmatrix} A_1 \\ 0 \end{bmatrix} x + \begin{bmatrix} B_1 \\ 0 \end{bmatrix} u + \begin{bmatrix} 0 \\ B_2 \end{bmatrix} \frac{d^\alpha u}{dt^\alpha}. \tag{36}$$

If the matrix $\begin{bmatrix} E_1 \\ -A_2 \end{bmatrix}$ is nonsingular then from (36) we have

$$\frac{d^\alpha x}{dt^\alpha} = \overline{A}_1 x + \overline{B}_0 u + \overline{B}_1 \frac{d^\alpha u}{dt^\alpha} \tag{37a}$$

where

$$\overline{A}_1 = \begin{bmatrix} E_1 \\ -A_2 \end{bmatrix}^{-1} \begin{bmatrix} A_1 \\ 0 \end{bmatrix},\ \overline{B}_0 = \begin{bmatrix} E_1 \\ -A_2 \end{bmatrix}^{-1} \begin{bmatrix} B_1 \\ 0 \end{bmatrix},\ \overline{B}_1 = \begin{bmatrix} E_1 \\ -A_2 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ B_2 \end{bmatrix}. \tag{37b}$$

If the matrix $\begin{bmatrix} E_1 \\ -A_2 \end{bmatrix}$ is singular then using elementary row operations we reduced the matrix $\begin{bmatrix} E_1 \\ -A_2 \end{bmatrix}$ to the form

$$L_2 \begin{bmatrix} E_1 \\ -A_2 \end{bmatrix} = \begin{bmatrix} E_2 \\ 0 \end{bmatrix} \tag{38}$$

and we repeat the procedure.

It is well known that if the condition (32) is satisfied then after μ steps of the procedure we obtain the nonsingular matrix

$$\begin{bmatrix} E_\mu \\ -A_\mu \end{bmatrix}. \tag{39}$$

Premultiplying the equation

$$\begin{bmatrix} E_\mu \\ -A_\mu \end{bmatrix} \frac{d^\alpha x}{dt^\alpha} = \begin{bmatrix} A_{\mu-1} \\ 0 \end{bmatrix} x + \begin{bmatrix} B_{\mu-1,0} \\ 0 \end{bmatrix} u + \begin{bmatrix} B_{\mu-1,1} \\ B_{\mu-1,0} \end{bmatrix} \frac{d^\alpha u}{dt^\alpha} + \ldots + \begin{bmatrix} 0 \\ B_\mu \end{bmatrix} \frac{d^{\alpha\mu} u}{dt^{\alpha\mu}} \tag{40}$$

by the inverse matrix $\begin{bmatrix} E_\mu \\ -A_\mu \end{bmatrix}^{-1}$ we obtain the desired equation

$$\frac{d^\alpha x}{dt^\alpha} = \overline{A}_\mu x + \overline{B}_0 u + \overline{B}_1 \frac{d^\alpha u}{dt^\alpha} + \ldots + \overline{B}_\mu \frac{d^{\alpha\mu} u}{dt^{\alpha\mu}} \tag{41a}$$

where

$$\begin{gathered} \overline{A}_\mu = \begin{bmatrix} E_\mu \\ -A_\mu \end{bmatrix}^{-1} \begin{bmatrix} A_{\mu-1} \\ 0 \end{bmatrix},\ \overline{B}_0 = \begin{bmatrix} E_\mu \\ -A_\mu \end{bmatrix}^{-1} \begin{bmatrix} B_{\mu-1,0} \\ 0 \end{bmatrix}, \\ \overline{B}_1 = \begin{bmatrix} E_\mu \\ -A_\mu \end{bmatrix}^{-1} \begin{bmatrix} B_{\mu-1,1} \\ B_{\mu-1,0} \end{bmatrix}, \ldots,\ \overline{B}_\mu = \begin{bmatrix} E_\mu \\ -A_\mu \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ B_\mu \end{bmatrix}. \end{gathered} \tag{41b}$$

The standard equation (41a) can be also obtained from the equation (40) by reducing the matrix (39) to the identity matrix I_n using the elementary row operations and this is equivalent to premultiplication of the equation (40) by suitable matrix of elementary row operations.

$$L_\mu \begin{bmatrix} E_\mu \\ -A_\mu \end{bmatrix} = I_n. \tag{42}$$

The desired polynomial matrix of elementary row operations (30) is given by

$$L(s^\alpha) = L_\mu \prod_{i=1}^{\mu} \text{diag}\,[I_{r_i}, I_{n-r_i} s^\alpha]. \tag{43}$$

Note that the matrix $I_{n-r_i} s^\alpha$ corresponds to the fractional differentiation of the algebraic equations.

The considerations can be easily extended to the linear electrical circuits described by the equation (13a).

Example 2. Consider the fractional descriptor electrical circuit shown on Figure 2 with given resistances R_1, R_2, R_3, inductances L_1, L_2, L_3 capacitance C and source voltages e_1, e_2.

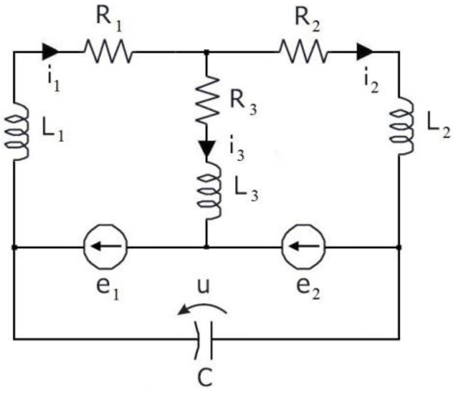


Fig. 2. Electrical circuit

Using the Kirchhoff's laws we can write the equations

$$e_1 = L_1 \frac{d^\beta i_1}{dt^\beta} + R_1 i_1 + L_3 \frac{d^\beta i_3}{dt^\beta} + R_3 i_3, \tag{44a}$$

$$e_2 = L_2 \frac{d^\beta i_2}{dt^\beta} + R_2 i_2 - L_3 \frac{d^\beta i_3}{dt^\beta} - R_3 i_3 \tag{44b}$$

$$i_3 = i_1 - i_2, \tag{44c}$$

$$u = e_1 + e_2. \tag{44d}$$

The equations can be written in the form

$$E \begin{bmatrix} \frac{d^\beta i_1}{dt^\beta} \\ \frac{d^\beta i_2}{dt^\beta} \\ \frac{d^\beta i_3}{dt^\beta} \\ \frac{d^\alpha u}{dt^\alpha} \end{bmatrix} = A \begin{bmatrix} i_1 \\ i_2 \\ i_3 \\ u \end{bmatrix} + B \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} \tag{45a}$$

where

$$E=\begin{bmatrix}L_1 & 0 & L_3 & 0\\ 0 & L_2 & -L_3 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix},\quad A=\begin{bmatrix}-R_1 & 0 & -R_3 & 0\\ 0 & -R_2 & R_3 & 0\\ 1 & -1 & -1 & 0\\ 0 & 0 & 0 & -1\end{bmatrix},\quad B=\begin{bmatrix}1 & 0\\ 0 & 1\\ 0 & 0\\ 1 & 1\end{bmatrix}. \tag{45b}$$

The pencil is regular since

$$\det[Es^\alpha - A]=\begin{vmatrix}s^\alpha L_1+R_1 & 0 & s^\alpha L_3+R_3 & 0\\ 0 & s^\alpha L_2+R_2 & -s^\alpha L_3-R_3 & 0\\ -1 & 1 & 1 & 0\\ 0 & 0 & 0 & 1\end{vmatrix} \tag{46}$$

$$=(s^\alpha L_1+R_1)[s^\alpha(L_2+L_3)+R_2+R_3]+(s^\alpha L_3+R_3)(s^\alpha L_2+R_2)\neq 0.$$

Defining

$$E_1=\begin{bmatrix}L_1 & 0 & L_3 & 0\\ 0 & L_2 & -L_3 & 0\end{bmatrix},\quad A_1=\begin{bmatrix}-R_1 & 0 & -R_3 & 0\\ 0 & -R_2 & R_3 & 0\end{bmatrix},\quad A_2=\begin{bmatrix}1 & -1 & -1 & 0\\ 0 & 0 & 0 & -1\end{bmatrix},$$
$$B_1=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix},\quad B_2=\begin{bmatrix}0 & 0\\ 1 & 1\end{bmatrix} \tag{47}$$

we can write the equation (45a) in the form

$$E_1\begin{bmatrix}\frac{d^\beta i_1}{dt^\beta}\\ \frac{d^\beta i_2}{dt^\beta}\\ \frac{d^\beta i_3}{dt^\beta}\\ \frac{d^\alpha u}{dt^\alpha}\end{bmatrix}=A_1\begin{bmatrix}i_1\\ i_2\\ i_3\\ u\end{bmatrix}+B_1\begin{bmatrix}e_1\\ e_2\end{bmatrix} \tag{48a}$$

and

$$0=A_2\begin{bmatrix}i_1\\ i_2\\ i_3\\ u\end{bmatrix}+B_2\begin{bmatrix}e_1\\ e_2\end{bmatrix}. \tag{48b}$$

The α fractional differentiation of (48b) yields

$$0=A_2\begin{bmatrix}\frac{d^\beta i_1}{dt^\beta}\\ \frac{d^\beta i_2}{dt^\beta}\\ \frac{d^\beta i_3}{dt^\beta}\\ \frac{d^\alpha u}{dt^\alpha}\end{bmatrix}+B_2\begin{bmatrix}\frac{d^\alpha e_1}{dt^\alpha}\\ \frac{d^\alpha e_2}{dt^\alpha}\end{bmatrix}. \tag{49}$$

From (48a) and (49) we have

$$\begin{bmatrix} E_1 \\ -A_2 \end{bmatrix} \begin{bmatrix} \frac{d^\beta i_1}{dt^\beta} \\ \frac{d^\beta i_2}{dt^\beta} \\ \frac{d^\beta i_3}{dt^\beta} \\ \frac{d^\alpha u}{dt^\alpha} \end{bmatrix} = \begin{bmatrix} A_1 \\ 0 \end{bmatrix} \begin{bmatrix} i_1 \\ i_2 \\ i_3 \\ u \end{bmatrix} + \begin{bmatrix} B_1 \\ 0 \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} + \begin{bmatrix} 0 \\ B_2 \end{bmatrix} \begin{bmatrix} \frac{d^\alpha e_1}{dt^\alpha} \\ \frac{d^\alpha e_2}{dt^\alpha} \end{bmatrix}. \tag{50}$$

The matrix

$$\begin{bmatrix} E_1 \\ -A_2 \end{bmatrix} = \begin{bmatrix} L_1 & 0 & L_3 & 0 \\ 0 & L_2 & -L_3 & 0 \\ -1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{51}$$

is nonsingular and premultiplying (50) by its inverse we obtain

$$\begin{bmatrix} \frac{d^\beta i_1}{dt^\beta} \\ \frac{d^\beta i_2}{dt^\beta} \\ \frac{d^\beta i_3}{dt^\beta} \\ \frac{d^\alpha u}{dt^\alpha} \end{bmatrix} = \overline{A} \begin{bmatrix} i_1 \\ i_2 \\ i_3 \\ u \end{bmatrix} + \overline{B}_0 \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} + \overline{B}_1 \begin{bmatrix} \frac{d^\alpha e_1}{dt^\alpha} \\ \frac{d^\alpha e_2}{dt^\alpha} \end{bmatrix} \tag{52a}$$

where

$$\overline{A}_1 = \begin{bmatrix} E_1 \\ -A_2 \end{bmatrix}^{-1} \begin{bmatrix} A_1 \\ 0 \end{bmatrix} = \frac{1}{L_1(L_2+L_3)+L_2L_3} \begin{bmatrix} L_2+L_3 & L_3 & L_2L_3 & 0 \\ L_3 & L_1+L_3 & -L_1L_3 & 0 \\ L_2 & -L_1 & -L_1L_2 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} -R_1 & 0 & -R_3 & 0 \\ 0 & -R_2 & R_3 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

$$\overline{B}_0 = \begin{bmatrix} E_1 \\ -A_2 \end{bmatrix}^{-1} \begin{bmatrix} B_1 \\ 0 \end{bmatrix} = \frac{1}{L_1(L_2+L_3)+L_2L_3} \begin{bmatrix} L_2+L_3 & L_3 & L_2L_3 & 0 \\ L_3 & L_1+L_3 & -L_1L_3 & 0 \\ L_2 & -L_1 & -L_1L_2 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix},$$

$$\overline{B}_1 = \begin{bmatrix} E_1 \\ -A_2 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ B_2 \end{bmatrix} = \frac{1}{L_1(L_2+L_3)+L_2L_3} \begin{bmatrix} L_2+L_3 & L_3 & L_2L_3 & 0 \\ L_3 & L_1+L_3 & -L_1L_3 & 0 \\ L_2 & -L_1 & -L_1L_2 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 1 & 1 \end{bmatrix}. \tag{52b}$$

The standard equation (52a) can be also obtained from the equation (50) by reducing the matrix (51) to the identity matrix I_4 using the elementary row operations

$$L[1+3\times(-L_3)],\ L[2+3\times L_3],\ L\left[1+2\times\left(\frac{L_3}{L_2+L_3}\right)\right],\ L\left[1\times\left(\frac{1}{L_1(L_2+L_3)+L_2L_3}\right)\right],$$

$$L[2+1\times L_3],\ L[3+1\times 1],\ L\left[1\times\left(\frac{1}{L_2+L_3}\right)\right],\ L[3+2\times(-1)], \qquad (53)$$

Using the elementary row operations (53) on the matrix

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & s^{\beta} & 0 \\ 0 & 0 & 0 & s^{\alpha} \end{bmatrix} \qquad (54)$$

we obtain the polynomial matrix

$$L[s^{\alpha}, s^{\beta}] = \begin{bmatrix} \frac{L_2+L_3}{L} & \frac{L_3}{L} & -\frac{L_2L_3}{L}s^{\beta} & 0 \\ \frac{L_3}{L} & \frac{L_1+L_3}{L} & \frac{L_1L_3}{L}s^{\beta} & 0 \\ \frac{L_2}{L} & -\frac{L_1}{L} & \frac{L_1L_2}{L}s^{\beta} & 0 \\ 0 & 0 & 0 & s^{\alpha} \end{bmatrix} \text{ and } L = L_1(L_2+L_3)+L_2L_3 \qquad (55)$$

satisfying the equations

$$L[s^{\alpha}][E\ \mathrm{diag}\,(s^{\beta}, s^{\beta}, s^{\beta}, s^{\alpha}) - A] = \mathrm{diag}\,(s^{\beta}, s^{\beta}, s^{\beta}, s^{\alpha}) - \overline{A}. \qquad (56)$$

Conclusion

A new polynomial approach is proposed to analysis of the standard and positive descriptor electrical circuits has been proposed. It has been shown (Theorem 4) that for given descriptor fractional electrical circuit the equivalent standard fractional electrical circuit can be found by premultiplication of the equation of the descriptor electrical circuit by suitable polynomial matrix of elementary row operations. The essence of the proposed method is demonstrated on simple positive fractional descriptor electrical circuit. The considerations can be easily extended to descriptor electrical circuits described by system of linear fractional equations with different orders [Kaczorek, 2010, 2011d]. An open problem is an extension of the approach to two-dimensional continuous-discrete fractional linear systems.

Acknowledgment

This work was supported by National Science Centre in Poland under work S/WE/1/11.

Bibliography

[Bru et. all, 2003a] R. Bru, C. Coll, S. Romero-Vivo and E. Sanchez. Some problems about structural properties of positive descriptor systems, Lecture Notes in Control and Inform. Sci., vol. 294, Springer, Berlin, 233-240, 2003.

[Bru et. all, 2003b] R. Bru, C. Coll and E. Sanchez. Structural properties of positive linear time-invariant difference-algebraic equations, Linear Algebra Appl., vol. 349, 1-10, 2003.

[Campbell et. all, 1976], S.L. Campbell, C.D. Meyer and N.J. Rose. Applications of the Drazin inverse to linear systems of differential equations with singular constant coefficients, SIAMJ Appl. Math., vol. 31, no. 3, 411-425, 1976.

[Dai, 1989] L. Dai. Singular control systems, Lectures Notes in Control and Information Sciences, Springer-Verlag, Berlin, 1989.

[Guang-Ren, 2010] D. Guang-Ren. Analysis and Design of Descriptor Linear Systems, Springer, New York, 2010.

[Farina et. all, 2000] L. Farina and S. Rinaldi. Positive Linear Systems; Theory and Applications, J. Wiley, New York, 2000.

[Kaczorek, 2013a] T. Kaczorek. Application of Drazin inverse to analysis of descriptor fractional discrete-time linear systems with regular pencils, Int. J. Appl. Math. Comput. Sci., vol. 23, no. 1, 29-34, 2013.

[Kaczorek, 2011a] T. Kaczorek. Checking of the positivity of descriptor linear systems by the use of the shuffle algorithm, Archives of Control Sciences, vol. 21, no. 3, 2011, 287-298, 2011.

[Kaczorek, 2014a] T. Kaczorek. Drazin inverse matrix method for fractional descriptor continuous-time linear systems, Submitted to Bull. Pol. Ac. Techn. Sci., 2014.

[Kaczorek, 2004] T. Kaczorek. Infinite eigenvalue assignment by output-feedbacks for singular systems, Int. J. Appl. Math. Comput. Sci., vol. 14, no. 1, 19-23, 2004.

[Kaczorek, 1992] T. Kaczorek. Linear Control Systems, vol. 1, Research Studies Press J. Wiley, New York, 1992.

[Kaczorek, 2014b] T. Kaczorek. Minimum energy control of positive fractional descriptor continuous-time linear systems, IET Control Theory and Applications, vol. 8, no. 4, 219–225, 2014.

[Kaczorek, 2008] T. Kaczorek. Fractional positive continuous-time systems and their reachability, Int. J Appl. Math. Comput. Sci., vol. 18, no. 2, 223-228, 2008.

[Kaczorek, 2011c] T. Kaczorek. Positive electrical circuits and their reachability, Archives of Electrical Engineering, vol. 60, no. 3, 283-301, 2011.

[Kaczorek, 2011b] T. Kaczorek. Pointwise completeness and pointwise degeneracy of 2D standard and positive Fornasini-Marchesini models, COMPEL, vol. 30, no. 2, 656-670, 2011.

[Kaczorek, 2009] T. Kaczorek. Pointwise completeness and pointwise degeneracy of standard and positive fractional linear systems with state-feedbacks, Archives of Control Sciences, vol. 19, 295-306, 2009.

[Kaczorek, 2010] T. Kaczorek. Positive linear systems with differetntial fractional order, Bull. Pol. Acad. Sci. Techn., vol. 58, no. 3., 453-458, 2010.

[Kaczorek, 2002] T. Kaczorek. Positive 1D and 2D Systems, Springer-Verlag, London, 2002.

[Kaczorek, 2011d] T. Kaczorek. Positive linear systems consisting of n subsystems with different fractional orders, IEEE Trans. Circuits and Systems, vol. 58, no. 6, 1203-1210, 2011.

[Kaczorek, 2011e] T. Kaczorek. Reduction and decomposition of singular fractional discrete-time linear systems, Acta Mechanica et Automatica, vol. 5, no. 4, 62-66, 2011.

[Kaczorek, 2011f] T. Kaczorek. Selected Problems of Fractional Systems Theory, Springer-Verlag, Berlin, 2011.

[Kaczorek, 2011g] T. Kaczorek. Singular fractional discrete-time linear systems, Control and Cybernetics, vol. 40, no. 3, 753-761, 2011.

[Virnik, 2008] E. Virnik. Stability analysis of positive descriptor systems, Linear Algebra and its Applications, vol. 429, 2640-2659, 2008.

Authors' Information

Tadeusz Kaczorek – Bialystok University of Technology, Faculty of Electrical Engineering, Wiejska 45D, 15-351 Bialystok, Poland; e-mail: kaczorek@isep.pw.edu.pl

Major Fields of Scientific Research: Control systems theory, fractional systems, positive systems

Computational Intelligence Models

FUZZY TECHNIQUES IN ROBOTIC SYSTEMS CONTROL

Arkady Yuschenko

Abstract: *New area of robotic systems application is in the field closely connected with the human society. That is the service robots, the robotic medical systems, robotics for risky environment etc. The most important problem in robotic control system now is not the technical realization but the compatibility the robotic system with human. One of the ways to approach the control techniques the possibilities of human-operator without any special knowledge in the field of robotics and control systems is the fuzzy logic techniques. Some examples of such approach are presented below.*

Keywords*: fuzzy logic, linguistic variables, mobile robot, fuzzy-neural net, operation planning, scene recognition, human-robot interface.*

Introduction

The modern trend of robotics is the inclusion of intelligent robots in human society. New area of robotic applications may be found in medicine, space investigation, agriculture etc. as service robotics. To create the robot compatible with human we suppose it is necessary to provide the anthropomorphism at the level of perception and reasoning. The most suitable approach to solve the problem are the fuzzy technologies. Some new results in this field are presented in the paper.

Three levels are possible to distinguish in the control system of intelligent robot: reflection level, tactical level and the level of the purpose designation. The lower level may be named by analogy to the living creature as reflection level. The stereotypes of reactions are forming here in accordance with the current situation. The situation itself may be determined as a composition of fuzzy conditions. Such approach makes it possible to use the human's perception and experience to determine the necessary rules for robot sensory system. Using the suitable algorithms of fuzzy inference the fuzzy controller may be formed to control the movements of manipulation or mobile robot. Mention that in the latter case the membership functions depend of current robot position because the scale of the image is continuously changing due the robot movement.

The next level named a tactical make it possible to compose the more complicate modes of behavior using the typical elements from the knowledge base. The latter accumulates the human experience in form of fuzzy rules. The principles of automatic inherence of the sequence of movements may be the same as in the case of human reasoning. Some of them are also under discussion.

We suppose that the hybrid neural-fuzzy approach is the most suitable for the realization of fuzzy controller for intelligent robot. It allows not only to form the basic rules of behavior but also to tune the fuzzy controller parameters using the backward error propagation algorithm. Usually to teach robot to fulfill the typical operations we use the principle "the teacher – the pupil". The new direction in this field is the self-teaching of robot using the "emotion block" as a part of the neural network.

The upper level of the system under discussion is the level of the purpose designation by human. The modern stage of the problem is the dialogue planning between human and robot using the language close to natural one. The dialogue is the most convenient form for human operator both for situation analysis and for robot control. The fuzzy form of relations in the external world and linguistic variables allow cooperating with robots the persons without a special preparation. That is why we hope that our approach will help to solve the problem of human – robot compliance. The theoretic background of the fuzzy logic control is well known to day. So the main attention we pay below to the applications of the fuzzy technique to the robotic control systems realization.

Representation of the external world by robotic system

A modern robotic system can to move in the external world and to fulfill the necessary complicated operations using manipulators. Robot is equipped with the computer vision system and a net of information devices proved the reconstruction of the image of real situation in the external world. So the robot can work autonomously. But at the same time the autonomy of robot is an illusion. The task for robot is stated by human – either previously or in the real time scale. The more deep is the robotics penetration through modern society life the more actual become the problem of compatibility between robots and human.

The first problem here is the "mutual" understanding of the external situation which has not been known beforehand. To reach such understanding it is necessary to apply in the human-robot dialogue the same relations which are natural for human in his usual life. That is the natural relations [Pospelov , 1989] such as <*a1 is far and a little to the right at a2*>. Such types of relations are extensional ones. To describe the current scene also the intentional relations are employed. For example *to be adjacent to; R_2 - to be inside of; R_3 - to be outside of; R_4 - to be in the centre of; R_5 - to be on the same line as; R_6 – to be on the same plane as; R_7 - to have zero projection on, R_8 - to be on the surface of.* Two unary relations are also proposed: R_{00} – *to be horizontal* and R_{01} – *to be vertical*, as well as 28 elementary spatial binary relations.

The set of specified objects in the current scene, the relations between them and transformation rules constitute a formal language for environment representation that is similar to a natural language. Scene description in this language allows for a formal semiotic representation that uses the spatial-temporal relations logic. For example a complex relation <*a_1 is on the surface S far and to the right*> can be written as $(a_1\ R_8\ S)\&(a_0\ d_5\ f_7\ a_1)$, where a_0 – is the observer, with respect to whom the distance *d* and orientation *f* relations are formulated.

Since the environment is ever-changing due the motion of the observed objects as well as to the motion of the robot itself, the scene description changes in time respectively. This circumstance requires that we take into account not only spatial but also temporal relations in the external world, such as *to be simultaneous with, to be prior to, to follow* etc.

Practically the task of the environment recognition may be solved by application of 3D computer video system (CVS) with the structural light [Mikhaylov, 2005]. The system consist of high definition TV-camera and pulse source, which generate special light flat matrix on work scene. In result CVS forms 3D description of the real work scene as Cartesian coordinates data file. Moreover, CVS forms set of the elementary triangles; every triangle includes three nearest points of scene.

The algorithms of the objects identification are based on the fuzzy features comparison of the objects detected on the working scene and of the objects accessed in the data. At first it is necessary to apply the fuzzification procedure for such geometric features measured by computer vision system as *D* – the distance for the obstacle, *W* – the width of the object, *L* , *R* – the distances from the main axis of robot platform to the left or to the right edge of the obstacle, *H* – the height of the obstacle etc. For example: *the object is* *extensive, flat, horizontal, does not high, to the right, does not distant*, etc. The linguistic variables allow the user easily to expand the list of the possible objects using their characteristic features.

The fuzzy classifier realized the Mamdany fuzzy inference procedure determined the type of the object [Volodin . 2011]. The first maximum approach is applied on the defuzzyfication stage. The choice of the fuzzy features for object classification as well of the production rules and membership functions depend of the task under consideration. It may be different depending on the importance of those either another features for the task to fulfill by the robotic system. Analysis of the typical situation for indoor work proved to determine the types of the obstacles such as *Wall, Door, Threshold, Block*, etc. The typical obstacles are presented in the data base by the vectors of fuzzy features obtained from the CVS. Among them are the coordinates and dimension of the object, the mutual disposition of the objects, their geometrical characteristics etc. Note that the geometric description of the obstacles and the separate objects may demand the special robot movements of cognitive type to obtain the information necessary to determine the type of unknown object. The fuzzy features allow identifying of the object also in the case when the object description is incomplete for obstacles or shadows. In such cases the cognitive behavior of the robot may be planned to seek the information necessary for the object classification. For example the special position of the robot vision system may be needed for the satisfactory observation.

The identification algorithms is the fuzzy logic inference using the previously determined classification rules and fuzzy features of the reference model of the objects possessed in the knowledge base. The base also contains the production rules of object classification and the corresponding maneuvers representation. The fuzzy features of the objects are represented by the membership functions for the corresponding linguistic variables. These functions are to be correlated with the robot and its computer vision system technical characteristics. For example the meaning *the doorway width is enough* is determined by the robot dimension itself. The meaning *the obstacle high does not enough* is determined by the dimension of the robot platform. The meaning *the obstacle is distant* is restricted by the CVS characteristics. The membership functions of the linguistic variables are to be assigned beforehand during the calibration procedure of the 3D CVS as a part of a definite mobile robot.

The situation in the working scene in the whole may be represented as a kind of fuzzy chart where binary relations between all the objects have been determined. Robot is one of these objects. Such a fuzzy chart allows the robot control system to determine its own position in relation for the bench-marks.

The important peculiarity of the mobile robot control system is in the fact that the scale of the image is continuously changing due the robot movement. That is why the membership functions for mobile robot sensing system depend of the distance. In this case the 3D membership functions had been introduced.

But the main advantage of a fuzzy identification system is the possibility of the robot –human dialogue about the environment and the objects around the robot, which is necessary for the prescribed operations fulfillment.

Mobile robot behavior in the 3D external world

It seems quite possible to use the term “behavior” for autonomous mobile robot work in the partially undetermined world. We can represent the stereotypes of behavior determined by the characteristic situation as production rules “*if the situation is S_i than the tactics is T_i*“. The *tactics* here is the list of production rules of typical stereotypes of behavior represented with the linguistic variables. These rules determine the previously described robot movement for ever typical situation. For example: “*if the obstacle is near and to the left then go around slow and anticlockwise*”. The typical movement (*go around slow and anticlockwise*) in its turn also may be determined by a list of production rules contained in the data base [Yuschenko, 2012].

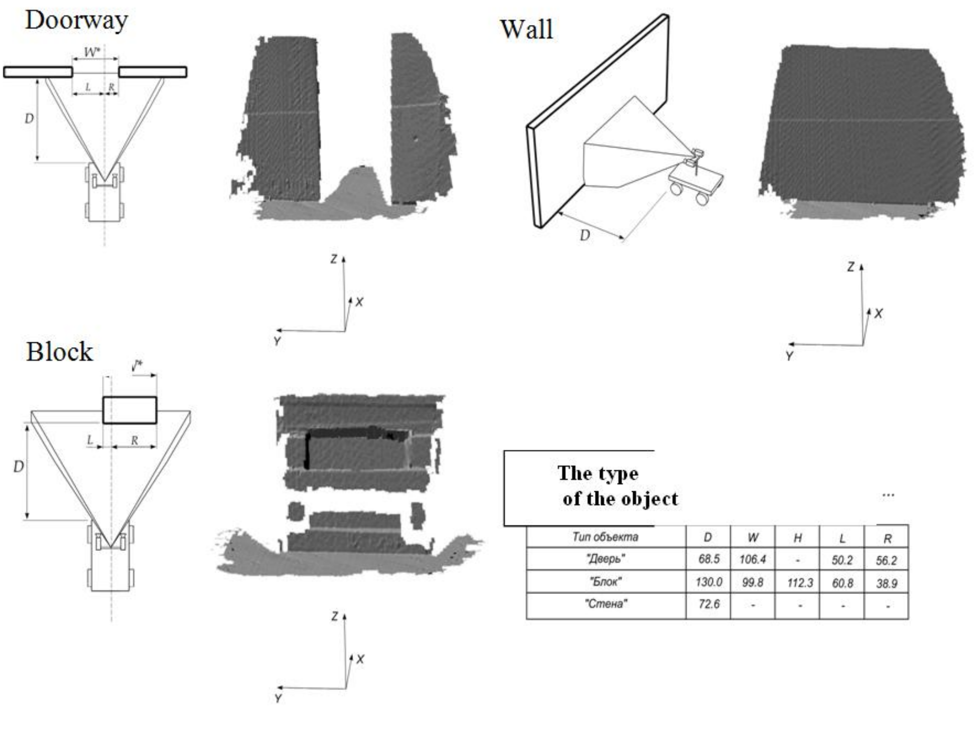
Such a data base allow to the user to develop the new tactics such as: “to pursuit the moving objects”, “move to the prescribed object”, “pass the doorway”, “go around the obstacle” etc.

In general the tactics of robot behavior may be determined with a task frame: <current situation > <robot characteristics > <tactics name > <the objects and obstacles > < the conditions of the operation feasibility> .

Robot’s characteristics such as dimension parameters (weight, power of drivers, possible velocity etc.) are contained in the data base. These characteristics together with the features of the environment (relief, the ground parameters, the types of the obstacles) determine the possibility of the prescribed operation feasibility. The latter may include also the post condition to be satisfied after operation fulfillment such as a stable robot position. The conditions of feasibility for such operation as the objects persecution or the obstacle avoidance are to be classified in accordance with their features beforehand.

In the case when some slots of the operation frame are empty the robot has itself to plan a movement of cognitive type to find the necessary information [Yuschenko, 2005] For example the parameters of the doorway to pass through. The cognitive operations are also contains in the robot knowledge base as a list of fuzzy production rules.

In our experimental investigation we supposed that the most part of indoor situations might be represented as a combination of the next typical objects: “Wall”, “Threshold”, “Block”, “Left angle”, ”Right angle”, “Doorway” (Fig.1). Every object detected by the 3D computer vision system is classified as one of the typical objects via the fuzzy inference procedure discussed above. Then the necessary movement trajectory may be planned.



Тип объекта	D	W	H	L	R
"Дверь"	68.5	106.4	-	50.2	56.2
"Блок"	130.0	99.8	112.3	60.8	38.9
"Стена"	72.6	-	-	-	-

Fig. 1. The typical obstacles and their parameters

The next linguistic variables were introduced: “The object position” (to the right, to the left, in the centre); “The object height” (tall, middle, law), “The object width”(wide, compact), “Distance to the object (dangerous, safety, the edge of the vision area)., The number of the terms of the linguistic variables may be easily expanded if necessary.

The experiment has been fulfilled in the laboratory rooms using a small mobile robot equipped with the 3D vision system described above (Fig. 2).

As it was mentioned above the scale of the image producing by the CVS based on the robot platform is changing during the robot movement. So it is important to take into consideration the membership functions as the functions of current distance to the obstacle. So for the terms of the linguistic variable “object height” the membership functions are formed taking in consideration the angle of the TV camera arrangement on board of the robot (Fig.3). The object classified as tall on the edge of the vision area may be classified as middle at the distance dimension. The same is for the linguistic variable “object position”. The object situated at the centre may be founded as to the right at small distance (Fig.4). It make possible to plan the maneuver beforehand. For example begin to plan the walking round the block at the safe distance from the obstacle.

The identification of the obstacles has been developed in accordance with the prescribed classification rules. For example: “slow and wide object is a Threshold”; “slow and compact object is a Block”; “ Tall or middle object to the left is a Left angle”;“ Tall and wide object in the center is a Wall” . Note that the conjunction of the features related both to the features of the object and to its position leads to description of the situation in the whole from the description of separate objects. An example of the real experiment presented on the Fig.1 where the robot determined the object “The doorway” as “The tall object to the left (Left angle) and The tall object to the right (Right angle)at almost the same distance”. To realize the operation: “To walk into the doorway” the condition of feasibility of operation “The doorway is free” is to be examined. The condition has not been fulfilled in the case when the system has found another obstacle (Block for example) between the Left and the Right angles.

After the situation has been determined completely the necessary trajectory may be planned as sequence of the typical operations contained in the data base as it described above. Consideration of 3D-membership functions (Fig.4,b) taking in consideration their relation from distance allowed the planning of the robot movement beforehand without stopping the robot. Such mode of control can improve the safety of operations and eliminate the possibility of collision. It may be extremely important in the case of appearance of moving objects allowing to correct the trajectory and to avoid the possible collision.

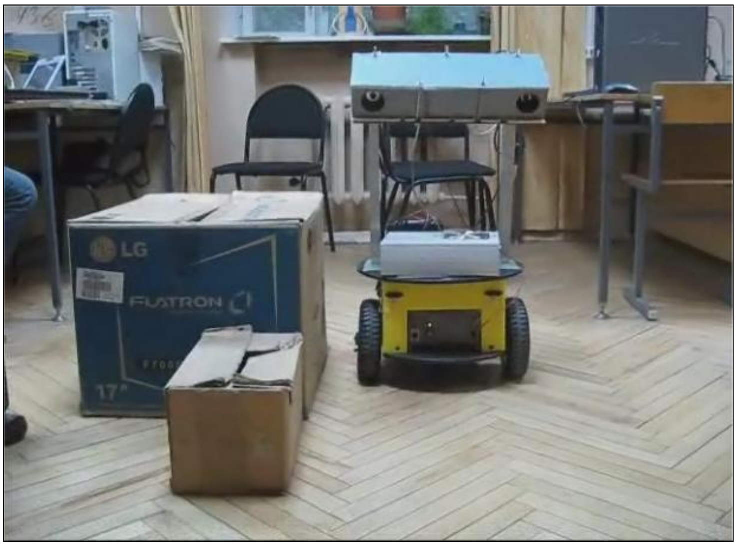


Fig. 2. Mobile robot equipped with the 3D computer vision system in experiment.

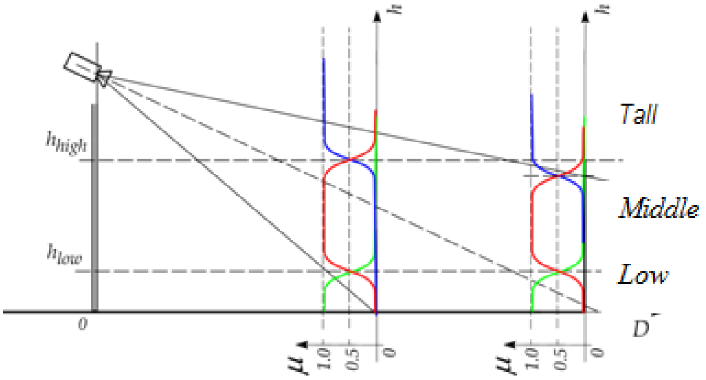
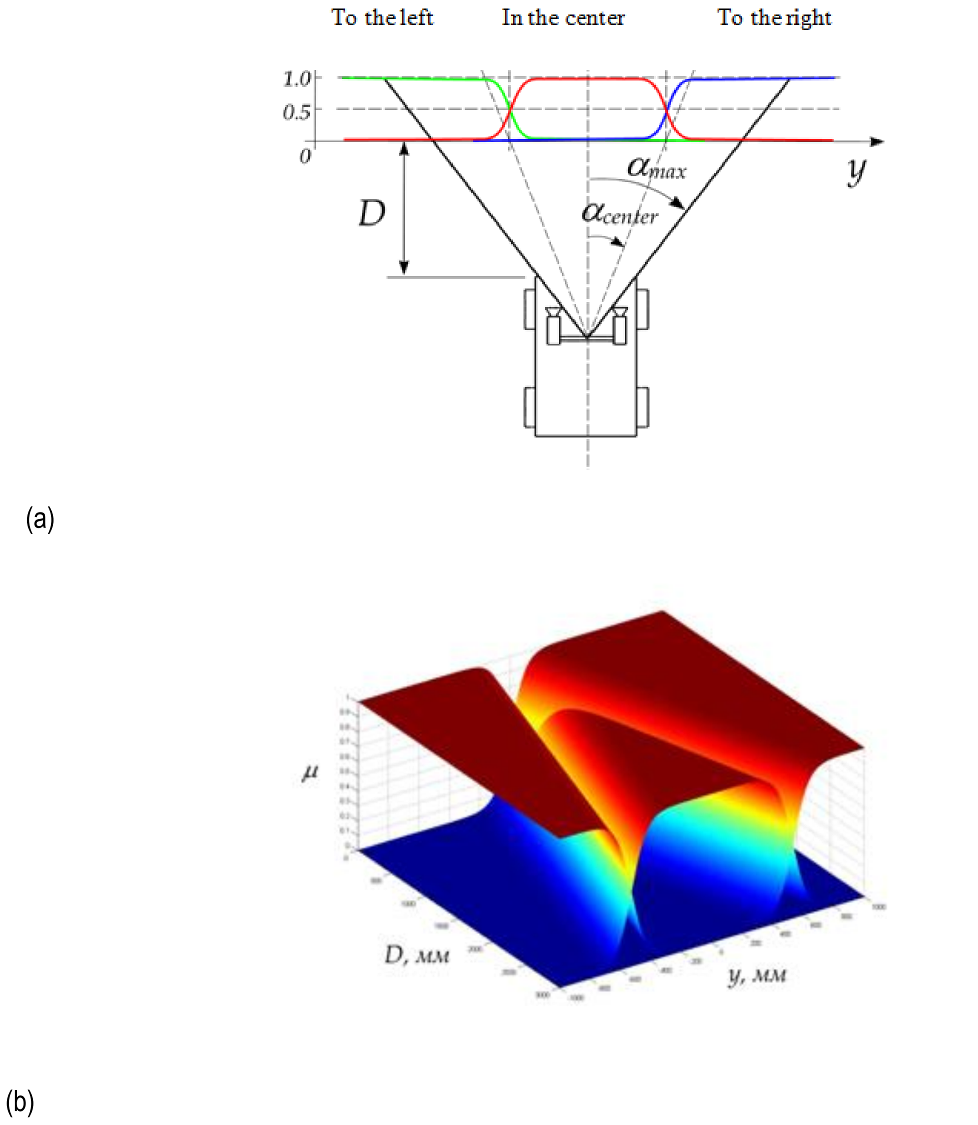


Fig. 3. Membership functions “Tall”, “Middle”, “Low”



(a)

(b)

Fig.4. Membership functions “To the left”, “To the right”, “In the centre” (a) and their relation from the distance (b)

Mobile fire reconnaissance robot control.

The image of the current situation may include not only the fuzzy variables determine the space relations and the linear dimensions of the objects but other fuzzy peculiarities of the environment. Such problem arrives in the task of autonomous control of a mobile robot applied for an indoor fire guarding. Such robot is equipped with the sensors of temperature, humidity, pressure, concentration of CO and CO_2 etc.. All the information proceeded in real time scale may be presented in the form of the fuzzy reports. For example " *the temperature is low, the humidity is high , the smoke is dense"*. Such information make it possible to understand the real position of the robot to the fire source. There are different strategies of the robot behavior are possible. In the case of real danger the robot has to leave the dangerous zone. Otherwise it has to move towards in the direction of the unobservable fire source and to recognize the type of the fire and the strategy of the firefighting [Tachkov, 2012]. The vector of the fire parameters is computed on board in real time scale. One of them (or some parameters forming a scalar estimation) is chosen as a most important. The robot movement now may be interpreted as the movement in the scalar field of the fire parameter along the gradient of the field. In this case at the every moment of time the robot moves toward the maximum of the chosen parameter – main fire factor MFF (for example, the temperature of the environment or the concentration of CO_2).

Let the gradient direction is determined with the angle $\varphi(\vec{r},t)$ oriented along the increasing meaning of the MFF (let it be a temperature) (Fig. 5) Then the equality takes place:

$$\varphi(\vec{r},t)=\theta(t)+\eta(t),$$

where $\eta(t)$ – is the angle of anticipation. From this equation we may obtain the velocity of the anticipation angle changing:

$$\frac{d\eta}{dt}=\upsilon\cdot\vec{u}(\theta)\cdot\vec{\nabla}\varphi-\omega+\frac{\partial\varphi}{\partial t}$$

Where $\vec{\nabla}\varphi$ is the MFF gradient in the point of the robot position in the direction determined by the angle φ, υ is the linear velocity of the robot, ω is the angular velocity and $\vec{u}(\theta)$ – is an ort of OX axis .

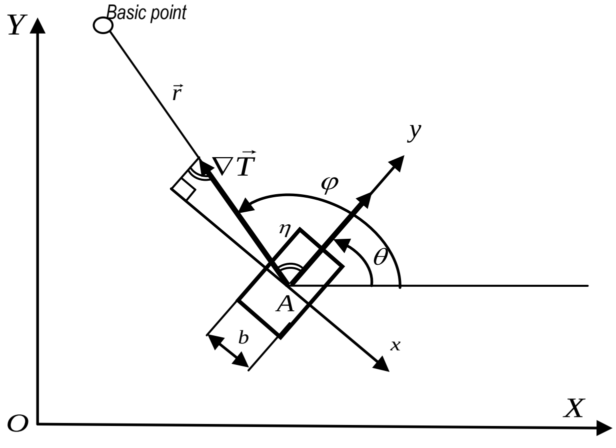


Fig.5. Self-guidance of the robot to the basic point ($\nabla\vec{T}$ is the gradient of the temperature field)

The kinematics equations of the robot are the next

$$\begin{cases} \dfrac{dr}{dt} = -\upsilon\cdot\cos(\varphi-\theta) \\ r\dfrac{d\varphi}{dt} = \upsilon\cdot\sin(\varphi-\theta) \end{cases}.$$

We may also put $r = 2K\cdot\frac{\upsilon}{\omega}\cdot\sin\eta$ and $\omega = \frac{2K\cdot\upsilon\cdot\sin\eta}{r}$, $\omega = K_H\cdot\dot{\varphi}$

The last equation is the law of proportional navigation with the coefficient K_H=2K , K>0.

For the fire in the closed room the proportional navigation law may be prescribed as

$$\omega = \frac{K_H\cdot\upsilon\cdot\Delta T}{b\cdot\left(T(r)-T_0+\varepsilon\right)},$$

where ΔT is the difference of the temperature from the left and the right boards of the robot platform $T(r)$ is the temperature in the point A, and T_0 is the initial temperature in the room and ε is a small value.

The formulas above allow to realize the fuzzy mode of control of the mobile robot based on the principle of the proportional navigation. The linguistic variables for input of the controller we choose the distance from obstacles to the right, to the left, and forward, the temperature, angular velocity of the MFF gradient. The output linguistic variables are the linear and the angular velocities of the robot movement.

The base of knowledge of the robot includes the lists of production rules for obstacles avoidance and rules for self-guidance to the virtual basic point. The membership functions for input variables connected with the distance, direction to the obstacles and temperature were determined with expert appraisal. The problem was to form such functions for the angular velocity of the MFF gradient. We choose three levels of the angular velocity – negative, positive and small (Fig. 6). The output membership functions are of singlton type. To tune the membership functions we used the linear equation between the output variable – the angular velocity of the robot platform from the input variable $\dot{\phi}$: $\omega = K_H\cdot\dot{\varphi}$. Defuzzification was produced by the centre of gravity method (Fig. 6).

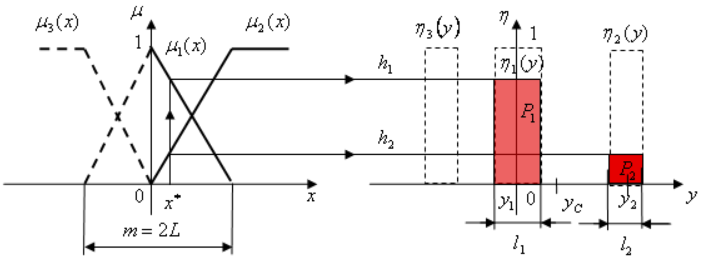


Fig. 6 The membership function tuning

The analytic presentation of the membership functions are the next:

$$\mu_1(x)=\begin{cases}0, & h_1(y_C)<0\\ \dfrac{-f\cdot(y_C-y_2)}{(y_C-y_1)-f\cdot(y_C-y_2)}, & 0\le h_1(y_C)\le 1\\ 1, & h_1(y_C)>1\end{cases}\quad \mu_2(x)=\begin{cases}0, & h_2(y_C)<0\\ \dfrac{y_C-y_1}{(y_C-y_1)-f\cdot(y_C-y_2)}, & 0\le h_2(y_C)\le 1\\ 1, & h_2(y_C)>1\end{cases},$$

where $f=l_2/l_1$ is the weight of the membership functions $\eta_1(y)$ and $\eta_2(y)$.

The experimental mobile robot with the temperature sensors at the right and the left board is shown at Fig.7. The mass of the robot is 1,26 kg and the width of the platform is about 0,1 m. The robot equipped with the visual sensors and the necessary algorithms to avoid the obstacles (see above).

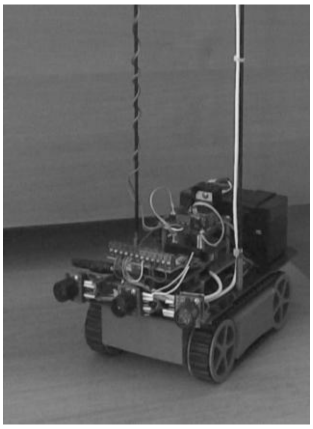


Fig. 7 The fire reconnaissance mobile robot

The robot is equipped with DC-drivers and can work autonomously. It is necessary to underline that the proportional navigation coefficient *K_H* determined in accordance with the parameters of the robot and of the control system. The experiments show that the stability and the quality of the transient responses also depend of this coefficient. To determine the stability interval the absolute stability criterion was applied.

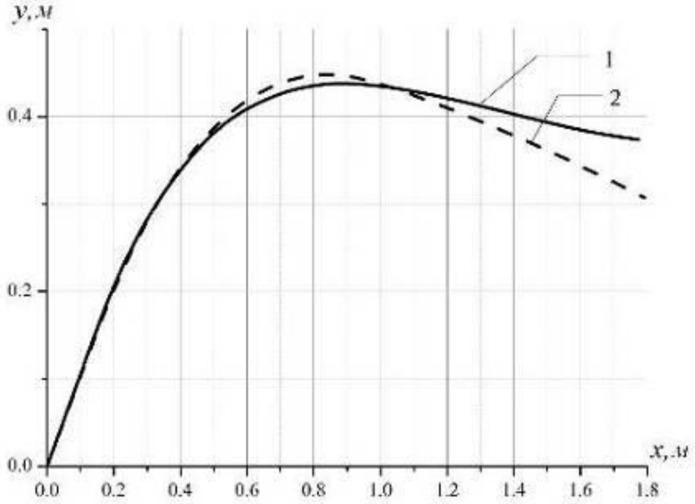


Fig 8. The divergence of the real trajectory 2 of the robot from the preplanned one 1.

Experiments showed that the divergence between the experimental robot trajectories and the simulation results is not more thay 5-7 cm (Fig.8), which is a satisfactory result taking in configuration the robot sizes.

Medical Intravascular microrobot control using a fuzzy finite state automata concept

The aim of the investigation is the development of the instruments and technologies for diagnostics and treatment of tube-like human's organs such as blood vessels and intestines. The medical microrobots may be applied to move along the tube-like organs by the same way as a worm (i.e. using a peristaltic principle). An experimental model of microrobot (Fig.9) has three segments which contracting successively to ensure progressive movement of the device [Savrasov, 2006]. The diameter of the robot is smaller than the same of the blood vessel. So it is pressed to the internal cover of the vessel by the special planes to avoid the thrombosis of the vessel. Every segment of robot contain three contact elements, pressure sensors and a regulator to control the pressure of the elements to the internal surface of the vessel. Aboard the robot a micro-video camera has been mounted to inform the surgeon of the situation inside the vessel and other micro-devices. The supporting plates carry tensometric sensors to control the contact forces. The driver of the robot is of hydraulic type with physiologic solution to avoid the danger of embolism.

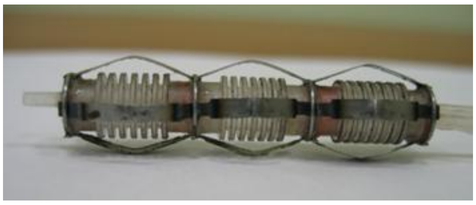


Fig.9. A microrobot model

Microrobot is a part of the robotic system including also a hydro-driver mounted in the stationary part of the system and intelligent interface of the operator. The surgeon-operator has opportunity to observe the inner surface of the vessel by visual sensors mounted aboard the robot and to control the robot movement along the vessel. The construction of the microrobot has to guarantee the stable position of the robot in the moving blood flow and its movement inside the vessel without any damage of the inner surface.

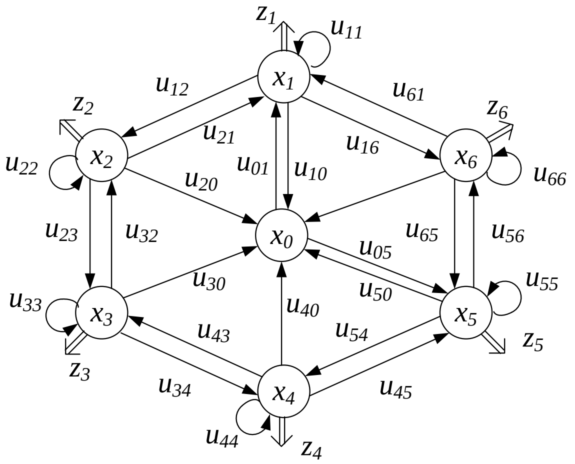


Fig.10. Finite state automata model determining the robot movement

The peculiarity of the microrobot movement is in its cyclic type. The segments of the robot contracts successively and during the cycle they may posses only one of two states – active (contracted) or passive (stretched). The conditions of the transition from one state to another depends of the contact forces values necessary to reach a stable position of the robot in the vessel. Such conditions may be formulated as a fuzzy production rules. So the mathematical model based on the fuzzy finite state automata concept has been proposed [Voynov , 2005].

Graph of the finite state fuzzy automata (Fig.10) has 6 states corresponing the states of microrobot on Fig.9. The states indicated as: x_i (i = 0…6); input alphabet is $U = \{u_{ij}\}$, and u_{ij} is a command to change the state i into state j (i = 0,…6, j = 0, …,6), output alphabet is Z = $\{z_i\}$, and z_i is a symbol of transition into state *i.* After receiving the operator's command such as "forward, backward, stop" etc. the control system forms a chain of operations to perform the command.

The active contact planes also have been presented as a finite fuzzy automata controlled with the corresponding regulator. The input linguistic variable is the force of pressure of the contact plane and output variable is the velocity of the pressure process. Individual tuning of the membership functions provide the smooth pressing process necessary for it safety. The same model describes the pressure sensors. So the logic level of robot control is a 3-levels net of finite state automata (Fig. 11). The lower level includes regulators of the contact sensors and controls the process of pressing the contact plates to the internal surface of the vessel. The middle level contains the monitors of three segments of robot providing their coordinated work in accordance with the prescribed task. The upper level is a monitor of the driver providing the robot movement by computing the control signals to the robot segments in accordance with the operator's command.

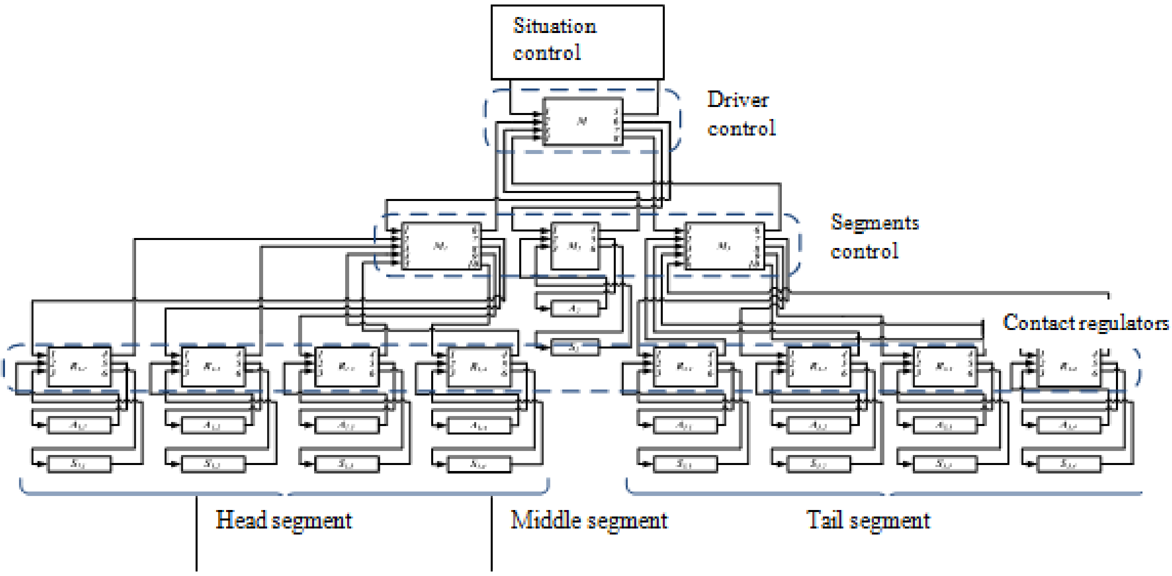


Fig. 11

In the problem cases of calcite deposition on the inner surface of the vessel the robotic system can support the surgeon's decision. The micro-video device forms an image of the situation which is analyzing by the computer expert system and possible decisions are presenting to the surgeon. In the crucial situations the control may be transferred to the surgeon.

Mention that the principles discussed above does not depend of the number of the robot segments. So the same model may be developed for robot with more number of segments which is more close to it biological prototype – the swarm. Such microrobots will possess more wide possibilities to penetrate to the distant parts of the human body to perform diagnostics or medical operation in the less traumatically way for the patient and make such operations safer.

Conclusion

The fuzzy control technique are wide spread to day for every application connected fith fuzzy information proceeding and with human factors. The approach allows to solve such problems as obstacles identification in the indoor navigation, the navigation in extremal condition with an information shortage, in medicine as a mean making the diagnostic and surgeon operation safer for patient. The fuzzy control system allows robot both to fulfill the stated task autonomously in partly determined environment and work under operator's control using the natural speech dialogue. We suppose that fuzzy logic technique makes it possible to control the service robots for the person without special knowledge which is one of the main goals of the future human-robotic society.

Bibliography

[Pospelov , 1989] Kandrashina E.Yu., Litvinceva L.V., Pospelov D.A. Representation of knowledge of time and space in intelligent systems. Nauka, Moscow, 1989.

[Mikhaylov , 2005] Mikhaylov B.B., Volodin Yu. S., Orlov A.V. Calibration of 3D computer vision system. Proc. of 16-th Conf. "Extreme Robotics", S-Pb, 2005: 314-322.

[Volodin, 2011] Mikhaylov B.B., Volodin Yu. S. Yuschenko A.S. Fuzzy obstacles classification by mobile robot using 3D computer vision system. Proc. of 6-th Conf. "Integrated models and soft computing in artificial intelligence", 2011, Kolomna: 372-380.

[Yuschenko, 2012] Yuschenko A.S. Fuzzy logic in mobile robots control systems. Vestnik BMSTU, Priborostroenie, 2012, № 6: 29-43.

[Yuschenko, 2005] Yuschenko A.S. Intelligent planning in robot operation. Mekhatronica, Vol. 3, 2005: 5-18.

[Tachkov, 2012] Tachkov A.A., Yuschenko A.S. Interactive System for Control of Fire-Fighting Reconnaissance Robot. Vestnik BMSTU, Priborostroenie, 2012,№ 6: 106-119.

[Savrasov, 2006] Savrasov G.V., Voynov V.V., Yuschenko A.S. et.al. Intravascular microrobot. Biomedical technology and radioelectronics, 2006, № 11: p.44-48.

[Voynov, 2005] Voynov V.V., Yuschenko A.S. Adaptive control of the microrobot for intravascular diagnostics. Transactions of the Conference "Extreme robotics", S-Pb, 2005,v.5: 126-133.

Authors' Information

Arkady Semenovich Yuschenko – Dr.Sc.(Techn.), Professor of the chair "Robotic Systems" , Moscow State Technical University n.a. N.E.Bauman, Russia, 105037, Moscow, Izmailovskaya sq.,7, Nauchno-Uchebny Centr "Robototechnika" , BMSTU, Tel. Office: +7(499) 165 17 01, Tel/ fax: +7(499) 367 05 81, e-mail: robot@bmstu.ru.

Major Fields of Scientific Research: Control theory, Control of "intelligent" robots, Human-robot interface and compliance

ROBUST ADAPTIVE FUZZY CLUSTERING FOR DATA WITH MISSING VALUES

Yevgeniy Bodyanskiy, Alina Shafronenko

Abstract: *the datasets clustering problem often encountered in many applications connected with Data Mining and Exploratory Data Analysis. Conventional approach to solving these problems requires that each observation may belong to only one cluster, although a more natural situation is when the vector of features with different levels of probabilities or possibilities can belong to several classes.* This situation is subject of consideration of fuzzy cluster analysis, intensively developing today.

In many practical Data Mining tasks, including clustering, data sets may contain gaps, information in which, for whatever reasons, is missing. More effective in this situation are approaches based on the mathematical apparatus of Computational Intelligence and first of all artificial neural networks and different modifications of classical fuzzy c-*means (FCM) method.*

Real data often contain abnormal outliers of different nature too, for example, measurement errors or distributions with "heavy tails". In this situation classic FCM is not effective because the objective function based on the Euclidean metric, only reinforces the impact of outliers. In such conditions it is advisable to use robust objective functions of special form that suppress influence of outliers. For information processing in a sequential mode adaptive procedures for on-line fuzzy clustering have been proposed, which are in fact on-line modifications of FCM, where instead of the Euclidean metric robust objective functions that weaken the influence of outliers were used.

Situation when data set contains missing values and outliers in the fuzzy clustering problem was not analyzed, although such a situation can arise in many practical applications. Therefore the development of twice robust (for missing values and outliers) fuzzy clustering algorithm has theoretical interest and practical sense.

The problem of fuzzy adaptive on-line clustering of data distorted by missing values and outliers sequentially supplied to the processing when the original sample volume and the number of distorted observations are unknown is considered. The probabilistic and possibilistic clustering algorithms for such data, that are based on the strategy of nearest prototype, partial distances and similarity measure of a special kind that weaken or overwhelming outliers are proposed.

Keywords: *Fuzzy clustering, Kohonen self-organizing network, learning rule, incomplete data with gaps and outliers.*

ACM Classification Keywords: *1.2.6 [Artificial Intelligence]: Learning – Connectionism and neural nets; 1.2.8 [Artificial Intelligence]: Problem Solving, Control Methods, and Search – Control theory; 1.5.1 [Pattern Recognition]: Clustering – Algorithms.*

Introduction

The problem of data sets clustering often occurs in many practical tasks, and for its solution has been successfully used mathematical apparatus of computational intelligence [Rutkowski, 2008] and first of all, artificial neural networks [Marwala, 2009] and soft computing methods [Klawonn, 2006] (in the case of overlapping classes) is usually assumed that original array is specified a priori and processing is made in batch mode. Here as one of the most effective approach based on using FCM [Bezdek, 1981], that is modified for the situation with missing values [Hathaway, 2001] which comes as a result to minimize the objective function with constraints of special form. In [Bodyanskiy, 2012; Bodyanskiy, 2013] adaptive fuzzy clustering procedures have been proposed for processing the data sequences containing an unknown quantity of missing values, realizing the problem in on-line mode and characterized by numerical simplicity. These procedures are in fact hybrid of T. Kohonen neural network [Kohonen, 1995] with the special form of a neighborhood function.

Real data often contain outliers of different nature, for example, measurement errors or distributions with "heavy tails". In this situation classic FCM is not effective because the objective function based on the Euclidean metric, only reinforces the impact of outliers. In such conditions it is advisable to use robust objective functions of special form [Dave, 1997], that suppress influence of outliers. For information processing in a sequential mode, in [Bodyanskiy, 2005; Kokshenev I., 2006] adaptive on-line fuzzy clustering procedures have been proposed, which are in fact on-line modifications of FCM, where instead of the Euclidean metric robust objective functions, weaken the influence of outliers where used.

Situation when data set contains both missing values and outliers in the fuzzy clustering problem are not considered, although such a situation can arise in many practical applications. Therefore the development of twice robust (for missing values and outliers) fuzzy clustering algorithms has theoretical interest and practical sense.

Problem statement

Baseline information for solving the tasks of clustering in a batch mode is the sample of observations, formed from N n-dimensional feature vectors $X = \{x_1, x_2, ..., x_N\} \subset R^n, x_k \in X, k = 1,2,...,N$. The result of clustering is the partition of original data set into m classes $(1 < m < N)$ with some level of membership $U_q(k)$ of k-th feature vector to the q-th cluster $(1 \le q \le m)$. Incoming data previously are centered and standardized by all features, so that all observations belong to the hypercube $[-1,1]^n$. Therefore, the data for clustering form array $\tilde{X} = \{\tilde{x}_1, ..., \tilde{x}_k, ..., \tilde{x}_N\} \subset R^n$, $\tilde{x}_k = (\tilde{x}_{k1}, ..., \tilde{x}_{ki}, ..., \tilde{x}_{kn})^T$, $-1 \le \tilde{x}_{ki} \le 1$, $1 < m < N$, $1 \le q \le m$, $1 \le i \le n$, $1 \le k \le N$. Note that traditionally adopted in Kohonen's maps (SOM) data transformation to the form $\|\tilde{x}_k\| = 1$ in this case does not make sense, because if x_k contains missing value - calculation rules of such vector is impossible, and if x_k contains outlier in one of the components - $\tilde{x}_k$ will be practically the same as the corresponding unit vector of the feature space. Transformation $-1 \le \tilde{x}_{ki} \le 1$ leads to the fact that the non deformed data concentrate in the vicinity of zero, and the data with outliers – near -1 and +1. Furthermore, we introduce additional sub arrays data [Hathaway, 2001]: $\tilde{X} = \{\tilde{x}_1, ..., \tilde{x}_k, ..., \tilde{x}_N\} \subset R^n$, $\tilde{x}_k = (\tilde{x}_{k1}, ..., \tilde{x}_{ki}, ..., \tilde{x}_{kn})^T$, $-1 \le \tilde{x}_{ki} \le 1$, $1 < m < N$, $1 \le q \le m$, $1 \le i \le n$, $1 \le k \le N$.

We have to develop numerically simple on-line procedure for partitioning in sequential mode to the data processing $\tilde{x}_k$ on m perhaps overlapping classes, while it is not known in advance whether $\tilde{x}_k$ is undistorted or contains missing values and outliers. Furthermore, it is assumed that the amount of information under processing is not known in advance and is increased with time.

Adaptive fuzzy clustering data with missing values based on the nearest prototype strategy

Nearest prototype strategy (NFS), proposed in [Hathaway, 2001], is a modification of FCM-algorithm and leads to the replacement of missing components of the vector observations $\tilde{x}_{ki} \in X_G$ by estimates of the corresponding component prototypes (centroids) of the clusters computed using FCM. Thus for each $\tilde{x}_{ki} \in X_G$ it's possible to find the prototype $w_q = (w_{q1,\ldots,}w_{qi,\ldots,}w_{qn})^T$ nearest to $\tilde{x}_k$ in the sense of the partial distance (PD)

$$D_P^2(\tilde{x}_k, w_q) = \frac{n}{\delta_{k\Sigma}} \sum_{i=1}^{n} (\tilde{x}_{ki} - w_{qi})^2 \delta_{ki} \tag{1}$$

where

$$\delta_{ki} = \begin{cases} 0 \mid \tilde{x}_{ki} \in X_G, \\ 1 \mid \tilde{x}_{ki} \in X_F, \end{cases}$$

$$\delta_{k\Sigma} = \sum_{i=1}^{n} \delta_{ki}$$

$w_q^{(\tau+1)} = \underset{q}{\arg\min}\{D_P^2(\tilde{x}_k, w_1^{(\tau+1)}), \ldots, D_P^2(\tilde{x}_k, w_m^{(\tau+1)})\}$, then instead $\tilde{x}_{ki} \in X_G$ input estimate $\hat{x}_{ki} \in w_{qi}$ used in place of the missing components.

In [Bodyanskiy, 2013] adaptive fuzzy clustering procedure based on the NPS was introduced:

$$\begin{cases} U_q^{(\tau+1)}(k) = \dfrac{(\|\hat{x}_k^{(\tau)} - w_q^{(\tau)}(k)\|^2)^{\frac{1}{1-\beta}}}{\sum\limits_{l=1}^{m} (\|\hat{x}_k^{(\tau)} - w_l(k)\|^2)^{\frac{1}{1-\beta}}}, \\ \hat{x}_{ki}^{(\tau)} = w_{qi}^{(\tau)}(k), \quad w_q^{(\tau)}(k) = \underset{q}{\arg\min}\{D_P^2(\tilde{x}_k, w_1^{(\tau)}(k)), \ldots, D_P^2(\tilde{x}_k, w_m^{(\tau)}(k))\}, \\ w_q^{(Q)}(k) = w_q^{(0)}(k+1), \\ w_q^{(\tau+1)}(k+1) = w_q^{(\tau)}(k+1) + \eta(k+1)(U_q^{(Q)}(k))^{\beta}(\hat{x}_k^{(\tau)} - w_q^{(\tau)}(k+1)) \ , \end{cases} \tag{2}$$

where $\beta > 1$ - parameter that is called fuzzyfier and defines "vagueness" of boundaries between classes, $\eta(k+1)$ - learning rate parameter, $\tau = 0,1,2,\ldots$ - accelerated computing time between two real-time instance k and $k+1$ occurs Q iteration in accelerated time.

From the last relation (2) it follows that centroids setting made using the Kohonen self-learning rule "Winner Takes More» (WTM) with the neighborhood function $(U_q^{(Q)}(k))^{\beta}$ having the Cauchian form.

The main disadvantage of FCM and other so-called probabilistic fuzzy clustering algorithms associated with a constraint on the levels of membership of each vector-image, which is equal to one, which gives sense of probability and membership, but it is not always correct in terms of the problem being solved. To remove this restriction in [Keller, 2005] possibilistic fuzzy clustering algorithm (PCM) was introduced, and in [Bodyanskiy, 2012; Bodyanskiy, 2013] - its adaptive version for the case of data containing missing values, having the form:

$$\begin{cases} U_q^{(\tau+1)}(k) = \dfrac{1}{1+(\dfrac{\left\|\hat{x}_k^{(\tau)} - w_q(k)\right\|^2}{\mu_q^{(\tau)}(k)})^{\frac{1}{\beta-1}}}, \\ \hat{x}_{ki}^{(\tau)} = w_{qi}^{(\tau)}(k), \quad w_q^{(\tau)}(k) = \underset{q}{\arg\min}\{D_P^2(\tilde{x}_k, w_1^{(\tau)}(k)),\ldots, D_P^2(\tilde{x}_k, w_m^{(\tau)}(k))\}, \\ w_q^{(Q)}(k) = w_q^{(0)}(k+1) \\ w_q^{(\tau+1)}(k+1) = w_q^{(\tau)}(k+1) + \eta(k+1)(U_q^{(Q)}(k))^\beta(\hat{x}_{k+1}^{(\tau)} - w_q^{(\tau)}(k+1)) \ , \\ \mu_q^{(\tau+1)} = \dfrac{\sum\limits_{p=1}^{k}(U_q^{(\tau+1)}(p))^\beta \left\|\hat{x}_p^{(\tau+1)} - w_q^{(\tau+1)}(k)\right\|^2}{\sum\limits_{p=1}^{k}(U_q^{(\tau+1)}(p))^\beta}, \end{cases} \tag{3}$$

where the scalar parameter $\mu \geq 0$ determines the distance at which level of membership equals to 0.5, i.e. if $\left\|\tilde{x}_k - w_q\right\|^2 = \mu_q(k)$, then $w_q(k) = 0.5$.

Algorithms (2), (3) have confirmed working capacity in solving a number of problems [Bodyanskiy, 2013], however, since they are based on the use of Euclidean distance, they do not possess stability to outliers.

Adaptive fuzzy robust data clustering based on the similarity measure

As already mentioned, to solve the problem of fuzzy clustering of data containing outliers the special objective functions of the form [Dave, 1997; Bodyanskiy, 2005; Kokshenev I., 2006] can be used, by some means these anomalies overwhelming, and the problem itself is associated with the minimization of these functions. From a practical point of view it is more convenient to use instead of the objective functions, based on the metrics, the so-called measures of similarity (SM) [Sepkovski, 1974], which are subject to more soft conditions than metrics:

$$\begin{cases} S(\tilde{x}_k, \tilde{x}_p) \geq 0, \\ S(\tilde{x}_k, \tilde{x}_p) = S(\tilde{x}_p, \tilde{x}_k), \\ S(\tilde{x}_k, \tilde{x}_k) = 1 \geq S(\tilde{x}_k, \tilde{x}_p) \end{cases}$$

(no triangle inequality), and clustering problem can be "tied" to maximize these measures.

If the data are transformed so that $-1 \leq \tilde{x}_{ki} \leq 1$ the measure of similarity can be structured so as to suppress unwanted data lying at the edges of interval $[-1,1]$. Figure 1 illustrates the use of similarity measure based on Cauchy function with different parameters width $\sigma^2 < 1$

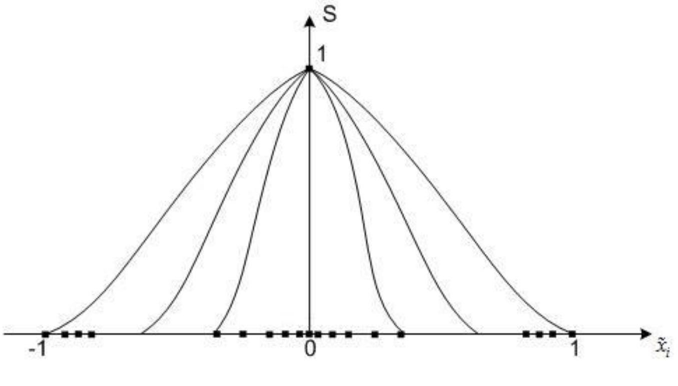


Fig. 1 Similarity measure based on the Cauchy function

By choosing the width parameter σ^2 of functions

$$S(\tilde{x}_k, w_q) = \frac{1}{1+\frac{\left\|\tilde{x}_k - w_q\right\|^2}{\sigma^2}} = \frac{\sigma^2}{\sigma^2 + \left\|\tilde{x}_k - w_q\right\|^2} = \qquad (4)$$
$$= \frac{\sigma^2}{\sigma^2 + D^2(\tilde{x}_k, w_q)}$$

is possible to exclude the effect outliers, that in principle cannot be done using the Euclidean metric

$$D^2(\tilde{x}_k, w_q) = \left\|\tilde{x}_k - w_q\right\|^2. \qquad (5)$$

Further, by introducing the objective function based on similarity measure (4),

$$E_S(U_q(k), w_q) = \sum_{k=1}^{N}\sum_{q=1}^{m} U_q^{\beta}(k) S(\tilde{x}_k, w_q) = \sum_{k=1}^{N}\sum_{q=1}^{m} \frac{U_q^{\beta}(k)\sigma^2}{\sigma^2 + \left\|\tilde{x}_k - w_q\right\|^2},$$

probabilistic constraints

$$\sum_{q=1}^{m} U_q(k) = 1,$$

Lagrange function

$$L_S(U_q(k), w_q, \lambda(k)) = \sum_{k=1}^{N}\sum_{q=1}^{m} \frac{U_q^{\beta}(k)\sigma^2}{\sigma^2 + \left\|\tilde{x}_k - w_q\right\|^2} + \sum_{k=1}^{N} \lambda(k)(\sum_{q=1}^{m} U_q(k) - 1) \qquad (6)$$

(here $\lambda(k)$ - indefinite Lagrange multipliers) and solving the system of Karush-Kuhn-Tucker equations, we get the solution

$$\begin{cases} U_q(k) = \frac{(S(\tilde{x}_k, w_q))^{\frac{1}{1-\beta}}}{\sum_{l=1}^{m} (S(\tilde{x}_k, w_l))^{\frac{1}{1-\beta}}}, \\ \lambda(k) = -(\sum_{l=1}^{m} (\beta S(\tilde{x}_k, w_l))^{\frac{1}{1-\beta}})^{1-\beta}, \\ \nabla_{w_q} L_S(U_q(k), w_q, \lambda(k)) = \sum_{k=1}^{N} U_q^{\beta}(k) \frac{\tilde{x}_k - w_q}{(\sigma^2 + \left\|\tilde{x}_k - w_q\right\|^2)^2} = \vec{0}. \end{cases} \qquad (7)$$

The last equation (7) has no analytic solution, so to find a saddle point of the Lagrangian (6) we can use the procedure of Arrow-Hurwitz-Uzawa, as a result of which we obtain the algorithm

$$\begin{cases} U_q(k+1) = \dfrac{(S(\tilde{x}_{k+1}, w_q))^{\frac{1}{1-\beta}}}{\sum\limits_{l=1}^{m} (S(\tilde{x}_{k+1}, w_l))^{\frac{1}{1-\beta}}}, \\ w_q(k+1) = w_q(k) + \eta(k+1) U_q^{\beta}(k+1) \dfrac{\tilde{x}_{k+1} - w_q}{(\sigma^2 + \|\tilde{x}_{k+1} - w_q\|^2)^2} = w_q(k) + \eta(k+1)\varphi_q(k+1)(\tilde{x}_{k+1} - w_q) \end{cases} \qquad (8)$$

where

$$\varphi_q(k+1) = \frac{\tilde{x}_{k+1} - w_q}{(\sigma^2 + \|\tilde{x}_{k+1} - w_q\|^2)^2}$$

neighbourhood robust functions of WTM-self-learning rule.

Assuming the fuzzifier value $\beta = 2$ we get a robust variant of FCM:

$$\begin{cases} U_q(k+1) = \dfrac{(S(\tilde{x}_{k+1}, w_q))}{\sum\limits_{l=1}^{m} (S(\tilde{x}_{k+1}, w_l))}, \\ w_q(k+1) = w_q(k) + \eta(k+1) \dfrac{U_q^2(k+1)}{(\sigma^2 + \|\tilde{x}_{k+1} - w_q\|^2)^2}. \end{cases}$$

Further, using the concept of accelerated time, it's possible to introduce robust adaptive probabilistic fuzzy clustering procedure in the form

$$\begin{cases} U_q^{(\tau+1)}(k) = \dfrac{(S(\tilde{x}_k, w_q^{(\tau)}(k)))^{\frac{1}{1-\beta}}}{\sum\limits_{l=1}^{m} (S(\tilde{x}_k, w_l^{(\tau)}))^{\frac{1}{1-\beta}}}, \\ w_q^{(Q)}(k) = w_q^{(0)}(k+1), \\ w_q^{(\tau+1)}(k+1) = w_q^{(\tau)}(k+1) + \eta(k+1) \dfrac{(U_q^{(Q)}(k))^{\beta}}{(\sigma^2 + \|\tilde{x}_{k+1} - w_q^{(\tau)}(k+1)\|^2)^2} (\tilde{x}_{k+1} - w_q^{(\tau)}(k+1)), \end{cases} \qquad (9)$$

with the decision of each membership $\tilde{x}_k$ to a specific cluster takes on the maximum value of similarity measure.

Similarly, it's possible to synthesize a robust adaptive algorithm for possibilistic [Klawonn, 1998] fuzzy clustering using criterion

$$E_S(U_q(k), w_q, \mu_q) = \sum_{k=1}^{N} \sum_{q=1}^{m} U_q^{\beta}(k) S(\tilde{x}_k, w_q) + \sum_{q=1}^{m} \mu_q \sum_{k=1}^{N} (1 - U_q(k))^{\beta}.$$

Solving the problem of optimization, we obtain the solution:

$$\begin{cases} U_q(k+1) = \left(1 + \left(\dfrac{S(\tilde{x}_{k+1}, w_q(k))}{\mu_q(k)}\right)\right)^{-1}, \\ w_q(k+1) = w_q(k) + \eta(k+1) U_q^{\beta}(k+1) \dfrac{\tilde{x}_{k+1} - w_q(k)}{(\sigma^2 + \left\| \tilde{x}_{k+1} - w_q(k) \right\|^2)^2}, \\ \mu_q(k+1) = \dfrac{\sum\limits_{p=1}^{k+1} U_q^{\beta}(p) S(\tilde{x}_p, w_q(k+1))}{\sum\limits_{p=1}^{k+1} U_q^{\beta}(p)}, \end{cases} \qquad (10)$$

receiving at $\beta = 2$ the form

$$\begin{cases} U_q(k+1) = \dfrac{1}{1 + \dfrac{S(\tilde{x}_{k+1}, w_q(k))}{\mu_q(k)}}, \\ w_q(k+1) = w_q(k) + \eta(k+1) \dfrac{U_q^2(k+1)}{(\sigma^2 + \left\| \tilde{x}_{k+1} - w_q(k) \right\|^2)^2} (\tilde{x}_{k+1} - w_q(k)), \\ \mu_q(k+1) = \dfrac{\sum\limits_{p=1}^{k+1} U_q^2(p) S(\tilde{x}_p, w_q(k+1))}{\sum\limits_{p=1}^{k+1} U_q^2(p)}. \end{cases}$$

And, finally, introducing the accelerated time we obtain the procedure

$$\begin{cases} U_q^{(\tau+1)}(k) = \dfrac{1}{1 + \left(\dfrac{S(\tilde{x}_k, w_q^{(\tau)}(k))}{\mu_q^{(\tau)}(k)}\right)^{\frac{1}{\beta-1}}}, \\ w_q^{(Q)}(k) = w_q^{(0)}(k+1), \\ w_q^{(\tau+1)}(k+1) = w_q^{(\tau)}(k+1) + \eta(k+1) \dfrac{(U_q^{(Q)}(k))^{\beta}}{(\sigma^2 + \left\| \tilde{x}_{k+1} - w_q^{(\tau)}(k+1) \right\|^2)^2} (\tilde{x}_{k+1} - w_q^{(\tau)}(k+1)), \\ \mu_q^{(\tau+1)}(k) = \dfrac{\sum\limits_{p=1}^{k} (U_q^{(\tau+1)}(p))^{\beta} S(\tilde{x}_p, w_q^{(\tau+1)}(k))}{\sum\limits_{p=1}^{k} (U_q^{(\tau+1)}(p))^{\beta}}. \end{cases} \qquad (11)$$

Adaptive fuzzy robust data clustering with missing values

For solving the problem of robust data clustering with missing values let’s introduce the partial similarity measure (PCM), which is a hybrid of a partial distance (PD) (1) and similarity measure (SM) (4). It is easily to see that such PSM has the form

$$S_P(\tilde{x}_k, w_q) = \frac{\sigma^2}{\sigma^2 + D_P^2(\tilde{x}_k, w_q)}, \tag{12}$$

that allows to obtain the desired properties of algorithms based on procedures described above.

So, on the basis of the procedures (2) and (9) we can introduce the robust adaptive probabilistic fuzzy clustering algorithm for data with missing values:

$$\begin{cases} U_q^{(\tau+1)}(k) = \dfrac{(S_P(\hat{x}_k^{(\tau)}, w_q^{(\tau)}(k)))^{\frac{1}{\beta-1}}}{\sum\limits_{l=1}^{m}(S_P(\hat{x}_k^{(\tau)}, w_l^{(\tau)}))^{\frac{1}{\beta-1}}}, \\ \hat{x}_{ki}^{(\tau)} = w_{qi}^{(\tau)},\ w_q^{(\tau)}(k) = \underset{q}{\arg\max}\{S_P(\tilde{x}_k^{(\tau)}, w_1^{(\tau)}(k)), \ldots, S_P(\tilde{x}_k^{(\tau)}, w_m^{(\tau)}(k))\}, \\ w_q^{(Q)}(k) = w_q^{(0)}(k+1), \\ w_q^{(\tau+1)}(k+1) = w_q^{(\tau)}(k+1) + \eta(k+1)\dfrac{(U_q^{(Q)}(k))^\beta}{(\sigma^2 + \left\|\hat{x}_{k+1}^{(\tau)} - w_q^{(\tau)}(k+1)\right\|^2)^2}(\hat{x}_{k+1}^{(\tau)} - w_q^{(\tau)}(k+1)), \end{cases} \tag{13}$$

based on procedures (3) and (11), also we can write the robust adaptive algorithm for possibilistic fuzzy clustering of data with missing values:

$$\begin{cases} U_q^{(\tau+1)}(k) = \dfrac{1}{1 + \left(\dfrac{S^{-1}(\hat{x}_k, w_q^{(\tau)}(k))}{\mu_q^{(\tau)}(k)}\right)^{\frac{1}{\beta-1}}}, \\ \hat{x}_{ki}^{(\tau)} = w_{qi}^{(\tau)},\ w_q^{(\tau)}(k) = \underset{q}{\arg\max}\{S_P(\tilde{x}_k^{(\tau)}, w_1^{(\tau)}(k)), \ldots, S_P(\tilde{x}_k^{(\tau)}, w_m^{(\tau)}(k))\} \\ w_q^{(Q)}(k) = w_q^{(0)}(k+1), \\ w_q^{(\tau+1)}(k+1) = w_q^{(\tau)}(k+1) + \eta(k+1)\dfrac{(U_q^{(Q)}(k))^\beta}{(\sigma^2 + \left\|\hat{x}_{k+1} - w_q^{(\tau)}(k+1)\right\|^2)^2}(\hat{x}_{k+1}^{(\tau)} - w_q^{(\tau)}(k+1)), \\ \mu_q^{(\tau+1)}(k) = \dfrac{\sum\limits_{p=1}^{k}(U_q^{(\tau+1)}(p))^\beta S_P^{-1}(\hat{x}_p, w_q^{(\tau+1)}(k))}{\sum\limits_{p=1}^{k}(U_q^{(\tau)}(p))^\beta}. \end{cases} \tag{14}$$

Thus, the use of partial similarity measure based on partial distance (1), allows us to solve the problem of fuzzy clustering of data containing both missing values and outliers.

Conclusion

The problem of robust adaptive fuzzy clustering algorithms is considered, allowing in on-line mode to process distorted data containing both outliers and missing values is considered. The basis of the proposed algorithms is using of classical procedures as fuzzy c-means of J. Bezdek, T. Kohonen self-learning, as well as specially introduced similarity measure allowing to process distorted information. The algorithms are simple in numerical implementation, being essentially gradient optimization procedures for objective functions of special form.

Bibliography

[Rutkowski, 2008] L.Rutkowski. Computational Intelligence. Methods and Techniques. Berlin-Heidelberg: Springer-Verlag, 2008.

[Marwala, 2009] T Marwala. Computational Intelligence for Missing Data Imputation, Estimation, and Management: Knowledge Optimization Techniques. Hershey-New York: Information Science Reference, 2009.

[Hathaway, 2001] R.J. Hathaway, J.C Bezdek. Fuzzy c-means clustering of incomplete data. IEEE Trans. on Systems, Man, and Cybernetics, 31, №5, 2001, P. 735-744.

[Klawonn, 2006] F. Klawonn. Reducing the Number of Parameters of a Fuzzy System Using Scaling Functions. Soft Computing 10, 2006, P 749-756

[Bezdek, 1981] J.C. Bezdek. Pattern Recognition with Fuzzy Objective Function Algorithms. – N.Y.: Plenum, 1981.

[Bodyanskiy, 2012] BodyanskiyYe., Shafronenko A., Volkova V. Adaptive clustering of incomplete data using neuro-fuzzy Kohonen network. In “Artificial Intelligence Methods and Techniques for Business and Engineering Applications” – Rzeszow-Sofia: ITHEA, 2012. – P. 287-296.

[Bodyanskiy, 2013] BodyanskiyYe., Shafronenko A., Volkova V. Adaptive fuzzy probabilistic clusrering of incomplete data. Int.J. Information Models and Analysis. – 2013. – 2. - №2. – P. 112-117.

[Bodyanskiy, 2013] BodyanskiyYe., Shafronenko A., Volkova V. Neuro fuzzy Kohonen network for incomplete data clustering using optimal completion strategy// Proc. East West Fuzzy Coll., 20th Zittau Fuzzy Coll. – Zittau / Goerlitz: HS, 2013. – P. 214 – 223.

[Kohonen, 1995] T. Kohonen. Self-Organizing Maps. Berlin: Springer-Verlag, 1995.

[Dave, 1997] Dave R.N., Krishnapuram R. Robust clustering methods: A unified view// IEEE Trans. on Fuzzy Systems. – 1997. – 5. - №2. – P.270-293.

[Bodyanskiy, 2005] BodyanskiyYe., Gorshkov Ye., Kokshenev I., Kolodyazhniy V. Robust recursive fuzzy clustering algorithms//Proc. East West Fuzzy Coll. – Zittau/Goerlitz: HS, 2005. – P. 301-308.

[Bodyanskiy, 2005] BodyanskiyYe. Computational intelligence techniques for data analisis. – Lectures Notes on Informatics Vol. P. – 72. – Bonn: GI, 2005. – P. 15-36.

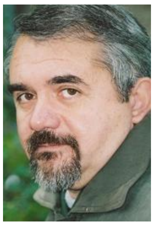
[Kokshenev I., 2006] Kokshenev I., BodyanskiyYe., Gorshkov Ye., Kolodyazhniy V. Outlier resistant recursive fuzzy clustering algorithm / In “Computational Intelligence: Theory and Application”. – Ed. by B.Reusch-Advances in Soft Computing, Vol. 38. – Berlin-Heidelberg: Springer-Verlag, 2006. – P. 647-652.

[Keller, 2005] L. Keller, R. Krishnapuram, N.R. Pal. Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. – N.Y.:Springer Science + Business Media, Inc., 2005.

[Sepkovski, 1974] Sepkovski J.J. Quantified coefficients of association and measurement of similarity // J. Int. Assoc. Math. – 1974. – 6. – №2. – P. 135-152.

[Klawonn, 1998] F. Klawonn, A. Keller. Fuzzy Clustering with Evolutionary Algorithms. Intelligent Systems 13, 1998, P. 975-991.

Authors' Information



Yevgeniy Bodyanskiy – Professor, Dr. – Ing. habil., Scientific Head of Control Systems Research Laboratory, Kharkiv National University of Radio Electronics, 14 Lenin Ave., Office 511, 61166 Kharkiv, Ukraine; e-mail: bodya@kture.kharkov.ua

Major Fields of Scientific Research: Artificial neural networks, Fuzzy systems, Hybrid systems of computational intelligence

Alina Shafronenko –Ph.D. student in Artificial Intelligence dept., Kharkiv National University of Radio Electronics, 14 Lenin Ave., Office 517, 61166 Kharkiv, Ukraine;

e-mail: alinashafronenko@gmail.com

Major Fields of Scientific Research: neural networks, neural network processing of data with missing values, fuzzy clustering, clustering of data

THE LEAST SQUARES SUPPORT VECTOR MACHINE BASED ON A NEO-FUZZY NEURON

Yevgeniy Bodyanskiy, Oleksii Tyshchenko, Daria Kopaliani

***Abstract**: The paper presents a fuzzy least squares support vector machine (LS-FSVM) which is implemented with the help of neo-fuzzy neurons (NFN) and which is essentially a zero order Takagi-Sugeno fuzzy inference system. The proposed LS-FSVM-NFN is numerically simple because it's generated with NFNs, it also has a small number of adjustable parameters and high speed associated with the possibility of applying the second order optimization learning procedures to process data in an online-mode.*

***Keywords**: Fuzzy support vector machine, neo-fuzzy neuron, learning procedure, time series.*

***ACM Classification Keywords**: 1.2.6 [Artificial Intelligence]: Learning – Connectionism and neural nets.*

Introduction

Currently, artificial neural networks (ANNs) are widely used for solving Data Mining, intelligent control, forecasting, pattern recognition tasks, etc. under uncertainty conditions, nonlinearity, stochasticity, randomness, various types of disturbances and noise, thanks to its universal approximating abilities and learning opportunities based on experimental data characterizing functioning of the investigated object [Haykin, 1999; Du, 2014].

ANNs' learning process is usually based on the use of the criterion optimization procedure, the convergence speed of this procedure can be quite low, especially while training multi-layer networks such as a multilayer perceptron (MLP), which creates a number of problems when a training sample is fed to the system in the form of an observations' sequence in an online mode, for example, adaptive control of non-stationary objects, Web Mining etc.

To accelerate the learning process in neural networks whose output signal is linearly dependent on adjustable synaptic weights is possible, for example, using radial basis (RBFN), normalized radial basis (NRBFN), polynomial (PNN) and the GMDH-neural networks (GMDH-ANN), however, their use is often complicated by the so-called "curse of dimensionality." The issue is not only in arising computational difficulties, but the reason is the available experimental data may be not enough for estimating a large number of synaptic weights.

An alternative to the optimization-based learning is the memory-based learning [Nelles, 2001] which is associated with a concept "neurons in the data points" [Zahirniak, 1990]. The most typical representative of neural networks whose training is based on this concept are generalized regression neural network (GRNN), but they solve the problem of interpolation and not approximation which complicates greatly their use while processing "noisy" data.

A hybrid of different neural networks, whose training is based both on optimization and memory, are support vector machines (SVM) [Vapnik, 1974; Vapnik, 1979; Cortes, 1995; Vapnik, 1995]. Their architecture coincides with RBFN and GRNN, synaptic weights are determined as a result of solving a nonlinear programming problem, and activation functions' centers are set according to the concept "neurons in the data points."

Thus, this network is a network with direct information transmission, which are generalizations of such popular constructions as MLP, RBFN, GRNN, which implement an empirical risk minimization method [Vapnik, 1974; Vapnik, 1979]. They have been widely applied to solving identification, pattern recognition and neurocontrol problems [Haykin, 1999; Du, 2014]. Although SVM-networks have a number of unquestionable advantages, their training is quite time consuming from a computational point of view, as it has to do with solving nonlinear programming problems of high dimensionality.

In this regard, least squares support vector machines (LS-SVM) [Suykens, 2002] were proposed as an alternative to the ordinary SVM, whose training is reduced to solving systems of linear equations. That's much easier from a computational point of view.

Neuro-fuzzy systems (NFS) [Jang, 1997] have more features compared to neural networks with their learning capability, approximation and linguistic interpretation of the results. Here, ANFIS [Jang, 1993] and TSK-systems [Takagi, 1985] are the most widely used systems, whose output layer is adjusted with the help of linear learning algorithms. It should be mentioned that the majority of neuro-fuzzy systems is trained with the help of optimization procedures.

A fuzzy analogue of a traditional SVM is a fuzzy support vector machine (FSVM) [Lin, 2002], where multidimensional kernel activation functions are replaced with one-dimensional bell-shaped membership functions. In [Abe, 2003], a least squares fuzzy support vector machine (LS-FSVM) was introduced to solve the tasks of pattern recognition based on binary training signals.

Although FSVM has a great potential compared to a traditional SVM, a training procedure is rather cumbersome from a computational point of view due to its implementation, which naturally limits its ability to solve real-time tasks.

It is advisable to develop rather simple neuro-fuzzy systems to realize the learning idea based on the empirical risk minimization when information is processed in an online mode.

A neo-fuzzy neuron [Yamakawa, 1992; Uchino, 1997; Miki, 1999] can be used as a basic element of such systems, which is characterized by high approximating properties, its simplicity and speed learning.

A Neo-Fuzzy Neuron

A neo-fuzzy neuron (NFN) is a nonlinear system with multiple inputs and a single output having the following mapping

$$\hat{y} = \sum_{i=1}^{n} f_i(x_i)$$

where x_i is the $i-$th component of a $n-$dimensional vector of input signals $x = (x_1, \ldots, x_i, \ldots, x_n)^T \in R^n$, $\hat{y}$ is a scalar NFN output. Structural units of the neo-fuzzy neuron are nonlinear synapses NS_i which transform the $i-$th input signal in the following way

$$f_i(x_i) = \sum_{l=1}^{h} w_{li}\mu_{li}(x_i)$$

where w_{li} is the $l-$th adjustable synaptic weight of the $i-$th nonlinear synapse, $l=1,2,\ldots h-$ the total quantity of synaptic weights and, respectively, membership functions $\mu_{li}(x_i)$ in the same nonlinear synapse. In this way transformation carried out by the NFN can be written as

$$\hat{y}=\sum_{i=1}^{n}\sum_{l=1}^{h}w_{li}\mu_{li}(x_i) \qquad (1)$$

and the fuzzy inference carried out by the same NFN has a form of

IF x_i IS X_{li} THEN THE OUTPUT IS w_{li}

which means that actually a nonlinear synapse implements a fuzzy zero-order Takagi-Sugeno reasoning [Takagi, 1985].

Authors of the neo-fuzzy neuron [Yamakawa, 1992; Uchino, 1997; Miki, 1999] used traditional triangular constructions meeting the conditions of unity partitioning as membership functions:

$$\mu_{li}(x_i)=\begin{cases}\dfrac{x_i-c_{l-1,i}}{c_{li}-c_{l-1,i}},\ if\ x_i\in\left[c_{l-1,i},c_{li}\right],\\ \dfrac{c_{l+1,i}-x_i}{c_{l+1,i}-c_{li}},\ if\ x_i\in\left[c_{li},c_{l+1,i}\right],\\ 0,\ otherwise\end{cases}$$

where c_{li} are relatively arbitrarily chosen (usually evenly distributed) centers of membership functions over the interval $[0,1]$ where, naturally, $0\le x_i\le 1$.

This choice of membership functions ensures that the input signal x_i activates only two neighboring membership functions, and their sum is always equal to 1 which means that

$$\mu_{li}(x_i)+\mu_{l+1,i}(x_i)=1$$

and

$$f_i(x_i)=w_{li}\mu_{li}(x_i)+w_{l+1,i}\mu_{l+1,i}(x_i).$$

Of course, other types of membership functions (except triangular) can be used like cubic and B-splines, polynomials, harmonic and orthogonal functions, wavelets etc. It should be noticed that the NFN contains nh membership functions and the same amount of adjustable synaptic weights.

Introducing a $(nh\times 1)-$vector of membership functions

$$\mu(x(k))=\left(\mu_{11}(x_1(k)),\ldots,\mu_{h1}(x_1(k)),\mu_{12}(x_2(k)),\ldots,\mu_{li}(x_i(k)),\ldots,\mu_{hn}(x_n(k))\right)^T$$

(here $k=1,2,\ldots,N$ is a number of the vector observation $x(k)$ in a training sample or current discrete time) and a corresponding vector of NFN synaptic weights

$$w=\left(w_{11},\ldots,w_{h1},w_{12},\ldots,w_{h2},\ldots,w_{li},\ldots,w_{hn}\right)^T,$$

the transformation (1) carried out by the NFN can be rewritten in the form

$$\hat{y}(k)=w^T\mu(x(k)).$$

The NFN authors used a gradient learning procedure

$$w_{li}(k)=w_{li}(k-1)+\eta e(k)\mu_{li}(x_i(k))=w_{li}(k-1)+\eta(y(k)-\hat{y}(k))\mu_{li}(x_i(k))=$$
$$=w_{li}(k-1)+\eta(y(k)-w^T(k-1)\mu(x(k)))\mu_{li}(x_i(k))$$

where $y(k)$ is a reference signal, η is a learning rate parameter.

In [Bodyanskiy, 2003], a learning algorithm was proposed that posses both tracking (non-stationary cases) and filtering («noisy» data) properties:

$$\begin{cases} w(k)=w(k-1)+r^{-1}(k)e(k)\mu(x(k)), \\ r(k)=\alpha r(k-1)+\|\mu(x(k))\|^2, 0\le\alpha\le 1, \end{cases} \tag{2}$$

when $\alpha=0$, the algorithm (2) coincides with the optimal Kaczmarz-Widrow-Hoff learning algorithm.

Basically, to set the NFN lots of other learning algorithms and identification [Nelles, 2001; Ljung, 1999] can be used including the standard least squares method

$$w(N)=\left(\sum_{k=1}^{N}\mu(x(k))\mu^T(x(k))\right)^{-1}\sum_{k=1}^{N}\mu(x(k))y(k) \tag{3}$$

and also his recurrent and exponentially-weighted versions.

The NFN training based on the empirical risk minimization

Training the NFN with the help of the least squares support vector machine approach (LS-SVM-NFN) leads to the quadratic criterion optimization

$$E(N)=\frac{1}{2}\|w\|^2+\frac{\gamma}{2}\sum_{k=1}^{N}e^2(k) \tag{4}$$

within the constraints as a system of N linear equations

$$y(k)=w^T\mu(x(k))+e(k) \tag{5}$$

where $\gamma>0$ is a regularization parameter (a momentum term).

The criterion optimization (4) without the constraints (5) leads to the expression

$$w(N)=\left(\sum_{k=1}^{N}\mu(x(k))\mu^T(x(k))+\gamma^{-1}I\right)^{-1}\sum_{k=1}^{N}\mu(x(k))y(k)$$

which is rather close to (3) and which is essentially a ridge estimator, where $I-(nh\times nh)$ is an identity matrix.

Let's introduce a Lagrange function to take into account the constraints' system (5)

$$L(w,e(k),\lambda(k))=E(k)+\sum_{k=1}^{N}\lambda(k)(y(k)-w^T\mu(x(k))-e(k))=$$
$$=\frac{1}{2}w^Tw+\frac{\gamma}{2}\sum_{k=1}^{N}e^2(k)+\sum_{k=1}^{N}\lambda(k)(y(k)-w^T\mu(x(k))-e(k))$$

(here $\lambda(k)$ stands for N undetermined Lagrange multipliers) and the Karush-Kuhn-Tucker system of equations

$$\begin{cases} \nabla_w L\left(w,e(k),\lambda(k)\right) = w - \sum_{k=1}^{N}\lambda(k)\mu\left(x(k)\right) = \vec{0}_N, \\ \dfrac{\partial L\left(w,e(k),\lambda(k)\right)}{\partial e(k)} = \gamma e(k) - \lambda(k) = 0, \\ \dfrac{\partial L\left(w,e(k),\lambda(k)\right)}{\partial \lambda(k)} = y(k) - w^T\mu\left(x(k)\right) - e(k) = 0 \end{cases} \tag{6}$$

where $\vec{0}_N - (N\times 1)$ is a vector formed with zeros.

The solution of the equation system (6) is:

$$\begin{cases} w(N) = \sum_{k=1}^{N}\lambda(k)\mu\left(x(k)\right), \\ \lambda(k) = \gamma e(k), \\ y(k) = w^T(N)\mu\left(x(k)\right) + e(k) \end{cases} \tag{7}$$

or in a matrix form

$$\left(\gamma^{-1}I_{NN} + \Omega_{NN}\right)\begin{pmatrix}\lambda(1)\\ \vdots \\ \lambda(N)\end{pmatrix} = \begin{pmatrix} y(1)\\ \vdots \\ y(N)\end{pmatrix}$$

(here $I_{NN} - (N\times N)$ is an identity matrix)

or

$$\left(\gamma^{-1}I_{NN} + \Omega_{NN}\right)\Lambda_N = Y_N$$

(here $\Omega_{NN} = \left\{\Omega_{pq} = \mu^T\left(x(p)\right)\mu\left(x(q)\right)\right\}, p = 1,2,\ldots,N; q = 1,2,\ldots,N$), whence

$$\Lambda_N = \left(\gamma^{-1}I_{NN} + \Omega_{NN}\right)^{-1}Y_N. \tag{8}$$

Then an output NFN signal

$$\hat{y}(x) = w^T(N)\mu(x)$$

for an arbitrary input signal x taking into account (7), (8) can be written in the form

$$\hat{y}(x) = \left(\sum_{k=1}^{N}\lambda(k)\mu\left(x(k)\right)\right)^T\mu(x). \tag{9}$$

If the processing data are consecutively supplied, the training process of the LS-SVM-NFN should be fulfilled in an online mode. Thus when a pair of $x(N+1), y(N+1)$ comes to the system, the expression (9) takes the form

$$\hat{y}(x) = \left(\sum_{k=1}^{N} \lambda(k)\mu(x(k)) + \lambda(N+1)\mu(x(N+1)) \right)^{T} \mu(x)$$

or in a matrix form

$$\left(\gamma^{-1} I_{N+1,N+1} + \Omega_{N+1,N+1}\right) \left(\begin{array}{c} \lambda(1) \\ \vdots \\ \lambda(N) \\ \hline \lambda(N+1) \end{array} \right) = \left(\begin{array}{c} y(1) \\ \vdots \\ y(N) \\ \hline y(N+1) \end{array} \right)$$

or

$$\left(\begin{array}{c|c} \Omega_{NN} & \omega_{N+1} \\ \hline \omega_{N+1}^{T} & \gamma^{-1} \end{array} \right) \left(\begin{array}{c} \Lambda_{N} \\ \hline \lambda(N+1) \end{array} \right) = \left(\begin{array}{c} Y_{N} \\ \hline y(N+1) \end{array} \right) \qquad (10)$$

where $\omega_{N+1} = \left(\mu^{T}(x(1))\mu(x(N+1)), \mu^{T}(x(2))\mu(x(N+1)), \ldots, \mu^{T}(x(N))\mu(x(N+1)) \right)^{T}$.

It comes from the expression (10) that

$$\Lambda_{N+1} = \left(\begin{array}{c} \Lambda_{N} \\ \hline \lambda(N+1) \end{array} \right) = \left(\begin{array}{c|c} \Omega_{NN} & \omega_{N+1} \\ \hline \omega_{N+1}^{T} & \gamma^{-1} \end{array} \right)^{-1} \left(\begin{array}{c} Y_{N} \\ \hline y(N+1) \end{array} \right). \qquad (11)$$

Using the Frobenius formula in the form of [Gantmacher, 2000]

$$M = \left(\begin{array}{c|c} A & B \\ \hline C & D \end{array} \right), \quad |D| \neq 0,$$

$$M^{-1} = \left(\begin{array}{c|c} A & B \\ \hline C & D \end{array} \right)^{-1} = \left(\begin{array}{c|c} K^{-1} & -K^{-1}BD^{-1} \\ \hline -D^{-1}CK^{-1} & D^{-1} + D^{-1}CK^{-1}BD^{-1} \end{array} \right),$$

$$K = A - BD^{-1}C$$

where taking into consideration (11)

$$K = \Omega_{NN} - \omega_{N+1}\gamma\omega_{N+1}^{T},\ K^{-1} = \left(\Omega_{NN} - \gamma\omega_{N+1}\omega_{N+1}^{T} \right)^{-1}$$

one can easily calculate the $(N+1)-$ th Lagrange multiplier with the help of the expression

$$\lambda(N+1) = -\gamma\omega_{N+1}^{T}K^{-1}Y_{N} + \gamma\left(1 + \gamma\omega_{N+1}^{T}K^{-1}\omega_{N+1}\right)y(N+1).$$

Then using the Sherman-Morrison formula of matrices inversion [Gantmacher, 2000], we finally get

$$\begin{cases} K^{-1} = \Omega_{NN}^{-1} + \dfrac{\Omega_{NN}^{-1}\omega_{N+1}\omega_{N+1}^{T}\Omega_{NN}^{-1}}{1 - \omega_{N+1}^{T}\Omega_{NN}^{-1}\omega_{N+1}}, \\ \lambda(N+1) = 1 + \gamma\omega_{N+1}^{T}K^{-1}\left(\omega_{N+1} - Y_{N}\right). \end{cases}$$

Conclusion

The paper presents a fuzzy least squares support vector machine (LS-FSVM) which is implemented with the help of neo-fuzzy neurons (NFN) and which is essentially a zero order Takagi-Sugeno fuzzy inference system. The proposed LS-FSVM-NFN is numerically simple because it's generated with NFNs, it also has a small number of adjustable parameters and high speed associated with the possibility of applying the second order optimization learning procedures to process data in an online-mode.

Bibliography

[Abe, 2003] S. Abe, D. Tsujinishi. Fuzzy Least Squares Support Vector Machines for multiclass problems. Neural Networks, 2003, №16, P. 785-792.

[Bodyanskiy, 2003] Ye. Bodyanskiy, I. Kokshenev, V. Kolodyazhniy. An adaptive learning algorithm for a neo-fuzzy neuron. Proc. 3rd Int. Conf. of European Union Soc. for Fuzzy Logic and Technology (EUSFLAT 2003), Zittau, Germany, 2003, P. 375-379.

[Cortes, 1995] C. Cortes, V. Vapnik. Support vector networks. Machine Learning, 1995, №20, P. 273-297.

[Du, 2014] K.-L. Du, M.N.S. Swamy. Neural Networks and Statistical Learning, London: Springer-Verlag, 2014, 816p.

[Gantmacher, 2000] F.R. Gantmacher. The Theory of Matrices, AMS Chelsea Publishing: Reprinted by American Mathematical Society, 2000, 660p.

[Haykin, 1999] S. Haykin. Neural Networks. A Comprehensive Foundation, Upper Saddle River, N.J.: Prentice Hall, 1999, 842 p.

[Jang, 1993] J.-S. R. Jang. ANFIS: Adaptive-network-based fuzzy inference systems. IEEE Trans. Syst., Man., and Cybern., 1993, № 23, P. 665-685.

[Jang, 1997] J.-S. R. Jang, C. T. Sun, E. Mizutani. Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence, N.J.: Prentice Hall, 1997, 640 p.

[Lin, 2002] Ch.-F. Lin, Sh.-D. Wang. Fuzzy Support Vector Machines. IEEE Trans. on Neural Networks, 2002, №13, P. 646-671.

[Ljung, 1999] L. Ljung. System Identification: Theory for the User, N.Y.: Prentice-Hall, 1999, 519p.

[Miki, 1999] T. Miki, T. Yamakawa. Analog implementation of neo-fuzzy neuron and its on-board learning. Computational Intelligence and Applications, ed. by N. E. Mastorakis, Piraeus: WSES Press, 1999, P. 144-149.

[Nelles, 2001] O. Nelles. Nonlinear System Identification, Berlin: Springer, 2001, 785p.

[Suykens, 2002] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle. Least Squares Support Vector Machines, Singapore: World Scientific, 2002, 294p.

[Takagi, 1985] T. Takagi, M. Sugeno. Fuzzy identification of systems and its application to modelling and control. IEEE Trans. Syst., Man., and Cybern., 1985, №15, P. 116-132.

[Uchino, 1997] E. Uchino, T. Yamakawa. Soft computing based signal prediction, restoration and filtering. Intelligent Hybrid Systems: Fuzzy Logic, Neural Networks and Genetic Algorithms, ed. by Da Ruan, Boston: Kluwer Academic Publisher, 1997, P. 331-349.

[Vapnik, 1974] V.N. Vapnik, A.Ya. Chervonenkis. Pattern Recognition Theory (statistical learning problems), M.: Nauka, 1974, 416p. (in Russian)

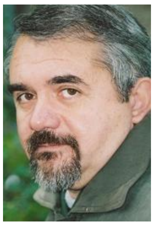
[Vapnik, 1979] V.N. Vapnik, A.Ya. Chervonenkis. Empirical data dependency restoration, M.: Nauka, 1979, 448p. (in Russian)

[Vapnik, 1995] V.N. Vapnik. The Nature of Statistical Learning Theory, N.Y.: Springer, 1995, 188p.

[Yamakawa, 1992] T. Yamakawa, E. Uchino, T. Miki, H. Kusanagi. A neo fuzzy neuron and its applications to system identification and prediction of the system behavior. Proc. 2nd Int. Conf. on Fuzzy Logic and Neural Networks "IIZUKA-92", Iizuka, Japan, 1992, P. 477-483.

[Zahirniak, 1990] D. Zahirniak, R. Chapman, S.K. Rogers, B.W. Suter, M. Kabrisky, V. Pyati. Pattern recognition using radial basis function network. Application of Artificial Intelligence Conf., Dayton, OH, 1990, P. 249-260.

Authors' Information



Yevgeniy Bodyanskiy – Professor, Dr. – Ing. habil., The Head of Control Systems Research Laboratory, Kharkiv National University of Radio Electronics, 14 Lenin Ave., Office 511, 61166 Kharkiv, Ukraine; e-mail: bodya@kture.kharkov.ua

Major Fields of Scientific Research: Artificial neural networks, Fuzzy systems, Hybrid systems of computational intelligence, Cascade neuro-fuzzy systems

Oleksii Tyshchenko - Ph.D., Senior Researcher at Control Systems Research Laboratory, Kharkiv National University of Radio Electronics, 14 Lenin Ave., Kharkiv, 61166, Ukraine; e-mail: lehatish@gmail.com

Major Fields of Scientific Research: Artificial neural networks, Hybrid systems of computational intelligence, Reservoir Computing, Cascade neuro-fuzzy systems

Daria Kopaliani – PhD student, Kharkiv National University of Radio Electronics, 14 Lenin Ave., Office 517, 61166 Kharkiv, Ukraine; e-mail: daria.kopaliani@gmail.com

Major Fields of Scientific Research: Artificial neural networks, Cascade neuro-fuzzy systems

MULTILAYER NEURO-FUZZY SYSTEM FOR SOLVING ON-LINE DIAGNOSTICS TASKS

Yevgeniy Bodyanskiy, Olena Vynokurova, Iryna Pliss, Dmytro Peleshko

***Abstract**: In the paper the problem of on-line diagnostics and properties change detection of systems whose output signal is multidimensional non-stationary stochastic sequence is considered. The six-layer diagnostic neuro-fuzzy system is proposed. The first layer of this system consists of membership functions blocks, the second layer provides aggregation of fuzzyfied inputs, the third one consists of tuning synaptic weights, the fourth one consists of summing blocks set, the fifth layer produces normalization (defuzzification) of output signals and finally sixth layer consists of nonlinear activation functions and provides the properties change detection. So the proposed neuro-fuzzy system is modification of L.Wang-J.Mendel system and solves diagnostics-classification problems in real time mode. For tuning synaptic weights of proposed system we have used special learning criterion, that is aimed at solving of pattern recognition, classification, diagnostics problems etc. The learning algorithm for synaptic weights is proposed and its speed optimization is performed. It allowed to design recurrent procedure, which is matrix hybrid of J.Shynk and S. Kaczmarz-B.Widrow-M.Hoff learning algorithm. It is important to notice that proposed system has significantly fewer number of tuning synaptic weights in comparison with conventional well-known neural network diagnosis systems based on multilayer perceptrons or radial basis function networks. This feature allows to reduce the learning set volume, to achieve optimal training rate, to provide linguistic interpretability and «transparency» of obtained results.*

***Keywords**: neuro-fuzzy-system, adaptive learning, data mining, diagnostic*

***ACM Classification Keywords**: I.2.6 Learning – Connectionism and neural nets.*

Introduction

For solving wide class problems of Data Mining which connected first of all with diagnostic, classification, pattern recognition etc the artificial neural networks are used increasingly frequently due to their universal approximation properties and learning ability based on experimental data set. Although for solving such problem the conventional multilayer perceptron is used in most cases, it should not go unnoticed such its common disadvantages as sufficiently large training set volume, low convergence rate of backpropagation learning algorithm, the necessity for using a large number of training epoch. And if especially computational problems we can solve, but necessity for a representative training sample significantly complicates the use of this neural network for solving many practical problems. This problem appears especially in the research where data set has short dimension and at that the object is described by set of different characteristics [Swamy, 2014, Kurse, 2013].

In this situation radial basis function neural networks are preferable [Haykin, 1999], whose output signal is linearly dependent on tuning synaptic weights. This fact allows to use for training these networks large range of well-known approaches from conventional least squares method to the popular linear adaptive identification algorithms [Ljung, 1999]. And although specificity of diagnosis-classification problems restricts to use conventional square learning criterion, using special Shynk criterion [Shynk, 1990], focused on pattern recognition tasks with binary training signal allows to design sufficient simple and effective diagnostic radial basis function neural network [Bodyanskiy, 2002, Bodyanskiy, 2005].

In spite of all its advantages radial basis neural networks is not panacea for all cases because of its possibilities are limited by so called "curse of dimensionality" that leads to the exponential increasing number of tuning synaptic weights in accordance with input signal-pattern dimension space.

Overcome this problem, the procedure of preliminary setting of radial basis function centers by one or another clustering methods is used. Hence, supervised learning is completed by self-learning of its centers, what makes such learning too tedious.

In [Bodyanskiy, 2010] for solving tasks of text documents processing in the context of Text Mining, which is characterized by large dimensionality of input signals, hierarchical radial basis function network with multilayer architecture was proposed. Such system uses usual RBFN in each unit and on the input of systems only part of features is fed, what allows to overcome problem of "course dimensionality". The main thing that has been achieved in this situation - it is the possibility to operate under conditions when input pattern dimension is comparable with training set volume. At the same time, it should not go unnoticed inconvenience of this system, impossibility to operate in sequential on-line mode, high level of subjectivity in partition of input pattern into subvector-vector set for each network unit.

Anyway the problem of processing data with high dimension of features vector under condition, when training set volume is comparable with this dimension, is attractive and especially for solving tasks of classification and diagnostic in Text Mining, Web Mining and medical-biology applications. Using of neuro-fuzzy systems (NFS) [Jang, 1997] is significantly future-oriented because such systems allow to provide not only good approximation properties and learning ability, but linguistic interpretability of obtained results. It is also necessary to note that obtained results of NFS-systems are equivalent to the results of radial basis function networks [Jang, 1993], this fact allows to use identical learning algorithms.

Thus this paper is devoted to synthesis of diagnostic neuro-fuzzy system for the case, when training set dimension is comparable with input patterns set volume, and these patterns are fed for processing in on-line mode.

Diagnostic neuro-fuzzy system architecture

Architecture of considered NFS is shown on Fig. 1 and consists of six sequentially-connected layers. In the input (null, receptive) layer of NFS $(n \times 1)$-dimensional vector of input signals-patterns $x(k) = (x_1(k), x_2(k), \dots x_n(k))^T$ is fed, where $k = 1, 2, \dots, N$ is observation number in initial data set. In this case it is supposed that all components $x_i(k)$ preliminary are modified so that

$$0 \le x_i(\mathrm{k}) \le 1, \forall i = 1, 2, \dots, n,$$

and binary input features have value 0 or 1.

The first layer consists of nh membership function $\mu_{li}(x_i(k))$, $i = 1, 2, \dots, n$; $l = 1, 2, \dots, h$ and provides fuzzyfication of input variables, at that the larger the number h, the better approximating properties NFS, although it is enough to have $h = 2$ for binary features.

The second hidden layer realizes aggregation of membership levels, which are computed in the first layer, and consists of h multiplication units Π.

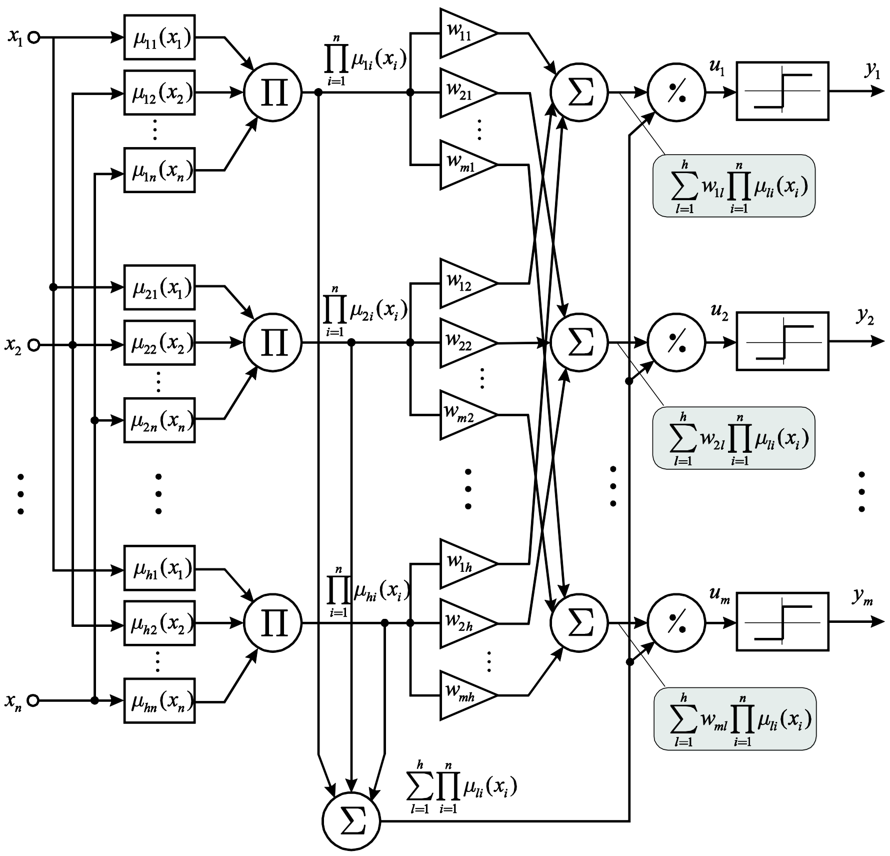

x_1
x_2
x_n
$\mu_{11}(x_1)$
$\mu_{12}(x_2)$
$\mu_{1n}(x_n)$
$\mu_{21}(x_1)$
$\mu_{22}(x_2)$
$\mu_{2n}(x_n)$
$\mu_{h1}(x_1)$
$\mu_{h2}(x_2)$
$\mu_{hn}(x_n)$
Π
$\prod_{i=1}^{n}\mu_{1i}(x_i)$
$\prod_{i=1}^{n}\mu_{2i}(x_i)$
$\prod_{i=1}^{n}\mu_{hi}(x_i)$
w_{11}
w_{21}
w_{m1}
w_{12}
w_{22}
w_{m2}
w_{1h}
w_{2h}
w_{mh}
Σ
$\sum_{l=1}^{h}\prod_{i=1}^{n}\mu_{li}(x_i)$
%
u_1
u_2
u_m
y_1
y_2
y_m
$\sum_{l=1}^{h}w_{1l}\prod_{i=1}^{n}\mu_{li}(x_i)$
$\sum_{l=1}^{h}w_{2l}\prod_{i=1}^{n}\mu_{li}(x_i)$
$\sum_{l=1}^{h}w_{ml}\prod_{i=1}^{n}\mu_{li}(x_i)$

Fig. 1. Diagnostic neuro-fuzzy system

The third hidden layer is one of synaptic weights w_{jl}, $j=1,2,\ldots,m$ which are adjusted during learning process. Proposed NFS consists of mh tuning weights, where m is a number of potential classes, one for each system output. It is clear that $mh << e^n$, i.e. number of NFS weights are significantly smaller than the number of RBFN weights.

The fourth hidden layer consists of $m+1$ summators Σ, which compute sum of output signal of the second and the third hidden layers.

In fifth hidden layer that consists of m division unit □/□ normalization of fourth layer output signals is realized.

And finally output (sixth) layer consists of m non-linear activation functions, at that in diagnosis tasks it is reasonable to use the simplest signum-functions, which takes $+1$ value in case of right diagnosis, and -1 – otherwise. Therefore output system signals $y_j(k)$ can take only two values ± 1.

Thus if vector signal $x(k)$ is fed on NFS input, the first layer elements compute membership levels $\mu_{li}(x_i(k))$, at that usually the bell-shaped (kernel) construction with as membership function nonstrictly local receptive field are used as membership functions. It allows to avoid appearing of “gaps” in fuzzyficated space [Friedman, 2003]. Most often it is conventional Gaussians.

$$\mu_{li}(x_i(k)) = \exp\left(-\frac{(x_i(k)-c_{li})^2}{2\sigma_i^2}\right) \quad (1)$$

where c_{li} is center parameter (in the simplest case the centers are located uniformly in the interval [0,1] with step $(h-1)^{-1}$), σ_i is width parameter, selected empirically or tuning with backpropagation algorithm [Osowski, 2006]. Fig. 2 shows membership functions.

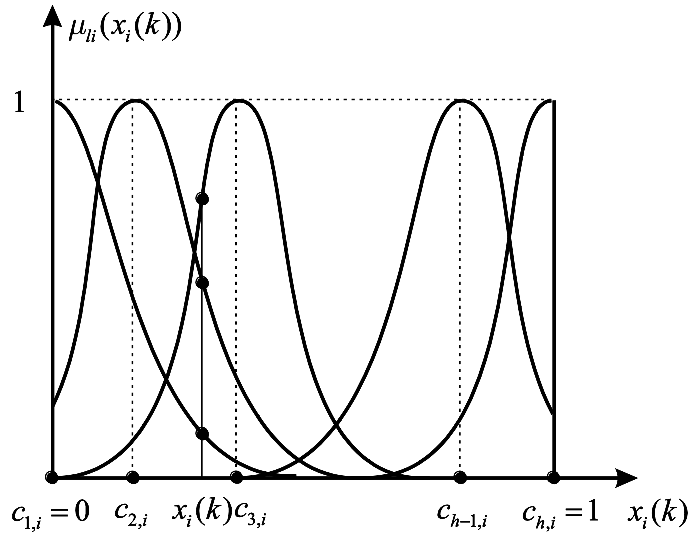


Fig. 2 – Bell-shaped membership functions

It is clear that for binary variables $x_i(k)$ it is enough to use only two triangular membership functions

$$\begin{cases}\mu_{1i}(x_i(k)) = 1 - x_i(k),\\ \mu_{2i}(x_i(k)) = x_i(k),\end{cases} \tag{2}$$

that are shown on Fig. 3.

We also have to notice that membership functions (2) in some cases with success can be used for features which have arbitrary number of values (see fig. 3), and number of synaptic weights take on minimally possible value $2m$.

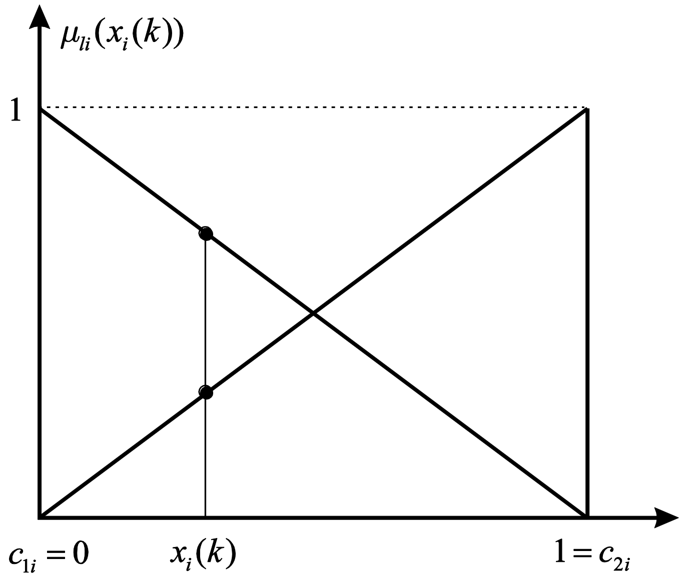


Fig. 3 – Membership functions for binary variables

On the outputs of second layer the aggregated values $\prod_{i=1}^{n}\mu_{li}(x_i(k))$ are appeared, at that it is simple to notice that if width parameters σ_i are the same for all features, i.e. $\sigma_i = \sigma$, that

$$\prod_{i=1}^{n}\mu_{li}(x_i(k)) = \prod_{i=1}^{n}\exp\left(-\frac{(x_i(k)-c_{li})^2}{2\sigma^2}\right) = \exp\left(-\frac{\|x_i(k)-c_{li}\|^2}{2\sigma^2}\right)$$

(here $c_l = (c_{l_1}, c_{l_2}, \ldots, c_{l_n})^T$) i.e. nonlinear transformation similar RBFN is realized.

Outputs of third hidden layer are values $w_{jl}\prod_{i=1}^{n}\mu_{li}(x_i(k))$, forth one $\sum_{l=1}^{h}w_{jl}\prod_{i=1}^{n}\mu_{li}(x_i(k))$ and $\sum_{l=1}^{h}\prod_{i=1}^{n}\mu_{li}(x_i(k))$, fifth one

$$u_j(k) = \frac{\sum_{l=1}^{h}w_{jl}\prod_{i=1}^{n}\mu_{li}(x_i(k))}{\sum_{l=1}^{h}\prod_{i=1}^{n}\mu_{li}(x_i(k))} = \sum_{l=1}^{h}w_{jl}\frac{\prod_{i=1}^{n}\mu_{li}(x_i(k))}{\sum_{l=1}^{h}\prod_{i=1}^{n}\mu_{li}(x_i(k))} =$$
$$= \sum_{l=1}^{h}w_{jl}\varphi_l(x(k)) = w_j^T\varphi(x(k))$$

(here $\varphi_l(x(k)) = \prod_{i=1}^{n}\mu_{li}(x_i(k))\left(\sum_{l=1}^{h}\prod_{i=1}^{n}\mu_{li}(x_i(k))\right)^{-1}$, $w_j = (w_{j1}, w_{j2}, \ldots, w_{jh})^T$, $\varphi(x(k)) = (\varphi_1(x(k)), \varphi_2(x(k)), \ldots, \varphi_h(x(k)))^T$) and, finally, sixth

$$y_j(k) = \operatorname{sign} u_j(k)$$

It is clearly to see that proposed NFS is modification of Wang-Mendel system [Wang, 1992, Wang, 1994], which oriented for solving on-line diagnostic-classification tasks.

Diagnostic neuro-fuzzy system learning

For training of synaptic weights on system under consideration we use learning algorithm based on specialized criterion [Shynk, 1990], which is aimed for solving pattern recognition, classification, diagnostic tasks etc.

Let us introduce m errors of learning

$$e_j(k) = d_j(k) - y_j(k) = d_j(k) - \operatorname{sign} u_j(k)$$

and m criterions based on these errors

$$\begin{aligned} E_j(k) &= e_j(k)u_j(k) = d_j(k)u_j(k) - |u_j(k)| = \\ &= \left(d_j(k) - \operatorname{sign} w_j^T\varphi(x(k))\right)\cdot w_j^T\varphi(x(k)), \end{aligned} \tag{3}$$

where $d_j(k) \in \{-1,1\}$ is training signal, having value 1, if input vector $x(k)$ belongs to j –th diagnosis, and -1 otherwise.

For synaptic weights tuning we can use conventional gradient procedure of criterion minimization (3)

$$w_{jl}(k+1) = w_{jl}(k) - \eta(k)\frac{\partial E_j(k)}{\partial w_{jl}}$$

(here $\eta(k)$ is learning rate parameter), which on vector form can be rewritten in the form

$$\begin{aligned} & w_j(k+1) = w_j(k) + \eta(k)e_j(k)\varphi(x(k)) = \\ & = w_j(k) + \eta(k)\left(d_j(k) - \text{sign } w_j^T(k)\varphi(x(k))\right)\cdot\varphi(x(k)), \\ & j = 1,2,\ldots,m. \end{aligned} \tag{4}$$

Introducing further general criterion for all system outputs

$$E(k) = \sum_{j=1}^{m} E_j(k) = \sum_{j=1}^{m} e_j(k)u_j(k),$$

we can write learning algorithm of all system synaptic weights in form

$$W(k+1) = W(k) + \eta(k)\left(d(k) - \text{sign } W(k)\varphi(x(k))\right)\cdot\varphi^T(x(k)), \tag{5}$$

where $\text{sign}(u_1(k), u_2(k), \ldots, u_m(k))^T = (\text{sign}u_1(k), \text{sign}u_2(k), \ldots, \text{sign}u_m(k))^T$, $d(k) = (d_1(k), d_2(k), \ldots, d_m(k))^T$,

$W(k) = \begin{pmatrix} w_1^T(k) \\ w_2^T(k) \\ \vdots \\ w_m^T(k) \end{pmatrix} - (m \times h)$ is matrix of tuning synaptic weights.

It is known that gradient algorithms (3)-(5) provide the convergence in enough wide range of variation of learning rate parameter $\eta(k)$ [Derevitskiy, 1981], however at that convergence rate can be nonsufficient. Increasing of learning rate we can use quasi-Newton learning algorithms [Shepherd, 1997], for example,

$$w_j(k+1) = w_j(k) + \left(\varphi(x(k))\varphi^T(x(k)) + \eta I\right)^{-1} e_j(k)\varphi(x(k)), \tag{6}$$

where $\eta > 0$ is momentum term, $I - (\text{h}\times\text{h})$ is unity matrix.

Using lemma of matrix inversion we can show that [7]

$$\left(\varphi(x(k))\varphi^T(x(k)) + \eta I\right)^{-1}\varphi(x(k)) = \frac{\varphi(x(k))}{\eta + \|\varphi(x(k))\|^2},$$

and rewrite (6) in compact form

$$w_j(k+1) = w_j(k) + \frac{e_j(k)\varphi(x(k))}{\eta + \|\varphi(x(k))\|^2}, \tag{7}$$

or

$$W(k+1) = W(k) + \frac{d(k) - signW(k)\varphi(x(k))}{\eta + \|\varphi(x(k))\|^2}\varphi^T(x(k)), \tag{8}$$

for $\eta = 0$ this algorithm is multidimensional modification of optimal algorithm, introduced in [Tsypkin, 1984].

Conclusions

The diagnostic neuro-fuzzy system and its adaptive learning algorithm are introduced for solving pattern recognition, classification, diagnostics tasks etc under condition when training set value is comparable with input patterns dimension, and these patterns are fed for processing in on-line mode. The feature of proposed systems is significant smaller number of tuning parameters in comparison with the artificial neural networks, which solve the same task.

The system is characterized by simplicity of computational implementation, high speed of learning process, possibility of processing information, which is described in different scales (interval, ordinal, binary).

Bibliography

[Bodyanskiy, 2002] Ye. Bodyanskiy, Ye. Kucherenko, O. Chaplanov Diagnostic and prediction of time series using multilayer radial-basis neural network. Proc. 8 Russian conf. with internal. participation “Neurocomputers and its Applying”, Moskow, 2002, P. 209-213 (in Russian).

[Bodyanskiy, 2005] Ye. Bodyanskiy, Ye. Kucherenko, O. Mikhalev Petri Neuro-Fuzzy Networks in Modelling Tasks of Complex Systems, Dnepropetrivsk: Systemni Technologii, 2005, 311 p. (in Russian).

[Bodyanskiy, 2010] Ye. Bodyanskiy, O. Shubkina Semantic annotation of text documents based on hierarchical radial basis function network. Eastern-European Journal of Enterprise Technologies, 2010, 9(90), P. 70-74 (in Russian).

[Derevitskiy, 1981] D.P. Derevitskiy, A.L. Fradkov Applied Discrete Adaptive Control System Theory. M. Nauka, 1981, 216 p.

[Friedman, 2003] J. Friedman, T. Hastie, R. Tibshirani. The Elements of Statistical Learning. Data Mining, Inference and Prediction, Berlin: Springer, 2003, 552 p.

[Haykin, 1999] S. Haykin Neural Networks. A Comprehensive Foundation. Upper Saddle River, NJ: Prentice Hall, 1999, 842 p.

[Jang, 1997] J.-S.R. Jang, C.-T. Sun, E. Mizutani Neuro-Fuzzy and Soft Computing. Prentice Hall, Upper Saddle River, NJ, 1997, 640 p.

[Jang, 1993] J.S.R. Jang, C.T. Sun Functional equivalence between radial basis function networks and fuzzy inference systems. IEEE Trans. on Neural Networks, 1993, 4, P.156-159.

[Kurse, 2013] R. Kruse, C. Borgelt, F. Klawonn, C. Moewes, M. Steinbrecher, P. Held Computational Intelligence. A Methodological Introduction, Springer, 2013, 488 p.

[Ljung, 1999] L. Ljung System Identification: Theory for the User. PTR Prentice Hall, Upper Saddle River, N.J., 1999, 672 p.

[Osowski, 2006] S. Osowski Sieci neuronowe do przetwarzania informacji. Oficyna Wydawnicza PW, Warszawa, 2006.

[Shepherd, 1997] A.J. Shepherd Second-Order Methods for Neural Networks. London: Springer-Verlag, 1997, 145 p.

[Shynk, 1990] J.J. Shynk Performance surfaces of a single-layer perceptron. IEEE Trans. on Neural Networks, 1990, 1, P. 268-274.

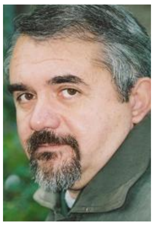
[Swamy, 2014] Ke-Lin Du, M.N.S. Swamy Neural Networks and Statistical Learning, Springer-Verlag London, 2014. - 824 p.

[Tsypkin, 1984] Ya.Z. Tsypkin Foundation of learning systems theory. M. Nauka, 1984, 320 p.

[Wang, 1992] L.X. Wang, J.M. Mendel Fuzzy basis functions, universal approximation, and orthogonal least squares learning. IEEE Trans. on Neural Network, 1992, 3, P. 807-814.

[Wang, 1994] L.-X. Wang Adaptive Fuzzy Systems and Control: Design and Stability Analysis. New Jersey: Prentice Hall, 1994, 256 p.

Authors' Information



Bodyanskiy Yevgeniy - Doctor of Technical Sciences, Professor of Artificial Intelligence Department and Scientific Head of the Control Systems Research Laboratory, Kharkiv National University of Radio Electronic, Lenina av. 14, Kharkiv, Ukraine 61166, Tel +380577021890, bodya@kture.kharkov.ua

Vynokurova Olena - Doctor of Technical Sciences, Leading Researcher of the Control Systems Research Laboratory, Kharkiv National University of Radio Electronic, Lenina av. 14, Kharkiv, Ukraine, 61166, Tel. +380577021890, vinokurova@kture.kharkov.ua

Pliss Iryna - Candidate of Technical Sciences (Ph.D.), Leading Researcher of the Control Systems Research Laboratory, Kharkiv National University of Radio Electronic, Lenina av. 14, Kharkiv, Ukraine, 61166, Tel. +380577021890, pliss@kture.kharkov.ua

Peleshko Dmytro - Doctor of Technical Sciences, Professor of Department of Information Technology Publishing, National University “Lviv Polytechnic”, Stepan Bandera 28, buld. 5, Lviv, Ukraine, 79013, Tel. +38032258-27-79, dpeleshko@gmail.com

Knowledge Discovery and Data Mining Models

MODELING EDUCATIONAL PROCESSES IN MODERN SOCIETY BY NAVIGATING MULTIDIMENSIONAL NETWORKS

Sergey Maruev, Eugene Levner, Dmitry Stefanovskyi, Alexander Troussov

***Abstract**: To graduate from a university and receive a diploma the student must follow curricula, have good command of certain topics, pass certain tests and exams. All the above mentioned artifacts of educational processes could be viewed as nodes in a large network where nodes of various kinds are connected by typed arcs, indicating, for instance, that the knowledge of a particular book or a research paper is required in a particular item of a particular curricula, or that before enrolling for a particular examination one needs to pass through particular tests. In this paradigm the process of education becomes the navigation from the initial nodes corresponding to the student knowledge and qualifications to the nodes which represent her goals. For some students the goal could be just one node representing diploma, for other students, especially for self-motivated life-long learners, the goal is a set of nodes.*

In this paper we present the initial results in modeling educational process as the navigation in multidimensional networks and the pertaining algorithms of optimization of that navigation. The results of our research could be useful for the building of educational resources (for instance, by finding structural weakness in existing networks), as well as for the personalization of the education.

The practical importance of our research stems from the processes of globalization, personalization of education, and from the explosive growth of the availability of good quality on-line training courses. Multidimensional networks of educational processes in modern are huge, naïve (common-sense) navigation tools are not sufficient for their analysis, and new computer-based navigation tools are to be designed.

***Keywords**: big data, graph-based methods, education.*

***ACM Classification Keywords**: Algorithms, Economics, Education, Experimentation, Theory.*

Introduction

We live in the interconnected space of socio-technical systems, where layers of technological infrastructures interact with the social context, which drives their everyday use and development. Most of the content is generated in public systems like LinkedIn, Facebook, Delicious, Twitter, blogs and microblogging systems, as well as in the social software used in the enterprises. These socio-technological systems already transformed computer information systems in all domains of human activities: document collections became a highly interconnected socio-semantic space, where documents are shared, discussed and edited collaboratively, and are filtered following the interests of individual users and social groups.

This process is now penetrating the modern educational systems by the explosion of corresponding internet resources and on-line platform for the distribution. Massive Open Online Courses (MOOCs), developed in universities, corporation and analytical centers are capable to compete with traditional methods of education. Such decentralized big data approaches influence methods of education; they allow individualization of education, building different trajectories of education. The problem of navigating and choosing the educational trajectory becomes actual. The trajectory must depend on the background, interests and expectations of students, on the availability of resources and various features of students and educational resources.

Resources, concepts, documents, individuals and organizations form multidimensional network, where nodes represent actors, concepts and other artifacts; links are also typed and weighted [Troussov et al. 2011]. Trajectories of education form paths in this network. This representation of modern types of education allows formalization and use of powerful abstraction provided by graph-based methods to optimize the trajectories according to given criteria.

The title of this paper has two notions characterizing the object of our investigation: Educational Processes in Modern Society. Both components are important and interdependent. In modern society many connections and relations function on the layers of technological structures and software. Interactions between agents become different, and these new types of interactions in their turn drive the future development of the technology. Educational processes evolve in the technological media, the trajectories becomes individual, though might be influenced by decisions taken by other students.

Wide spread of internet technologies and educational tools make actual the task of choosing the on-line objects relevant to educational curricula. Traditional methods for the solution of this problem, based on the titles and metainformation of the resources, do not allow discriminate, for instance, between textbooks using different approach, different conceptualization and granularity of topics.

Modern age information technologies drastically changed forms of communications, information retrieval and management, as well as social interactions between people. During last decade, these innovation affected the education. Explosive emergence of massive open on-line courses (MOOCs) demonstrates one of the trend inherent in modern education. More than 4,5 millions of students across the globe use on-line courses EdX, Udacity and Coursera [Carr, 2012]

Availability of huge amount of ready to use educational data leads to a more disruptive and far-reaching changes related to the notion of “big data”. According to [Guthrie, 2013], universities can customize courses and learning modules for student’s needs. At the same time, phenomenon of big data poses the question of how to use these opportunities in the big data environment.

Individualization of education becomes the key success factor in modern education [Robinson,2010]. The student is given the opportunity to choose between so many courses, learning materials and tutors according to their own interests. “An embarrassment of riches” - overabundance of new good opportunities - generates new troubles for students in navigating and building educational trajectories.

Enterprises and recruiting agencies now also are interested to detect how the qualification obtained by the individualized education corresponds to requirements. They need tools to evaluate and compare various courses and modules.

Learning trajectory is a complicated concept using in many domains. Educators and cognitive psychologists define learning trajectory as a sequence of knowledge units (paradigm, concepts, methods etc.) internalized by students.

Speaking on a different scale, one can formalize the learning trajectory, for instance, simply as the sequence of courses finished by a student. Learning trajectory naturally lends itself to the formalized trajectory in a corresponding network of educational artifacts. Nodes represent units of knowledge, courses and resources needed to get the knowledge. Graph-based methods allow addressing the problem of learning trajectory optimization, for instance, by minimizing the duration of education. By attaching additional information to network nodes, one can compute optimization using additional criteria. Network models allow to compute the traffic in the network, and to solve the capacitated transportation problem.

Greedy strategy in building the trajectory suggests the usage of the iterative process. On each step of iteration, the agent chooses the most suitable object. In doing so we must consider two levels, scales of analysis, as it is depicted on the Fig. 1. The trajectory is build on the higher level of objects, the optimal choice of the element must be done by analyzing objects on the micro level (which because of their complexity usually also requires application of graph-based methods).

Information technology has changed the usual forms of communication, information work and social interaction between people. In the past 10 years, these technological innovations are changing education. Phenomenon MOOC (massive open on-line courses) demonstrates one of the trends characteristic of modern education. More than 4.5 million students in the world use online courses EdX, Udacity and Coursera [Carr, 2012]. A huge amount of data available belongs to a more disruptive and far-reaching change - "big data". According to [Guthrie, 2013] universities can use big data to customize courses and learning modules for student's needs. Individualization of education is becoming a key factor in the success of modern education [Robinson, 2010]. Individualization of education is an opportunity for students to choose courses of study, teaching materials, teachers in accordance with their interests.

New possibilities give rise to new problems. Opportunity to take courses or other learning resources by different authors from different sources creates a problem of resource or courses selection on the condition of the student interests and therefore the problem of constructing a trajectory in the space of learning resources and navigation between educational resources.

Recruitment agencies also became interested in what the applicant studied courses meet the requirements of a career position. They need to know the content of student learning modules or courses in different universities or training centers and need a tool to solve this problem.

Learning trajectory is a complicated concept using in many meanings. Educators and psychologists call the learning trajectory sequence of elements of knowledge (ideas, concepts, methods, etc.) that develops student. Changing the scale we come to understand the learning trajectory as a sequence of courses that the student is studying. Learning trajectory is naturally represented as a graph. Nodes of the graph represent the knowledge, training or resources necessary for their development. Methods of the graph theory can solve some problems by optimizing learning trajectory, such as optimization of training time. Adding a description of the node of her prerequisites, you can automate the construction of possible learning trajectories. Graph models allow to analyze the capacity of the resultant structure [Maruev and Shilin, 2012] and evaluate the necessary resources [Maruev and Gorbunova, 2012] for her work.

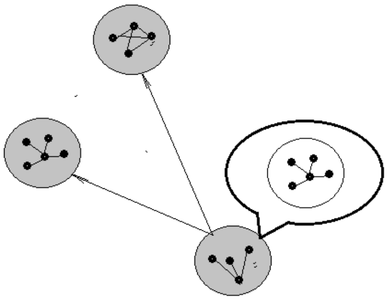


Fig. 1. Levels of modelling. Gray nodes – mezo-level, black nodes – micro-level.

Agent builds its learning trajectory as an iterative process. He selects the next element at each step of the process. Choice is optimal if the selected item is more resemble than other to the expectation of other agent. We are working on two levels of scale in Fig.1. Building a path of generalized elements on mezo-level and doing the optimal choice using their internal structure on micro-level.

Application of Graph-mining for the Selection of the Most Suitable Educational Resource

In this section we demonstrate the application of network modeling and graph-based methods for the selection of the most relevant educational module to cover a particular topic. To show the proposed methods in sufficient detail and to validate the results we used the following use case.

We took real life Russian language curriculum on the topic of macroeconomics, and two Russian textbooks on this topics; this curriculum and textbooks we will refer to as *C*, *T1*, and *T2*. We use *C* to automatically extract vocabulary for the macroeconomics and to build a network or semantic relations between concepts as it is seen from the text and the structure of the curricula. This vocabulary is then used to analyze *T1* and *T2* and to build networks corresponding to the relations between terms in these resources. We present a novel generic method for comparing graphs which allows us to quantify how well an educational resource covers the program.

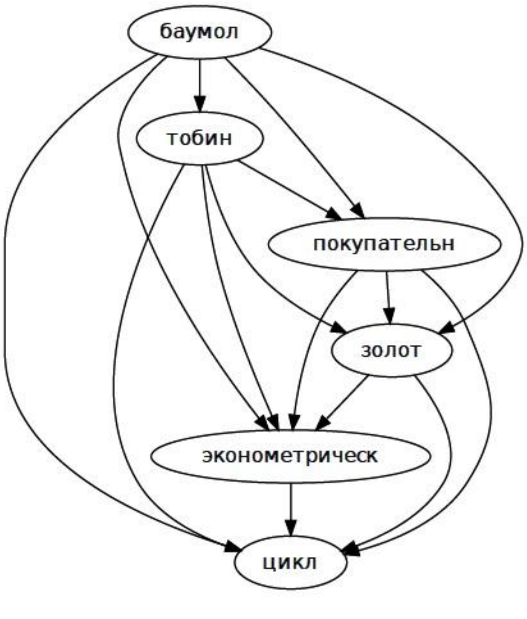


Fig. 2. A fragment of the network of concepts extracted from the curriculum on microeconomics. The nodes represent stems for several Russian words, terms and names of prominent economists (William Baumol, James Tobin, gold, econometrics, cycle). The links represent collocations of the words in the textbook 1. The whole configuration suggests, that the curriculum includes the Baumol–Tobin economic model of the transactions demand for money. Although our modeling doesn't use ontologies and recognition of multiword terms, our methods of modeling and mining are capable to indirectly capture and detect graph configuration of terms used in this model.

CONSTRUCTION OF THE NETWORK MODELS OF CARRICULUM AND EDUCATIONAL RESOURCES

We preprocess the text of the curriculum to find orthographic words, to filter out so called "stop" words (using Google's list of stop words for Russian), and to stem words to index forms using Porter stemmer for Russian (see, for instance, [Jurafsky and Martin 2009]). Words which are met within the window of three sentences are connected with an arc; the weight of the arc is calculated based on the number of the connections. An example of such collocation graph is shown on the Fig. 2.

Collocations per se do not represent semantic relations (since collocations captures various relations, including, for instance, sintagmatic relations), but the resulting network in the context of our study might be considered as semantic network.

MINING – FINDING STRUCTURAL SIMILARITIES IN NETWORKS WITH THE SAME SET OF NODES

Apparently, the global topology of these three network models – C, T1, and T2 - could be quite different. However, there must be some similarity at the level of the local topology, especially at the level of micro- and mezzo-clustering. To measure similarities which might be related to the coverage of topics, we propose the following methods.

We assume that the local clusterization in the graph *C* must be high. For instance, all multiword expressions like *real balance effect*, form complete graphs, where each word used in the multiword term is connected to another words in the expression. We also might put forward a hypothesis that if we select at random a set of concepts from the vocabulary of *C*, and compute the number of nodes which are within the distance two from the initial set in all three graphs *C*, *T1*, and *T2*, these numbers should be approximately the same. It is also clear that, from the other hand, if instead of a real text *T1* we'll take a random list of word forms from the vocabulary in *C*, most of the multiword terms will not be seen in the topology of such meaningless text. Therefore, If, for instance, this number is high in the network *C* and *T1*, but small in the network *T2*, we might suppose that a certain topic (or topics) of the curriculum is not covered in the textbook *T2* in sufficient details.

To find the neighborhood of a set of nodes, and to ensure the extensibility of the method to work with fuzzy sets of nodes, we use method of spreading activation (see, for instance, [Crestani 1997], [Troussov et al. 2009]) with two iterations. From the considerations above, we suppose that the high number of activated nodes in our experiments is a good predictor that the textbook covers the curriculum in sufficient details.

EVALUATION

The number of nodes in networks *C, T1*, and *T2* is the same and equals to 260. We split this number into 52 sets with 5 elements. Correspondingly, we generated 66 sets of nodes and used spreading activation to propagate the activation to neighbor nodes in networks *T1* and *T2*. The cumulated number of activated nodes in the network *T1* is 617, for the network *T2* this number equals to 1152. Therefore, according to our metrics, we conclude that the textbook *T2* is better learning object for the curriculum in the question. This result computed by our graph-based method metrics is consistent with the manual evaluation provided by experts in the field. However, the amount of the experimentation is too small to conclude that the feasibility of our method is proven.

Conclusions and Future Work

The practical importance of our research stems from the processes of globalization, personalization of education, and from the explosive growth of the availability of on-line training courses. Multidimensional networks of modern educational processes are huge; therefore naïve (common-sense) navigation tools are not sufficient for their analysis, and new computer-based navigation tools are to be designed.

The central element of the navigation tools is a block decision making choosing one of the many educational elements based on a comparison of their contents. If the elements exist in the form of free texts, someone must read, understand all the texts and make a decision based on comparison of them. Such endeavor is impossible when we move into big data. It is therefore necessary to develop methods of automated processing big volumes of training resources for the decision to select an element of a learning trajectory.

In this paper we present the initial results in modeling educational process as the navigation in multidimensional networks and the pertaining algorithms of optimization of that navigation. The results of our research could be useful for the building of educational resources (for instance, by finding structural weaknesses in existing networks), as well as for the personalization of the education.

The presented model of elements of a learning trajectory is a multidimensional network. We have developed a method of constructing such a multidimensional network and the algorithm for selecting the most similar element on each step of the iterative algorithm of a learning trajectory constructing. Our algorithm used the spread activation method, which has been successfully used for mining multidimensional networks. The efficiency of the algorithm is shown in the example of choosing a textbook for a particular program (curriculum). Further research in this direction involves experiments with big data, construction of models with more dimensions and improvement of the spreading activation algorithm (an example of a new strand of spreading activation algorithms is provided in [Troussov et al. 2011a]).

Bibliography

[Carr, N., 2012]. The Crisis in Higher Education MIT Technology Review magazine November/December 2012/ http:// technologyrReview.com/featuredstory/429376/ the-crisis- in-higher-education

[Crestani, F., 1997]. Application of Spreading Activation Techniques in Information Retrieval. Artificial Intelligence Review, 11(6), 453-482.

[Guthrie, D., 2013]. The coming Big Data Education Revolution. US News. http://www.usnews.com/opinion/articles/2013/08/15/why-big-data-not-mooc-will-revolutionize-education

[Jurafsky, D. and Martin, J.H., 2009]. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition (2ed.), Prentice Hall.

[Maruev S. and Gorbunova E., 2012b]. Analysis of the Throughput of the Process of Distance Learning. Artificial Intelligence Driven Solutions to Business and Engineering Problems. ITHEA Rzeszow-Sofia, 2012. P.7-11.

[Maruev S. and Shilin K., 2012a]. Model of Resource Potential Estimation for Quality of Education Ensuring. Economic Policy. #1, 2012, P.78-86. (In Russian)

[Robinson, K., 2010]. Changing Education Paradigms http://www.ted.com/talks/ken_robinson_changing_education_paradigms

[Troussov, A., Darena, F., Zizka, J., Parra, D., and Brusilovsky, P., 2011a]. "Vectorised Spreading Activation Algorithm for Centrality Measurement". Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis. sv. LIX, no. 7, s. 469--476. ISSN 1211-8516.

[Troussov, A., Jundge, J., Alexandrov, M., and Levner, E., 2011]. Social Context as Machine-Processable Knowledge. Proceedings of the International Conference on Intelligent Information and Engineering Systems INFOS 2011, Rzeszów - Polańczyk, Poland, pp. 104-114, ISBN: 978-954-16-0053-5.

[Troussov, A., Levner, E., Bogdan, C., Judge, J., and Botvich, D., 2009]. Spread of Activation Methods. In Dynamic and Advanced Data Mining for Progressing Technological Development, Y. Xiang and S. Ali (eds) IGI Global.

Authors' Information

Sergey Maruev – Professor, The Russian Presidential Academy of National Economy and Public Administration (RANEPA), Prosp. Vernadskogo 82, bld. 5, Moscow, 119571, Russia; e-mail: maruev@rane.ru

Major Fields of Scientific Research: mathematical modeling, economics of education, business process modeling, social network analysis, operations research

Eugene Levner – Ashkelon Academic College, 52, Golomb St, Holon 68102 Israel; e-mail: 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

Dmitry Stefanovskyi – Assoc. Prof., Ph.D, The Russian Presidential Academy of national economy and public administration (RANEPA), Prosp. Vernadskogo 82, bld. 5, Moscow, 119571, Russian Federation; e-mail: dstefanovskiy@gmail.com

Major Fields of Scientific Research: mathematical modeling, world economy

Alexander Troussov – Director of the International Research Laboratory for Mathematical Methods for Social Network Mining, The Russian Presidential Academy of National Economy and Public Administration (RANEPA), Prosp. Vernadskogo 82, bld. 5, Moscow, 119571, Russian Federation; e-mail: troussov@gmail.com

Major Fields of Scientific Research: Natural language processing, Information Retrieval, Social and semantic web, social network analysis, graph-based methods

MULTIDIMENSIONAL NETWORKS FOR HETEROGENEOUS DATA MODELING

Sergey Maruev, Dmitry Stefanovskyi, Alexander Troussov, John Curry, Alexey Frolov

***Abstract**: Big data frequently come in tabular form of rows and columns of numbers, special codes and short textual descriptions, in strict, structured, disciplined formats generated by a variety of transactional and operational business systems. In this paper we discuss the advantages of modeling heterogeneous data by multidimensional networks in line with the concept known as "Graph databases". Graph-based methods provide a powerful abstraction for mining such data; however, it is hard to achieve good results in mining using of the shelf methods. In this paper we show how empirical methods of fuzzy logic could be injected into abstract graph-based methods to achieve desirable results. We outline the wide range of applications of that modeling and mining, and present our results on the use of our methods of modeling and mining for processing of custom declarations for commercial goods. We examine several use cases, including recommendations to custom officers and participants of the international trade. The feasibility of the approach was tested by application to 2500 custom records collected during a continuous period of one month at eight border checkpoints between Russian Federation and two EU countries. In several use cases the algorithm achieved high accuracy under experimental conditions.*

***Keywords**: big data, graph-based methods, custom declarations.*

***ACM Classification Keywords**: Algorithms, Economics, Experimentation, Theory.*

Introduction

Big data frequently come in tabular form of rows and columns of numbers, special codes and short textual descriptions, in strict, structured, disciplined formats generated by a variety of transactional and operational business systems. In this paper we discuss the advantages of modeling heterogeneous data by multidimensional networks in line with the concept known as "Graph databases". Graph-based methods provide a powerful abstraction for mining such data; however, it is hard to achieve good results in mining using of the shelf methods. In this paper we show how empirical methods of fuzzy logic could be injected into abstract graph-based methods to achieve desirable results. We outline the wide range of applications of that modeling and mining, and present our results on the use of our methods of modeling and mining for important area of applications - processing of custom declarations for commercial goods.

International trade is one of the most important drivers of the global economy. Therefore, the study of impediments to this trade is of interest to the field of international economics. International trade is typically more costly than domestic trade due to the imposition of extra direct and indirect costs including tariffs, time costs due to border delays and processing costs that are exacerbated by differences in language, legal system and culture, see, for instance, [Zvetkov et al. 2013] in Russian.

We examine several use cases, including recommendations to custom officers and participants of the international trade. The feasibility of the approach was tested by application to 2500 custom records (which have 12043 items of goods) collected during a continuous period of one month at eight border checkpoints between Russian Federation and two EU countries, the same data set that was used in [Maruev et al. 2014].

We tested our approach on the use case of computing the code of custom goods based on the textual description provided in the declaration; the algorithm achieved high accuracy under experimental conditions.

The rest of the paper is organized as follows. Representing data in rows and columns probably has been the most pervasive formal method for data collection, representation and analysis in all areas of human activities, notably including the use in computer data bases. In section entitled "Network Modeling vs. Tabular Representation" we provide a brief description of modeling using multidimensional networks and comparison of such modeling with the table representation. We show how tables could be converted into multidimensional networks, and argue that such modeling naturally lends itself to the discovery of patterns. The bulk of the paper is the demonstration how real world data about custom declarations could be modeled by networks and explored using methods of graph mining.

In the next section - "Custom Declarations data" - we describe the data used in this paper. In section "Network Data Representation and Mining" we present a particular way of network modeling tailored to the task of prediction of the nomenclature code of goods from the textual description. The resulting network is a multidimensional network with two types of nodes: nodes corresponding to the nomenclature codes of goods, and nodes corresponding to the words used in natural language descriptions of goods. Assuming that all the data in our collection have correct nomenclature codes, we can consider the obtained network as an encapsulation of the knowledge about the relations between codes and words in the textual descriptions of the goods. New textual descriptions could be mapped into nodes of this network, the results of the mapping might be considered as a fuzzy set of nodes. Measuring "proximity" of this set to nodes representing nomenclature codes one can quantify the relevancy of the description to certain codes. In this paper we use the generic computational scheme on networks called spreading activation [Troussov et al., 2009] for this purpose. In section "Evaluation" we present the results of experimental validation of recommending codes based on the textual description of goods. Finally, section "Conclusions and Future Work" describes the conclusions and future work.

Network Modeling vs. Tabular Representation

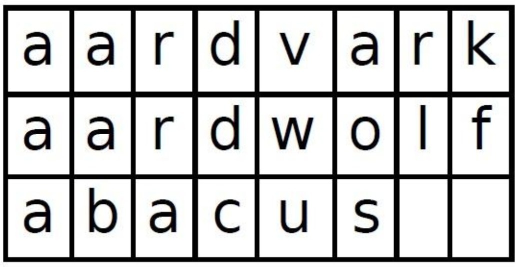
Network modeling is endemic throughout various domains of applications, including social and semantic web, communications. To introduce network method for custom declaration's data representation, we juxtapose this method with the applications in computational linguistics, where we can visually show the difference between network and tabular representations.

Many data in office applications comes as "tables", which could be processed by spreadsheet applications such as Microsoft Excel. Data which are viewed as networks, such as social networks, in many cases are converted to matrix form (to a "table") as incidence matrix and processed using linear algebra methods (in mathematics, an incidence matrix is a matrix that shows the relationship between two classes of objects.). However, linear algebra provides only a subclass of useful graph-mining techniques.

In this paper we argue that the usefulness of graph-based methods for mining of unstructured heterogeneous data (usually represented as tabular data) is underappreciated. This statement is somewhat similar to the ideas which led to the coinage of the term "graph databases" [Rodriguez, 2011], although our emphasize is solely on the data representation and mining algorithms relying on the navigation through a network using links between neighbors, not on the methods of storage of graphs or spars matrix.

To illustrate our point let us consider computer dictionaries for natural languages.

Many lists of common English words starts from aardvark, aardwolf, abacus, … . To use such lists for solving crosswords in would be suitable to model the data in a tabular form like this:



a	a	r	d	v	a	r	k
a	a	r	d	w	o	l	f
a	b	a	c	u	s		

Fig. 1. A list of a few English words could be represented in a tabular form, which could be suitable for certain applications, like solving crosswords or computing of statistics of letters in certain positions

The same table could be redrawn as a graph where nodes represent letters.

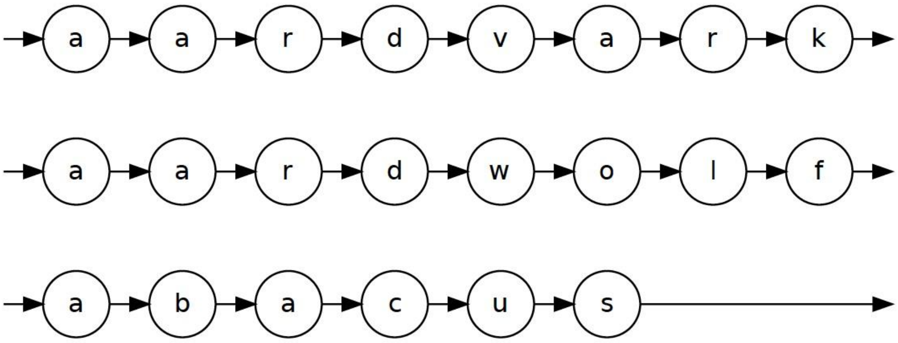


Fig. 2. The list of English words from the Fig. 1 could be visualized as a graph.

If this graph is used to construct a computer dictionary, it can be compactified to the following form:

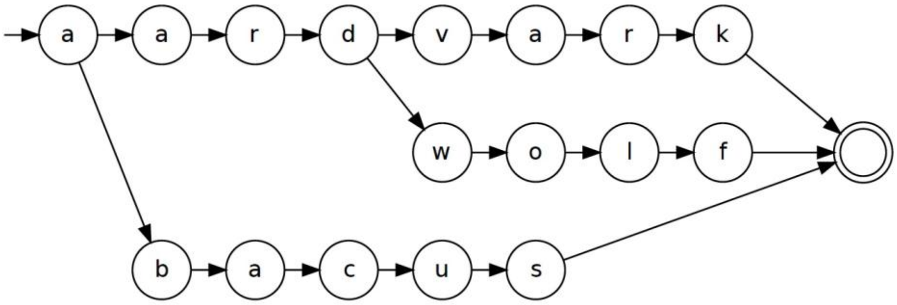


Fig. 3. This graph model has exactly the same data as the original list of common English words on the Fig. 1, and is usually called the Mealy finite-state machine.

For computer science problems, such as designing the data structure for dictionaries which support search operation, a standard solution using hash tables could be used as well as the Mealy finite-state machine shown on the Fig. 3; and one can argue about advantages and disadvantages of both methods. However, the situation drastically changes when we move from lists of random character strings to the lists of words from the vocabulary of a natural language.

When the strings are words from a natural language, graph-based representation of data has at least one crucial advantage. Firstly, the graph-representation becomes very compact since common prefixes and affixes of words are conflated, and this leads to non-functional advantages (in memory footprint and processing speed). More importantly, even when such compactification is provided by formal mathematical methods, which are unaware of the morphology, effectively they produce a representation of the initial list of strings in a graph form which shows patterns of the morphology of the language; therefore it becomes possible to process out-of-vocabulary words, like *trichloroisocyanuric*, and to construct morphological guessers [Jurafsky and Martin, 2009]. For instance, a morphological guesser might infer that the word *ontologization* is a well formed English noun, and to find that it is related to the noun *ontology*. Moreover, using graph-representation one can infer that the relation between the pair of words *ontology-ontologization* is the same as the relation between words *industry-industrialization* (see, for instance, [Troussov and O'Donovan, 2003]).

The procedure of the processing out of vocabulary words, like the word *ontologisation*, might be summarized in the following diagram:

```
      Input: a new word: ontologization  → Network model
              of  existing  words   →   Patterns  which  the  input
follows
```

When the network model of custom declaration has been already constructed, the scheme of processing new declarations is the same as it is in the above mentioned morphological applications:

```
      Input: a new custom declaration record →
                → Network model of custom declaration data →
                             → Patterns which the input follows
```

There is also a significant difference between these two use cases. In the computational linguistic all the patterns of the language could be discovered using the human insight and the methods of the computational linguistics. In case of mining custom declarations, on the fly and on demand discovery of emerging patterns moves to the foreground.

Custom Declarations Data Description

The data used in this paper are the same as in the paper [Maruev et al. 2014]. In this section we provide a short overview of this data.

The raw data for this study originates with traders shipping goods into the Russian Federation through land borders. They comprise a description of the itemized contents of a shipment of goods in a particular vehicle (always a truck in this study). These data are used to produce custom goods declarations and to compute taxes. When the truck crosses the border, the data become part of the custom service’s electronic data archive.

Each item record describes a specific type of goods, and has several numeric and alphanumeric fields, including identification numbers of consignee, consignor, and carrier; gross weight, invoiced cost; currency code and currency rate. Fields relevant to this paper are:

- GoodsTNVEDCode – ten digits code for the commodity. This nomenclature is used in the Customs Union of Belarus, Kazakhstan, and Russia and is also consistent with the codes used in the European Union
- GoodsDescription – goods description;

Table 1 shows an example of an item record which uses Russian language description of goods related to printing machinery.

Table 1. An example of goods item with the nomenclature code (GoodsTNVEDCode) 8443999009.

№	DI	E	C1	C2	C3	Goods-TNVEDCode	GoodsDescription	GW	InvoicedCost	CC	CR
1	1	1	385	389	785	8443999009	ЧАСТИ И ПРИНАДЛЕЖНОСТИ ПЕЧАТНЫХ МАШИН	7482.320	531495.03	USD	33.2474

Network Data Representation and Mining

We represent these data as multidimensional networks (see [Troussov et al., 2011]). Nodes and links are typed. Nodes represent complete data fields or, as in the case of the “GoodsDescription”, particular words from that textual field. An example of such a network is given on Fig. 4 below. This network and the methods of its construction will be described in Subsection “Network Construction”.

In this paper we focus on a particular task – to automatically detect the code of the commodity based on its textual description. The same as in the paper [Maruev et al., 2014], we can consider this task as a task of supervised machine learning. We split the data set into two halves, learn the rules based on the first half, and apply the rules to the second half (not used in the procedure of learning). We then validate the results against the known outcomes to assess the predictive power of our rules.

In our approach, learning is done in two distinctive stages. Firstly, we build the network from the data. Secondly, we use graph algorithms to find patterns in the network.

Fig.4 shows the fragment of the network constructed in our validation experiment. The network was generated from the original tabular data in the reduced feature space where only nomenclature codes of the commodity and words from textual descriptions are used.

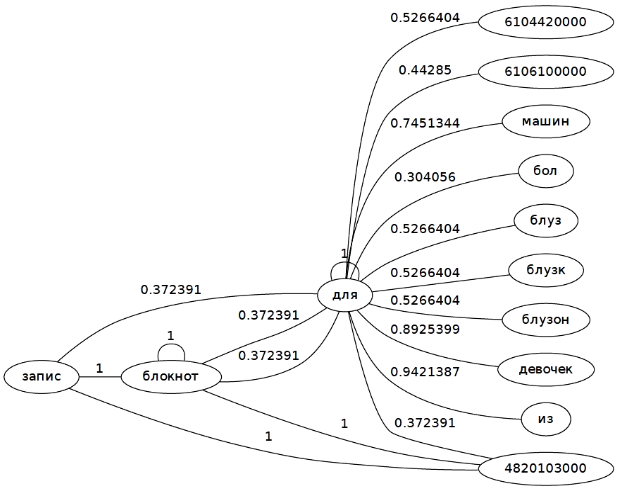


Fig. 4. A fragment of the network generated from the original list of goods items. Nodes labeled with numbers represent goods codes; nodes labeled with strings of alphabetical characters represent words used in textual descriptions. Links are weighted. This network shows, for example, that the word “для” (“for” in English) has been met in many goods descriptions and with many goods codes. Most of the formal methods, used to compute structural importance of the nodes in the network, will rate this word very high; one can also wrongly assume that this word could be a good predictor for many different codes. The mining method described in this paper allows preventing such nodes to dominate the results.

NETWORK CONSTRUCTION

The construction and the use of this network consist of the following generic steps which could be adjusted for use in other scenarios.

1. The first 8000 items description rows were used to construct the network
2. We removed all the fields except the commodity codes ('GoodsTNVEDCode"), and the goods description ("GoodsDescription").

The removal is not a necessary step, but we have done it in order to reduce the volume of learning data and test our algorithm under harsh conditions.

3. Each word has been reduced to its normalized form by the procedure known as stemming, see, for instance, [Jurafsky and Martin 2009]. Stemming is an empirical natural language processing procedure allowing to map inflected and derived words into one index form; for instance, to map "fishing", "fished", and "fisher" to the form "fish". We used Porter stemmer for Russian to perform this procedure.
4. As the result, we obtained 4352 entities, which were merged into one network with 4352 nodes.

If the two entities are met at least once in one same shipment document, the corresponding pair of nodes is connected by an arc. The weight of that arc represents how frequently the two entities corresponding to the pair of nodes are met in a shipment document, i.e. the number of co-occurrences divided by the number of items.

MODELING NEW DOCUMENTS AS SETS OF NODES ON THE NETWORK

The network represents the learning data. Each new document could now be modeled as a set of nodes on this network. I.e. for each new item we need to perform steps 1-3 described above. Each new entity is mapped into a corresponding node on the network. If such a node is not found, the entity is ignored. It could not be usefully present in the model because our "learning" has no knowledge about such entities.

MINING

When we encode a set of data as a network, such as described in the previous subsection, mining now can be done by various graph-based methods. For the recommendation tasks, one can model the situation as a set of nodes on this network, and based on the graph-topology discover other nodes which might be relevant (close) to the initial conditions. Specifically for our task - computing GoodsTNVEDCode based on the textual description – we model the description as a set of nodes (corresponding to individual orthographic words in the description) on the network, and using graph-methods find the most relevant nodes representing GoodsTNVEDCodes. To find and rank related nodes, we used the set of operations based on the Spreading Activation Method, described in [Troussov et al. 2009], and its generalization in the paper [Troussov et al. 2011].

The Spreading Activation Method has its origin in neurophysiology: "In neurophysiology interactions between neurons is modeled by way of activation which propagates from one neuron to another via connections called synapses to transmit information using chemical signals. The first spreading activation models were used in cognitive psychology to model this processes of memory retrieval." – [Troussov et al. 2009]. Later this framework was exploited in Artificial Intelligence as a method for searching associative, neural or semantic networks; see, for example,[Crestani 1997], [Aleman-Meza et al. 2003], [Rocha et al. 2004].

In terms of the spreading activation, our mining could be explained as follows: we put the initial activation at those network nodes which correspond to words used in the description, and compute how much activation comes to the nodes corresponding GoodsTNVEDCodes. Spreading activation serves as a search method in the work, which also allows to compute the cumulative strengths of connections between the words in the description and the GoodsTNVEDCode.

Depending on the task of mining and the structural properties of the network, a few up to several dozen iterations of spreading activation are normally sufficient to achieve the goal. We found that one iteration of spreading activation is enough for our purposes. In other words, the number of arcs and their weights between the model and the node GoodsTNVEDCode is a good predictor that the goods code is consistent with the given textual description. The larger the weights, likelier it is that the goods code is a correct one. The effect of the number of arcs here is much less evident, because, logically, a high number of arcs with small weights indicate that the goods code is wrong.

To aggregate weights one can use the arithmetic mean of the weights of arcs. However, for the use case of recommendations to assign an armed convoy for the shipment, [Maruev et al. 2014] found that arcs with high weights close to *1.0* were important, while arcs with small weights were not reliable predictors. Therefore, instead of arithmetic mean for *n* real numbers x_1, x_2, …, x_n representing weights, [Maruev et al. 2014] used the mean computed as the L^p –norm of the vector { x_1, x_2, …, x_n } with the parameter *p* empirically taken with the value *3.5* to favor links with high weights and to ignore links with very small weights:

$$\| \mathrm{x} \|_{\mathrm{p}} = (| x_1 |^{\mathrm{p}} + | x_2 |^{\mathrm{p}} + \ldots + | x_n |^{\mathrm{p}})^{1/p} \qquad (1)$$

In this paper we use the same formula 1 for the aggregation; however, we found that the parameter *p* in our case should be different.

Evaluation

We used 8000 chronologically first data records to model the data as a network of custom goods codes and words used in goods description; and the rest of the data (4043 records) to test our algorithm. Each new record was broken into words and mapped onto the network, and the strength of its connection to various codes was computed using the spreading activation method. The output of this algorithm is the list of the nodes corresponding to the strength of the connection with the set of words, the strength of the connection in this context is called the level of activation (see [Troussov et al., 2009]). If two or more nodes have the same activation, the order in which they appear in the list is random.

If the most activated node is the same as the node corresponding to the code in the record in question, the code is considered to be predicted correctly. We investigated the accuracy of the prediction under the various values of the parameter *p* in formula 1.

If the recall is measured in top two goods codes (that is the result of the prediction is considered as to be correct if either the most activated node or the second most activated node is the node corresponding to the code in the record), the recall is 100%. In all cases where the correct answer was the second most activated node, the textual description was not sufficient to predict the nomenclature code; for instance, the code for tomato fruits depends on the additional information absent in the data we used, such as the season when the cargo enters the country. One can easily fix this particular problem with “hard rules”, but the goal of our pilot project is to create proof-of-the-concept scalable technology platform for mining custom declarations for cases not covered by hard rules of custom regulation, as opposed to another people-intensive manual solution.

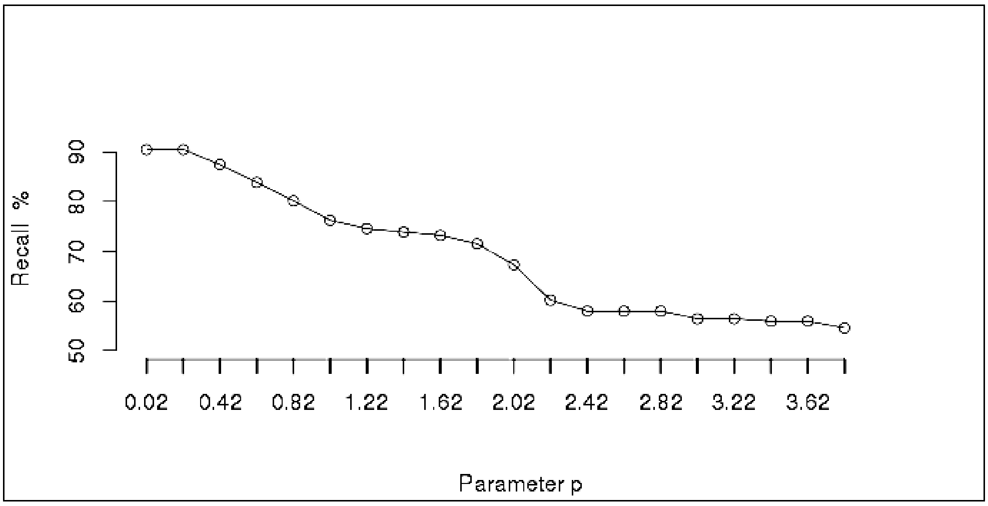


Fig. 5. The recall of the prediction of goods nomenclature codes in testing 4023 custom records based on the network created using different set of 8000 records. With the parameter p close to zero the recall is 90.5 %.

To understand if the achieved results are good or not in the paradigm of the machine learning, we need to take into consideration that the amount of the data used for learning (that is the data used to construct the network) is small (8000 records) as contrasted with the number of possible goods codes (about ten thousands). In addition, the data which we consider as the gold standard, might contain factual errors (such as the wrong codes), and definitely have certain number of misspellings in goods descriptions.

We conclude that the results of the experiments are the best possible to achieve under what might be considered as "harsh" conditions.

The most interesting theoretical result for us was the following. In this paper we used spreading activation as mining method, the same as was used in the paper [Maruev et al, 2014], where this method has been applied for a different use case of mining, namely recommendation of assigning armed convoy to the track when it crosses the border. Spreading of the activation algorithms are based on the iterative re-computation of activation of the network nodes. On each iteration the new level of activation is computed based on current activation in the node and the activation transferred to the node by its neighbors; in [Troussov et al., 2009], this stage is called "Computation of the New Level of Activation".

Firstly, we discover that the same formula (formula 1) of re-computation used in [Maruev et al. 2014] works well for our task, however good results are achieved with different value of the parameter *p*. Formula 1 with paremeter p>=1 is used in fuzzy logic to implement logical operation AND using Jager's *t*-norms (see [Chen, 1996]).

In [Maruev et al., 2014], the parameter *p* has been empirically taken with the value *3.5*. The rationale of this could be explained as follows. During the re-computation stage, the activation at each neighbor node around the node representing assigning of armed convoy could be considered as a predictor of convoy assignment; some predictors are more important than other. Formula 1 effectively aggregates these predictors into one real number using a specific type of logical operation AND. If this number is bigger than a certain threshold the algorithm recommends assigning the armed convoy. Aggregation of the several numbers into one could be done in many different ways, for instance one can take the arithmetic mean of these numbers. When the aggregation is done using formula 1 with $p=2.0$ the result become more dependent on the most important predictors; with the parameter $p=3.5$ the result become heavily dependent on the most important predictors, while less important predictors are practically ignored. Indeed, the escort is assigned to the track, so the fact that some of the goods in the track previously were not provided by the escort is not important, but the fact that some of the goods previously were escorted, is very important.

The experimental results depicted at the Fig. 5 show that the best results are achieved with small value of *p*, which essentially means that the number of the predictors for the particular goods code is important, while giving more importance to strong predictors quickly degrades the results. At the moment we can't explain this phenomenon in a rigorous way. Tentatively, we speculate along the following two lines.

Firstly, in our experiments to predict the code based on the textual description, we filter out some words using an empirical technique known in information retrieval as the removal of "stop words"; typically these are functional words forming so called closed word classes (see [Jurafsky and Martin, 2009]), like prepositions, conjunctions, particles, etc. However, after the removal of stop words we still have a big number of certain words (like *weight*, *mkm*) which describe quantity, weight and methods of packaging of goods, and are applied to many different goods codes. Such words technically become strong predictors to many completely unrelated goods codes. To improve the quality of graph mining one can manually remove such words and/or to develop graph-based methods which prevents these words from dominating the outcome of the algorithm.

Secondly, although some words are strong predictors of particular codes, and the noun phrases used in goods descriptions are mostly semantically compositional (that is the meaning of the expression is composed from the meaning of individual words (see [Jurafsky and Martin, 2009]), only the particular combination of all words in the description eventually determines the code. Based on the results of our experiments, it seems that our method of modeling captures this peculiarity very well in the topology of the network, and our mining method - spreading activation with formula 1 – could be quite sensitive to such combinations, and is capable to implicitly find such combinations and use them for the correct prediction of the goods code. To illustrate this let us consider the real example from our data – the item described as "*jacket for computational devices*" (*čexól dljz vuchisliteljnuch ustrojstv* in Russian, Russian equivalent for *jacked* actually means removable or replaceable protective or insulating cover for an object). Both words *computational* and *devices* are very strong predictors for the goods codes related to electronics. However "*jacket for computational devices*" is not an electronics goods item.

Conclusions and Future Work

We introduced a generic method for modeling tabular data as a multidimensional network, where nodes represent various codes and alphanumeric fields, as well as the terms extracted from the fields containing natural language phrases. The network form of representation provides ease of merge of heterogeneous information; external knowledge could be added on the top of the network obtained from data as new nodes and new weighted arcs. We validated our approach on a task of finding patterns in 2500 custom records, containing 12043 items of goods, collected during a continuous period of one month at eight border checkpoints between Russian Federation and two EU countries.

In this paper we described the application of this modeling method to a network form of data representation of custom goods declarations. The experimental results of the paper, in conjunction with the results of the paper [Maruev et al. 2014], demonstrate the applicability of spreading activation based algorithms for mining this data for two polar use cases: the assignment of an armed escort to a shipping vehicle in cases of elevated risk of theft, and the prediction of the nomenclature code of goods based on the textual description. Obtained results show potential applications of our method for building recommender systems for use by customs officers, traders, carriers and insurers.

Our modification of the traditional spreading activation technique [Troussov et al., 2009] allows the explicit injection of fuzzy logic into the computational scheme to address specific problems of mining. Future development of network process methods might be driven by “physics-inspired” (see [Troussov et al., 2011a]) and “logic-inspired” (including cellular automata and fuzzy logic) approaches, which will allow synthesizing algorithms with desirable outcomes.

Bibliography

[Aleman-Meza, B., Halaschek, C., Arpinar, I., Sheth, A., 2003]. Context-Aware Semantic Association Ranking. Proceedings of SWDB'03, Berlin, Germany, 33-50.

[Chen, C.H. (ed.), 1996]. Fuzzy Logic and Neural Network Handbook. McGraw-Hill, 700 pp.

[Crestani, F., 1997]. Application of Spreading Activation Techniques in Information Retrieval. Artificial Intelligence Review, 11(6), 453-482.

[Jurafsky, D. and Martin, J.H., 2009]. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition (2ed.), Prentice Hall.

[Maruev, S., Stefanovskyi, D., Frolov, A., Troussov, A., and Curry, J.] Deep mining of custom declarations for commercial goods. To appear at Procedia Economics and Finance, 2014.

[Rocha, C, Schwabe, D., Poggi de Aragao, M., 2004]. A Hybrid Approach for Searching in the Semantic Web. Proceedings of the 13th international conference on World Wide Web, May 17-20, 2004, New York, NY, USA, 374-383.

[Rodriguez, M. Knowledge Representation and Reasoning with Graph Databases, 2011]. http://markorodriguez.com/2011/02/23/knowledge-representation-and-reasoning-with-graph-databases/

[Troussov, A., and O'Donovan, B]. "Morphosyntactic Annotation and Lemmatization Based on the Finite-State Dictionary of Wordformation Elements". Proceeding of the International Conference Speech and Computer (SPECOM’ 2003), October 27-29 2003, Moscow, Russia

[Troussov, A., Darena, F., Zizka, J., Parra, D., and Brusilovsky, P., 2011a]. "Vectorised Spreading Activation Algorithm for Centrality Measurement". Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis. sv. LIX, no. 7, s. 469--476. ISSN 1211-8516.

[Troussov, A., Jundge, J., Alexandrov, M., and Levner, E., 2011]. Social Context as Machine-Processable Knowledge. Proceedings of the International Conference on Intelligent Information and Engineering Systems INFOS 2011, Rzeszów - Polańczyk, Poland, pp. 104-114, ISBN: 978-954-16-0053-5.

[Troussov, A., Levner, E., Bogdan, C., Judge, J., Botvich, D., 2009]. Spread of Activation Methods. In Dynamic and Advanced Data Mining for Progressing Technological Development, Y. Xiang and S. Ali (eds) IGI Global.

[Zvetkov, V, Zoidov, K, and Medkov, A., 2013]. In Russian. Цветков, В, Зоидов, К., Медков, А. О возможности и целесообразности организации транзита через Россию грузов между странами Тихоокеанского региона и Европы. Депонирована в системе Соционет, февраль 2013 г. Retrieved February 27, 2014, from http://www.cemi.rssi.ru/mei/articles/tsvetkov-and13-01.pdf

Authors' Information

Sergey Maruev – Professor, The Russian Presidential Academy of National Economy and Public Administration, Prosp. Vernadskogo 82, bld. 5, Moscow, 119571, Russia; e-mail: maruev@gmail.com

Major Fields of Scientific Research: mathematical modeling, economics of education, business process modeling, social network analysis

Dmitry Stefanovskyi – Assoc. Prof., Ph.D, The Russian Presidential Academy of national economy and public administration; Prosp. Vernadskogo 82, bld. 1, Moscow, 119571, Russian Federation; e-mail: dstefanovskiy@gmail.com

Major Fields of Scientific Research: mathematical modeling, world economy

Alexander Troussov – Director of the International Research Laboratory for Mathematical Methods for Social Network Mining, The Russian Presidential Academy of National Economy and Public Administration (RANEPA), Prosp. Vernadskogo 82, bld. 5, Moscow, 119571, Russian Federation; e-mail: troussov@gmail.com

Major Fields of Scientific Research: Natural language processing, Information Retrieval, Social and semantic web, social network analysis, graph-based methods

John Curry – Office of the Revenue Commissioners, Central Revenue Information Office, Cathedral Street, Dublin 1, Ireland, e-mail: curry.john@gmail.com

Major Fields of Scientific Research: Data mining, Software technologies

Alexey Frolov – The Federal Customs Service of Russia, ROSTEK-Pskov, Director, Ul. Vorovskogo, bld.20, 180000, Pskov, Russian Federation, e-mail: info@rostek-pskov.ru

Major Fields of Scientific Research: Economics, World economy

FORECAST OF FORRESTER'S VARIABLES USING GMDH TECHNIQUE[1]

Olga Proncheva, Mikhail Alexandrov, Volodymyr Stepashko, Oleksiy Koshulko

Abstract: *In distant 1970 the MIT professor J. Forrester proposed a model of world dynamics in the form of five differential equations for 5 basic variables, namely: population, capital assets, agricultural assets ratio, pollution and natural resources. In the paper, we show the possibility to build a predictive model based on non-lineal difference equations using technique of inductive modeling. Unlike Forrester's model, this model provides a quantitative prediction. The model was tested on real data presented by World Bank and showed its high accuracy when forecasting for 1997-2012 years.*

Keywords: *inductive modeling, world dynamics.*

ACM Classification Keywords: *1.6.4. Model validation and analysis*

Introduction

Forrester´s macroeconomic variables are as follows:

- population (P);
- capital assets(K);
- agricultural assets ratio(X);
- pollution (Z);
- natural resources (R).

They relate to overpopulation of our planet, lack of basic resources, critical level of pollution, food shortages and industrialization as well as the related industrial growth. For these variables a differential model has built, named "classical J.Forrester model" [Forrester, 1979] presented here in a generalized form:

$$\frac{dP}{dt} = f_1(P,K,X,Z,R) \tag{1}$$

$$\frac{dK}{dt} = f_2(P,K,X,Z,R) \tag{2}$$

$$\frac{dX}{dt} = f_3(P,K,X,Z,R) \tag{3}$$

$$\frac{dZ}{dt} = f_4(P,K,X,Z,R) \tag{4}$$

$$\frac{dR}{dt} = f_5(P,K,X,Z,R) \tag{5}$$

Many other researchers also have built their own models of world dynamics [Egorov, 1980; Matrosov, 1999; Matrosov, 1999; Makhov, 2010; Meadows, 2007]. All of them were based on differential equations and have the following two properties:

- they provide long-term qualitative prediction because they take into account only tendencies;
- they are very subjective because model forms are specified and based only on author´s vision.

[1] The work done under support of the British Petroleum grant (BP-RANEPA 2013).

A topical problem is to build models which could

- provide quantitative predictionfor short-term and middle-term period;
- be free from subjective preferences of authors.

An adequate tool to solve this problem is well-known Group Method of Data Handling (GMDH) which is one of the most successful techniques of Inductive Modeling (IM). This method was proposed in 1968 by Prof. A. Ivakhnenko and now it is developed by his disciples and followers [Stepashko, 2013]. Now GMDH has numerous applications in natural sciences, economical, technical and social areas [Bulakh, 2013; Bulgakova, 2013; Kovalchuk, 2013; Pavlov, 2013; Samoilenko, 2013; Tutova, 2013; Zubov, 2013]. In the paper we build IM-models for middle-term forecast of Forrester's variables and test their accuracy according to the data for the period 1998-2012. By IM-models we mean here the models built using GMDH.

The paper structure is the following. Section 2 contains a short description of GMDH and here we build IM-model. In section 3 we present the results of experiments. Section 4 includes short discussion and directions of future research.

GMDH technique

General description

GMDH has long enough history [Ivakhnenko, 1968; Ivakhnenko, 1971; Madala, 1994; Stepashko, 1983] and presented in many publications. In what follows, we only remind its generalized scheme:

1. Certain model class is chosen. Models of gradually increasing complexity are generated in this class.
2. Learning data set is divided into training set and checking set to form an external criterion. Models are built on the first set and tested in predictive mode on the second one.
3. Model parameters are estimated using any internal criterion (e.g. the least squares) on the checking set.
4. Model quality is determined using any external criterion (e.g. the regularity criterion) on the checking set.
5. Model complexity is increased until the external criterion would reach its minimum.

It is easy to see that:

- GMDH belongs to the data-driven technique of inductive modeling, which means the models to be built are based on given instances; they are not theory-driven ones.
- GMDH realizes so-called model self-organization approach that is the models change their structure trying to be adjusted according to a given data of observations.
- GMDH is an evolutionary algorithm because model complexity is increased to reach the best one reflecting properties of observation data.

The final IM-model has an optimal complexity with regard to a) reflection of object properties contained in observation data and b) robustness to unknown factors which we call noise.

The most common class of traditional IM-models being considered when using GMDH is so called Kolmogorov-Gabor polynomial (Ivakhnenko, 1971):

$$Y(x_1, x_2, \dots, x_n) = a_0 + \sum_{i=1}^{n} a_i x_i + \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} x_i x_j + \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} a_{ijk} x_i x_j x_k + \cdots \quad (6)$$

During preprocessing stageone can consider not only natural powers of variables but also factional powers and functions from initial variables, for example exponential or trigonometric.

The classes of predictive models

The models to be built belong to the class of nonlinear difference equations. We decided to check predictive properties of two model types. The first one contains separate variables in different powers without their combinations. The second one additionally contains pairwise multiplications to take into account their mutual influence. Consequently, these structures are "simple model" and "model with pairwise multiplications".

The first experiments did not give any satisfied results, so we decided to make an additional preprocessing:

- Normalization: the values of variables were changed to belong to the interval [0; 1] to have the same scale. To do this, the numerical values of the population size, the capital assets and agricultural assets ratio in each year were divided by 10^{10}, the quantity of remaining natural resources –by10^{12}.
- Transformations: the variables were allowed to have not only natural powers but also fractional ones to provide that any small change in the variables did not lead to the sharp jump in the model parameters.

Thus, the new model type had the following form:

$$A_{i,t+1}=f(A_{j,t}^{-3/2}, A_{j,t}^{-1}, A_{j,t}^{-1/2}, A_{j,t}^{1/2}, A_{j,t}, A_{j,t}^{3/2}, A_{j,t-1}, ..., A_{j,t-T}), \tag{7}$$

$$A_{i,t+1}=f(A_{j,t}^{-3/2}, A_{j,t}^{-1}, A_{j,t}^{-1/2}, A_{j,t}^{1/2}, A_{j,t}, A_{j,t}^{3/2}, A_{j,t-1}, ..., A_{j,t-T}, A_{j,t}^{1/2}\cdot A_{l,t}^{1/2}, A_{j,t}^{-1/2}\cdot A_{l,t}^{-1/2)} \tag{8}$$

here: i=1,...,5; j=1,...,5; l=1,...,5; l≠ j; T –the needed forecast horizon; $A_{i,t}$– the value of i-th variable in a year t; f() – a function linear in parameters.

Using these models we produced forecast for the period 15 years from 1997 to 2012. Then the result of prediction was compared with the real data provided by the World Bank. The World Bank provides only data on population, the capital assets and agricultural assets ratio, so the data on pollution and natural resources have been restored. Therefore, we could compare real and forecasted dynamics only for these three variables.

To build IM-models we used the program package GMDH Shell (GS) developed by GEOS company to solve problems of time series forecast, function approximation and data classification [http://www.gmdhshell.com/]. It contains several GMDH-based algorithms and broad possibilities for visualization as well as pre- and post-processing. In our research we used the iterative GMDH algorithm of neural network type.

Experimental study of two models

Selection of best models

We completed two series of experiments, in each series the best models from classes (7) and (8) were chosen. The limitations on the models were: powers no more than 3/2, number of lags no more than 15. The models were trained on the interval 1900-1998 and then tested on the data 1998-2012. Model of each class with the greatest accuracy of forecast was considered as "the best model".

For the experiments we prepared data for training and checking:

a) Period 1900-1960: we used data from Forrester's model.

b) Period 1961-1998: we used data from World Bank for population, capital assets and agricultural assets ratio. The other variables (pollution and resources) were restored with Inductive Modeling technique. Data from Forrester model were used up to 1970 and then forecast was made for one year. Thereafter, the numerical values of the population, capital assets and agricultural assets ratio were replaced with real data and then forecast for 1 year was made and so on.

c) Period 1999-2013: we restored data with Inductive Modeling technique and then compared the results with real data provided by the World Bank.

Forecast for 15 years with the best simple model

In the next experiments the forecast horizon was equal to 15 years. The best IM-model in the class of "simple" models is:

$$P_{t+1} = -\ 0.00603058 + 0.00521011 \cdot Z_t^{-3/2} + 0.0255463 \cdot X_{t-7} + 1.22781 \cdot P_t - 0.182304 \cdot P_{t-2}$$

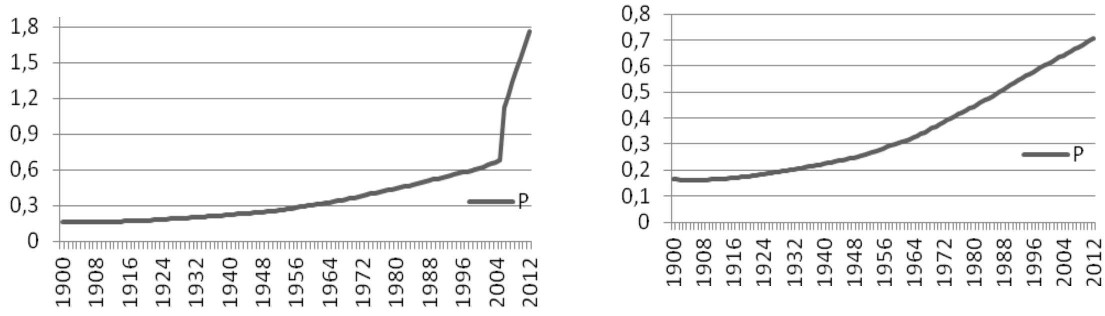
$$K_{t+1} = 0.00361846 + 1.38141 \cdot K_t - 0.372398 \cdot K_{t-1}$$

$$X_{t+1} = -\ 0.01487678 - 0.2118857 \cdot X_t + 0.943634 \cdot X_{t-5} + 1.6358 \cdot X_t^{3/2}$$

$$Z_{t+1} = -\ 0.00551411 + 0.787061 \cdot P_t^{-1/2}$$

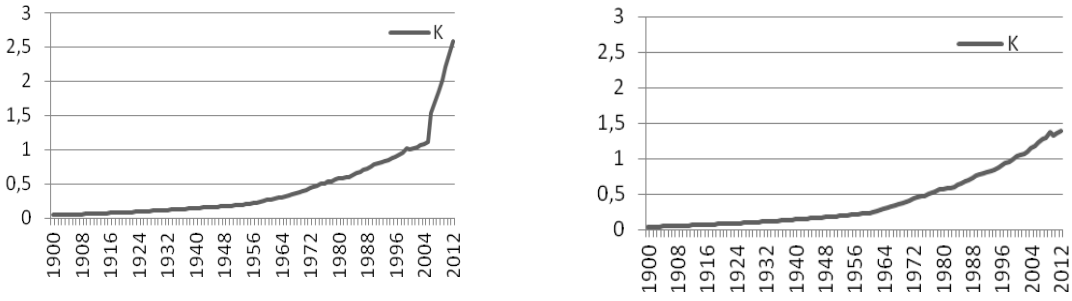
$$R_{t+1} = -\ 0.852927 - 0.0110949 \cdot P_{t-10} - 1.94371 \cdot Z_t^{1/2} + 0.965339 \cdot P_t^{1/2}$$

The real and forecast dynamics is compared below(fig. 7-8). The accuracy results of modeling are: R^2=0.98, RMSE=1.54%. Here R^2 is the coefficient of determination; RMSE is the root-mean-square-error.



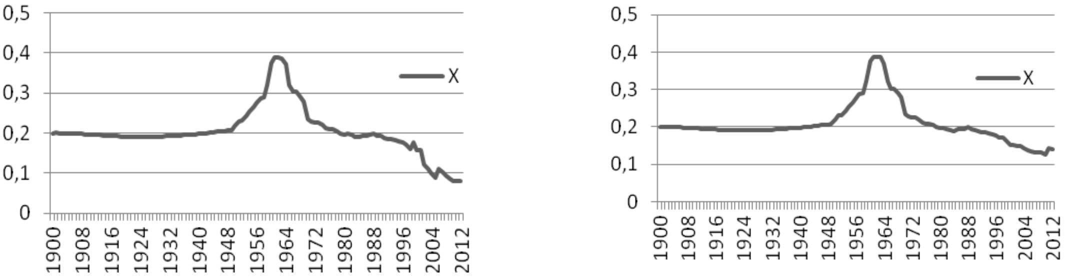
Forecast Real data

Fig. 7. Comparison of real and forecast dynamics of population



Forecast Real data

Fig. 8. Comparison of real and forecast dynamics of capital assets



Forecast Real data

Fig. 9. Comparison of real and forecast dynamics of the agricultural assets ratio

Forecast on 15 years with the best model having pairwise multiplications

The best IM-model in the class of models with pairwise multiplications is:

$$P_{t+1}= – 0.001603058 – 0.0468069\cdot P_{t–2}+0.0255463\cdot X_{t–7}+0.09195\cdot Z_t^{–3/2}$$

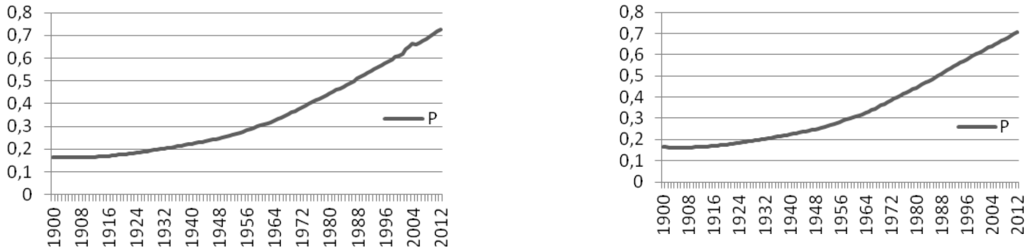
$$K_{t+1}= – 0.00187356 – 0.165445\cdot K_t\cdot X_t+1.00087\cdot K_t$$

$$X_{t+1} = – 0.000409618 – 0.132082\cdot t^{1/2}– 3.34978\cdot X_t^{1/2}+1.08047\ X_t^{3/2}$$

$$Z_{t+1} = – 1.98521e\text{-}16 + 0.787061\cdot Z_t\cdot P_t^{–1/2}$$

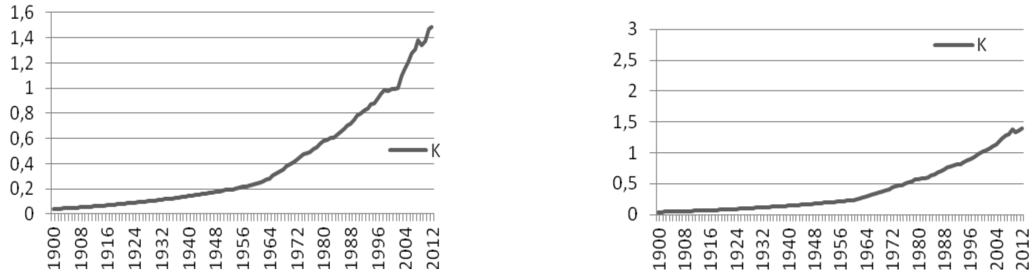
$$R_{t+1} = 0.852927 – 0.0110949\cdot P_{t–10}– 1.94371\cdot Z_t^{1/2}+Z_t^{1/2}\cdot P_t^{1/2}$$

The real and forecast dynamics is compared below(fig. 10-12). The accuracy results of modeling are: R^2=0.99, RMSE = 1,08%.



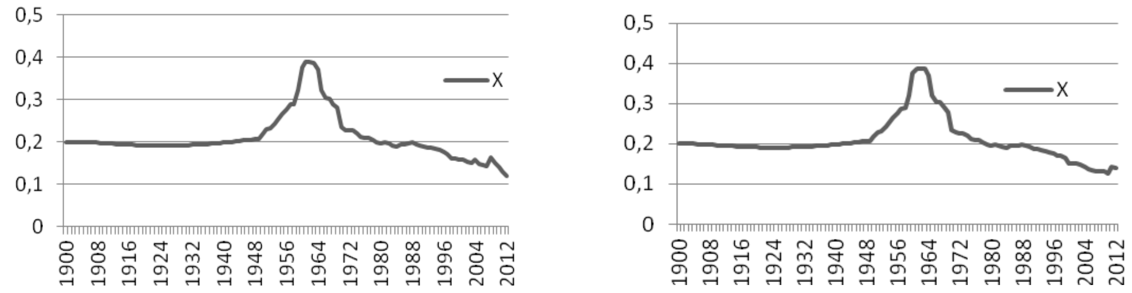
Forecast Real data

Fig. 10. Comparison of real and forecast dynamics of population



Forecast Real data

Fig. 11. Comparison of real and forecast dynamics of capital assets



Forecast Real data

Fig. 12. Comparison of real and forecast dynamics of agricultural assets ratio

One can see that IM-model with pairwise multiplications imitates real data much better.

Conclusion

This paper presents some middle-term forecasting models of world dynamics with Forrester´s variables, which were built using GMDH technique. The main results are the followings:

- There is a fundamental possibility to obtain enough accurate forecasts using GMDH with a polynomial-type class of nonlinear difference models;
- Successful modeling needs careful preprocessing including normalization and variable transformation;
- Models with pairwise multiplications of variables prove to be essentially better than the models without them.

In the future we plan to study in detail various classes of models, various GMDH algorithms, and give forecasts of Forrester's variables for future periods.

Bibliography

[Bulakh, 2013] Bulakh V., Perep'olkina L., Sinelnikova O. Models and Estimation Methods for the city in the real estate market. ICIM 2013, 4th international conference on inductive modeling, pp. 251-253.

[Bulgakova, 2013] Bulgakova O. Modeling the Ukraine Black Sea economic region GRP dependence on socio-economic indicators. ICIM 2013, 4th international conference oninductive modeling, pp. 254-260.

[Egorov, 1980] Egorov V., Kallistov N., Mitrofanov V., Piontkovskii A. Mathematical models of global development: a critical analysis of the patterns of nature. - Gidrometeoizdat, 1980. - 192p.

[Forrester, 1979] Forrester J. World Dynamics. Productivity Pr; 2nd edition, 1979, 142 p.

[Ivakhnenko, 1982] Ivakhnenko A.: The inductive method of self-organizing models of complex systems. Kiev, Naukova Dumka, 1982, 296 p.(In Russian)

[Kovalchuk, 2013] Kovalchuk P.I., Gerus A.V., Kovalchuk V.P. Perceptron model of system environmental assessment of water quality in river basins. ICIM 2013, 4th international conference on inductive modeling, pp. 279-284.

[Makhov, 2010] Makhov S. Long-term trends and forecasts from the standpoint of the new global dynamics models / simulations and forecast crises and global dynamics / Ed. Akayev, Korotaev, Malinetskii / Future Russia. - Moscow: LKI, 2010. - P. 262 - 276. (In Russian)

[Matrosov, 1999] Matrosov V., Matrosov I. Global modeling taking into account biomass dynamics and scenarios for sustainable development / New development paradigm forRussia (Comprehensive study of the problems of sustainable development). - Moscow: Academia, MGUK, 1999. - P. 18 - 24. (In Russian)

[Matrosova, 1999] Matrosova K. Sustainable development in the modified mathematical model "World Dynamics" / New development paradigm for Russia (Comprehensive study of the problems of sustainable development). - Moscow: Academia, MGUK, 1999. - P. 344-353. (In Russian)

[Meadows, 1973] Donella M. Meadows, Donella H. Meadows, Dennis L. Meadows, Tzonis. Toward Global Equilibrium: Collected Papers. Productivity Press Inc, 1973, 358 p.

[Pavlov, 2013] Pavlov A., Kondrashova N. Application of generalized relaxational-iterative algorithm for solving the "space weather" forecast problem. ICIM 2013, 4th international conference on inductive modeling, pp. 302-306.

[Samoilenko, 2013] Samoilenko O.A. Combinatorial GMDH Algorithm Application to predict the Dynamic of Demographic Characteristics in Ukraine. ICIM 2013, 4th international conference on inductive modeling, pp. 312-314.

[Stepashko, 2013] Stepashko, V. Ideas of academician O. Ivakhnenko in Inductive Modeling field from historical perspective / Proc. of 4th Intern. Conf. on Inductive Modeling (ICIM-2013), IRTC of NAS of Ukraine, Kyiv, 2013, pp. 31-37.

[Tutova, 2013] Tutova O.,Savchenko Ie. Modeling of impact of macroeconomic indicators on the growth of national income. ICIM 2013, 4th international conference on inductive modeling, pp. 337-343.

[Zubov, 2013] Zubov D. Average daily temperature's long-range forecast using inductive modeling and satellite datasets. ICIM 2013, 4th international conference on inductive modeling, pp. 348-353.

Authors' Information

Olga Proncheva – M.Sc, Russian Presidential Academy of national economy and public administration; Prosp. Vernadskogo 82, bld. 1, Moscow, 119571, Russia; Moscow Institute of Physics and Technology (State University); Institutskii per 9., Dolgoprudny, MoscowRegion, 141700, Russia

e-mail: olga.proncheva@gmail.com

Major Fields of Scientific Research: macroeconomics, mathematical modelling, applied mathematics

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

Major Fields of Scientific Research: data mining, text mining, mathematical modelling

Volodymyr Stepashko –Head of Department for Information Technologies ofInductive Modeling of IRTC ITS, Professor, Dr Sci, P.A.: 40, Akademik Glushkov Prospect,Kyiv, Ukraine, 03680;

e-mail: stepashko@irtc.org.ua

Main Fields of Scientific Research: data mining, knowledge discovery, information technologies,inductive modelling, Group Method of Data Handling (GMDH)

OleksiyKoshulko– *Senior Research Associate, Glushkov Institute of Cybernetics of NAS of Ukraine; Glushkova str. 40, Kyiv, 03680, Ukraine;*

e-mail: koshulko@gmail.com

Major Fields of Scientific Research: data science, parallel processing.

APPLICATION OF DATA MINING TECHNIQUES FOR DIRECT MARKETING

Anatoli Nachev

***Abstract**: This paper presents a case study of data mining modeling techniques for direct marketing. It focuses to three stages of the CRISP-DM process for data mining projects: data preparation, modeling, and evaluation. We address some gaps in previous studies, namely: selection of model hyper-parameters and controlling the problem of under-fitting and over-fitting; dealing with randomness and 'lucky' set composition; the role of variable selection and data saturation. In order to avoid overestimation of the model performance, we applied a double-testing procedure, which combines cross-validation, multiple runs over random selection of the folds and hyper-parameters, and multiple runs over random selection of partitions. The paper compares modeling techniques, such as neural nets, logistic regression, naive Bayes, linear and quadratic discriminant analysis, all tested at different levels of data saturation. To illustrate the issues discussed, we built predictive models, which outperform those proposed by other studies.*

***Keywords**: direct marketing, data mining, modelling, classification, variable selection, neural networks.*

***ACM Classification Keywords**: I.5.2- Computing Methodologies - Pattern Recognition – Design Methodology - Classifier design and evaluation.*

Introduction

Today, banks are faced with various challenges offering products and service to their customers, such as increasing competition, continually rising marketing costs, decreased response rates, at the same time not having a direct relationship with their customers. In order to address these problems, banks aim to select those customers who are most likely to be potential buyers of the new product or service and make a direct relationship with them. In simple words, banks want to select the customers who should be contacted in the next marketing campaigns.

Response modeling is usually formulated as a binary classification problem. The customers are divided into two classes, respondents and non-respondents. Various classification methods (classifiers) have been used for response modeling such as statistical and machine learning methods. They use historical purchase data to train and then identify customers who are likely to respond by purchasing a product.

Many data mining and machine learning techniques have been involved to build decision support models capable of predicting the likelihood if a customer will respond to the offering or not. These models can perform well or not that well depending on many factors, an important of which is how training of the model has been planned and executed. Recently, neural networks have been studied in [Elsalamon & Elsayad, 2013], [Elsalamony, 2014], [Moro et al., 2011], [Yu & Cho, 2006] and regarded as an efficient modelling technique. Decision trees have been explored in [Elsalamon & Elsayad, 2013], [Elsalamony, 2014], [Moro et al., 2011], [Sing'oei & Wang, 2013]. Support vector machines are also well performing models discussed in [Moro et al., 2011], [Shin & Cho, 2006], [Yu & Cho, 2006]. Many other modelling techniques and approaches, both statistical and machine learning, have been studied and used in the domain.

In this paper, we explore five modeling techniques and discuss factors, which affect their performance and capabilities to predict. We extend the methodology used in [Elsalamon & Elsayad, 2013], [Elsalamony, 2014], [Moro et al., 2011] addressing certain gaps.

The remainder of the paper is organized as follows: section 2 provides an overview of the data mining techniques used; section 3 discusses the dataset used in the study, its features, and the preprocessing steps needed to prepare the data for experiments; section 4 presents and discuses the experimental results; and section 5 gives the conclusions.

Data Mining Techniques

We often suspect some relationships among the data we wish to process. However, in order to make more precise statements, draw conclusions, or predict from the measured data, we have to set up a model which represents the nature of the underlying relationship. Here we use several modeling technique, namely: neural networks, logistic regression, naïve Bayes, and linear / quadratic discriminant analysis. This section briefly outlines each.

Neural Networks

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. Among the models, the most common is the multilayer perceptron (MLP), which has a feed-forward topology. 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 θ. The input layer, being the first layer, has size (number of nodes), which corresponds to the size of the input samples. Each hidden and output node computes its activation level by s_i (1) and then transforms it to output by an activation function $f_i(s_i)$. The NN we use in this study works with the logistic activation function (1), where β is slope parameter.

$$s_i = \sum_j w_{ij} x_j + \theta \qquad f_i(s_i) = \frac{1}{1+e^{-\beta s_i}} \tag{1}$$

The overall NN model is given in the form:

$$y_i = f_i(w_{i,\theta} + \sum_{j=I+1}^{I+H} f_j(\sum_{n=1}^{I} x_n w_{m,n} + w_{m,\theta}) w_{i,n}) \tag{2}$$

where y_i is the output of the network for node i, $w_{i,j}$ is the weight of the connection from node j to i and f_j is the activation function for node j. For a binary classification, there is one output neuron with logistic activation function. The training algorithm we use for the MLP is the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm [Broyden , 1970], [Fletcher , 1970]. The BFGS method approximates the Newton's method, a class of hill-climbing optimization techniques. The algorithm stops when the error slope approaches zero or after a maximum of epochs.

Logistic Regression

Logistic regression extends the ideas of linear regression. As with multiple linear regression, the independent variables $x_1, \ldots, x_q$ may be categorical or continuous variables or a mixture of these two types, but the dependent output variable *Y* is categorical, as we use logistic regression for classification. The idea behind logistic regression is straightforward: instead of using Y as the dependent variable, we use a function of it, which is called the *logit*. To understand the logit, we take two intermediate steps: First, we look at *P*, the probability of belonging to class 1, *P* can take any value in the interval [0, 1]. However, if we express *P* as a linear function, it is not guaranteed that the right hand side will lead to values within the interval [0, 1]. The fix is to use a non-linear function of the predictors in the form:

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q)}} \tag{3}$$

This is called the logistic response function. For any values of $x_1, \ldots, x_q$ the right hand side will always lead to values in the interval [0, 1].

The next step is to use a cutoff value on these probabilities in order to map each case to one of the class labels. For example, in a binary case, a cutoff of 0.5 means that cases with an estimated probability of *P(Y = 1) > 0.5* are classified as belonging to class 1, whereas cases with *P(Y = 1) < 0.5* are classified as belonging to class 0. This cutoff need not be set at 0.5.

Naive Bayes

Bayesian classifiers [Clark & Niblett, 1989] operate by using the Bayes theorem, saying that: Let *X* be the data record (case) whose class label is unknown. Let *H* be some hypothesis, such as "data record *X* belongs to a specified class *C*." For classification, we want to determine *P(H|X)* - the probability that the hypothesis *H* holds, given the observed data record *X. P(H|X)* is the posterior probability of *H* conditioned on *X*. Similarly, *P(X|H)* is posterior probability of *X* conditioned on *H. P(X)* is the prior probability of *X*. Bayes theorem is useful in that it provides a way of calculating the posterior probability, *P(H|X)*, from *P(H), P(X)*, and *P(X|H)*. Bayes theorem is

$$P(H \mid X) = \frac{P(X \mid H) P(H)}{P(X)} \tag{4}$$

A difficulty arises when we have more than a few variables and classes - we would require an enormous number of observations (records) to estimate these probabilities. Naive Bayes classification gets around this problem by not requiring that we have lots of observations for each possible combination of the variables. In other words, Naive Bayes classifiers assume that the effect of a variable value on a given class is independent of the values of other variable. This assumption is called class conditional independence. It is made to simplify the computation and in this sense considered to be Naive. This assumption is a fairly strong assumption and is often not applicable, but bias in estimating probabilities often may not make a difference in practice - it is the order of the probabilities, not their exact values, which determine the classifications. Studies comparing classification algorithms have found the Naive Bayesian classifier to be comparable in performance with classification trees and with neural network classifiers. They have also exhibited high accuracy and speed when applied to large databases.

Linear and Quadratic Discriminant Analysis

Discriminant analysis [Fisher, 1936] is a technique for classifying a set of observations into predefined classes. Based on the training set, the technique constructs a set of functions of the predictors, known as discriminant functions. In principle, any mathematical function may be used as a discriminating function. In case of the linear discriminant analysis (LDA), a linear function (5) is used, where x_i are variables describing the data set.

$$y = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_n x_n \tag{5}$$

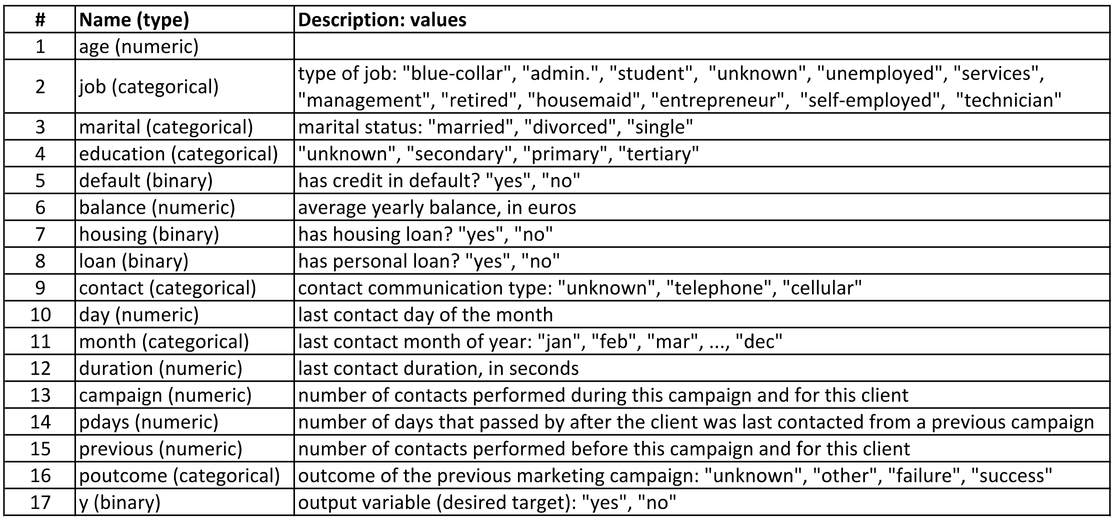
The parameters a_i have to be determined in such a way that the discrimination between the groups is best, which means that the separation (distance) between the groups is maximized, and the distance within the groups is minimized. Quadratic discriminant analysis (QDA) is a generalization of LDA. Both LDA and QDA assume that the observations come from a multivariate normal distribution. LDA assumes that the groups have equal covariance matrices. QDA removes this assumption, allowing the groups to have different covariance matrices. LDA is simpler, faster, and more accurate than QDA, but performing well mainly with linear problems - that is where the decision boundaries are linear. In contrast, QDA is suitable for problems with non-linear decision boundaries.

Dataset and Preprocessing

The direct marketing dataset used in this study was provided by Moro et al. [Moro et al., 2011], also available in [Bache, & Lichman, 2013]. It consists of 45,211 samples, each having 17 attributes (see Table 1), one of which, *y*, s the class label. The attributes are both categorical and numeric and can be grouped as:

- demographical (*age*, *education*, *job*, *marital status*);
- bank information (*balance*, *prior defaults*, *loans*);
- direct marketing campaign information (*contact type*, *duration*, *days* since last contact, *outcome* of the prior campaign for that client, etc.)

Table 1: Dataset attribute names, types, descriptions, and values.



#	Name (type)	Description: values
1	age (numeric)	
2	job (categorical)	type of job: "blue-collar", "admin.", "student", "unknown", "unemployed", "services", "management", "retired", "housemaid", "entrepreneur", "self-employed", "technician"
3	marital (categorical)	marital status: "married", "divorced", "single"
4	education (categorical)	"unknown", "secondary", "primary", "tertiary"
5	default (binary)	has credit in default? "yes", "no"
6	balance (numeric)	average yearly balance, in euros
7	housing (binary)	has housing loan? "yes", "no"
8	loan (binary)	has personal loan? "yes", "no"
9	contact (categorical)	contact communication type: "unknown", "telephone", "cellular"
10	day (numeric)	last contact day of the month
11	month (categorical)	last contact month of year: "jan", "feb", "mar", ..., "dec"
12	duration (numeric)	last contact duration, in seconds
13	campaign (numeric)	number of contacts performed during this campaign and for this client
14	pdays (numeric)	number of days that passed by after the client was last contacted from a previous campaign
15	previous (numeric)	number of contacts performed before this campaign and for this client
16	poutcome (categorical)	outcome of the previous marketing campaign: "unknown", "other", "failure", "success"
17	y (binary)	output variable (desired target): "yes", "no"

There are no missing values. Further details about data collection, understanding, and initial preprocessing steps can be found in [Moro et al., 2011]. Referring to the data understanding stage of the CRISP-DM, we explored each variable distribution using histograms, but in order to understand value distributions in the details, we did marginal distribution plots. They show relationship between any two variables examining the distribution of one, using the other one for grouping. Figure 1 illustrates marginal distribution plots of all 17 variables, using the class variable for grouping. Each variable distribution is represented by two histograms - one belonging to the class label 'no' and one to 'yes'. The last plot corresponding to the output variable *y* shows that the dataset is unbalanced. Indeed, the successful samples corresponding to the class 'yes' are 5,289, which is 11.7% of all samples; all other samples belong to the 'no' class, which is 88.3% of the dataset.

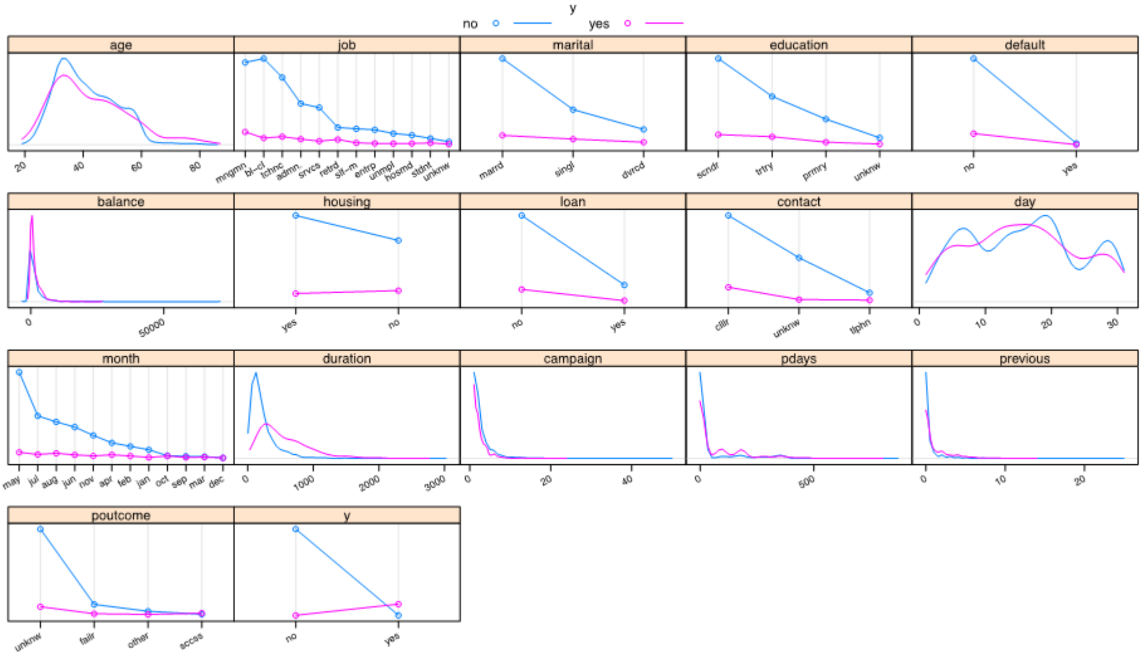


Fig. 1. Marginal distribution plots of all 17 variables, using the class variable y for grouping. Each variable distribution is represented by two histograms - one belonging to the class 'no' and one to 'yes'.

Some modelling techniques, like neural nets, process numeric data only in a fairly limited range, usually [0,1]. This presents a problem, as the dataset we use contains both numeric values out of the usual range and non-numeric. The data transformations needed to sort that out are part of the data preparation stage of the CRISP-DM project model [Chapman et al., 2000]. We did two transformations: mapping non-numeric data to binary dummies and normalization/scaling into the [0,1] interval.

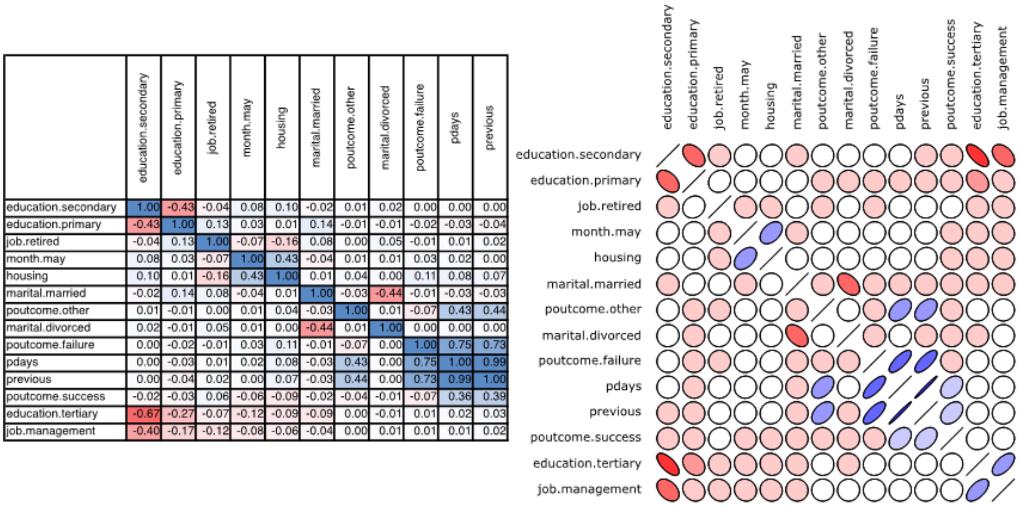
Non-numeric categorical variables cannot be used as they are and must be decomposed into a series of dummy binary variables. For example, a single variable, such as *education* having possible values of "unknown", "primary", "secondary", and "tertiary" would be decomposed into four separate variables: *unknown* - 0/1; *primary* - 0/1; *secondary* - 0/1; and *tertiary* - 0/1. This is a full set of dummy variables, which number corresponds to the number of possible values. In this example, however, only three of the dummy variables need - if the values of three are known, the fourth is also known. Thus, we can map a categorical variable into dummies, which are one less than the number of possible values. Using reduced number of dummies we converted the original dataset variables into 42 numeric variables altogether, which is 6 less than the 48 variables used in [Elsalamon & Elsayad, 2013] and [Elsalamony, 2014]. There are two benefits of that: first, the neural network architecture becomes simpler and faster; secondly, in some modeling algorithms, such as multiple linear regression or logistic regression, the full set of dummy variables will cause the algorithm to fail due to the redundancy.

The second data transformation we did is related to normalization/scaling. This procedure attempts to give all data attributes equal weight, regardless of the different nature of data and/or different measurement units, e.g. *day* (1-31) vs. *duration* in seconds (0-4918). If the data are left as they are, the training process is getting influenced and biased by some 'dominating' variables with large values. In order to address this, we did normalization (z-scoring) and scaling down to [0, 1] by (6):

$$x_i^{new} = \frac{x_i - \mu}{\sigma}, \qquad x_i^{new} = \frac{x_i - a}{b - a} \tag{6}$$

where μ is the mean and σ is the standard deviation of the variable in question; [a,b] is the range of values for that variable. The two transformations were applied to each of the variables independently and separately.

Another part of the data understanding stage of CRISP-DM is to measure correlations between all variables of the dataset. We did pairwise correlation analysis to determine the extent to which values of pairs of variables are proportional to each other, as this may influence the model performance. Correlation coefficients can range in [-1,+1], where -1 represents a perfect negative correlation while +1 represents a perfect positive correlation. A value of 0 represents a lack of correlation. We used Spearman's rho statistic to estimate a rank-based measure of association. This method is robust and recommended if the data do not necessarily come from a bivariate normal distribution. After doing complete analysis of all 42 variables, we further focused to only 11 of them, those with significant correlation coefficients out of the interval [-0.4,+0.4]. Figure 2 shows the correlation coefficients in tabular format and by plot where ellipses converging to circles represent no correlations; the level of stretching of the ellipses shows the level of correlation. Colors and shades also illustrate whether correlations are positive or negative.



	education.secondary	education.primary	job.retired	month.may	housing	marital.married	poutcome.other	marital.divorced	poutcome.failure	pdays	previous
education.secondary	1.00	-0.43	-0.04	0.08	0.10	-0.02	0.01	0.02	0.00	0.00	0.00
education.primary	-0.43	1.00	0.13	0.03	0.01	0.14	-0.01	-0.01	-0.02	-0.03	-0.04
job.retired	-0.04	0.13	1.00	-0.07	-0.16	0.08	0.00	0.05	-0.01	0.01	0.02
month.may	0.08	0.03	-0.07	1.00	0.43	-0.04	0.01	0.01	0.03	0.02	0.00
housing	0.10	0.01	-0.16	0.43	1.00	0.01	0.04	0.00	0.11	0.08	0.07
marital.married	-0.02	0.14	0.08	-0.04	0.01	1.00	-0.03	-0.44	-0.01	-0.03	-0.03
poutcome.other	0.01	-0.01	0.00	0.01	0.04	-0.03	1.00	0.01	-0.07	0.43	0.44
marital.divorced	0.02	-0.01	0.05	0.01	0.00	-0.44	0.01	1.00	0.00	0.00	0.00
poutcome.failure	0.00	-0.02	-0.01	0.03	0.11	-0.01	-0.07	0.00	1.00	0.75	0.73
pdays	0.00	-0.03	0.01	0.02	0.08	-0.03	0.43	0.00	0.75	1.00	0.99
previous	0.00	-0.04	0.02	0.00	0.07	-0.03	0.44	0.00	0.73	0.99	1.00
poutcome.success	-0.02	-0.03	0.06	-0.06	-0.09	-0.02	-0.04	-0.01	-0.07	0.36	0.39
education.tertiary	-0.67	-0.27	-0.07	-0.12	-0.09	-0.09	0.00	-0.01	0.01	0.02	0.03
job.management	-0.40	-0.17	-0.12	-0.08	-0.06	-0.04	0.00	0.01	0.01	0.01	0.02

Fig. 2. Selection of most correlated variables (with absulute correlation coefficints above 0.4) and displayed tabularly and by a plot. Circles reprsent no correlation; ellipses represenit different level of correlation. Colors and shades represent how positive or negative correlations are

This analysis led us to the conclusion that the variables *pdays* and *previous* are in strong correlation, both related to past contacts with the client. There are also correlations between these two variables and the *poutcome* variable, which measures the outcome form previous campaigns. These variables can be identified as candidates for elimination at the modeling stage. We further did sensitivity analysis in order to measure the discriminatory power of each variable and its contribution to the classification task. We found that the least significant variable is *loan*, preceded by *marital*. On the other end, the most significant variables are *duration* and *month*.

Empirical Results and Discussion

In order to build models for direct marketing application and compare their characteristics with those discussed in other studies [Moro et al., 2011], [Elsalamon & Elsayad, 2013], [Elsalamony, 2014], we used the same dataset as used before and did experiments consistently. We, however, extended the methodology addressing the following issues:

- *Optimization of neural network architecture.* Simplifying the NN architectures may lead to better performance, easier to train, and faster NNs.
- *Validation and testing.* Using validation and test sets in a double-testing procedure helps to avoid overestimation of the model performance which is a gap in the studies mentioned above.
- *Randomness and 'lucky' set composition*. Random sampling is a fair way to select training and testing sets, but some 'lucky' draws can train the model much better than others. Using rigorous testing and validation procedures can solidify the conclusions made.
- *Variable selection*. Further to identifying importance of variables and their contribution to the classification task, also reported in the previous studies, we take the next step considering elimination of some input variables, which may lead to improvement of the model performance.
- *Data saturation*. We also explored the capacity of the modelling techniques to act in early stages of data collection where lack of sufficient data may lead to underfitting of the models.

All experiments were conducted using R environment [Cortez, 2010], [R Development Core Team, 2009], [Sing et al., 2005]. The model performances were measured by accuracy as the most common figure of merit, but on the other hand, accuracy can vary dramatically depending on class prevalence, thus being misleading estimator in cases where the most important class is underrepresented - that is our case, because the dataset is unbalanced with underrepresented class 'yes'. In order to address this problem, we used sensitivity, specificity, and ROC analysis [Fawcett, 2005] as more relevant performance estimators.

For the sake of consistency with the previous studies, we used 98 % of the dataset for training and validation, which was split randomly in ratio 2/3: 1/3. The rest of 2% were retained for test. Search of optimal NN architecture was made exploring models with one hidden layer of size from 0 to 13. In order to validate the results and reduce the effect of lucky set composition, each architecture was tested 300 times: internally, the fit algorithm runs 10 times with different random selection of training and validation sets and initial weights. For each of those set compositions, the 3-fold cross-validation creates 3 model instances and averages their results. We iterated all those procedures 10 times per architecture, recording and averaging accuracy and AUC.

H	*ACC*	*ACCmax*	*AUC*	*AUCmax*
0	89.514	92.040	0.895	0.937
1	89.011	92.150	0.898	0.910
2	89.849	93.143	0.903	0.924
3	90.489	93.143	0.906	0.939
4	90.289	91.107	0.908	0.913
5	90.250	90.606	0.910	0.921
6	90.090	90.701	0.912	0.919
7	90.285	90.606	0.913	0.919
8	90.025	90.700	0.915	0.923
9	90.049	90.505	0.915	0.922
10	90.050	90.505	0.915	0.920
11	90.091	90.800	0.913	0.918
12	89.528	90.300	0.913	0.923
13	90.120	90.403	0.914	0.918

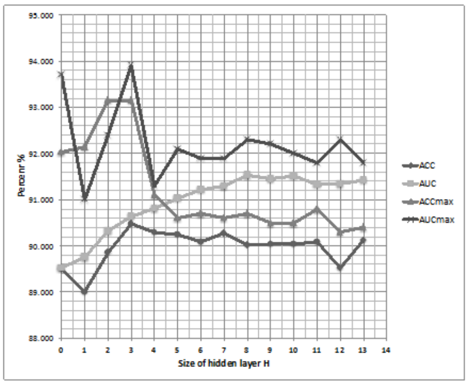


Fig. 3. Validated performance metrics of neural networks with H hidden nodes and architecture 42-H-1, where accuracy (ACC) and area under ROC curve (AUC) values are average of 300 model instances of that architecture. ACC_{max} and AUC_{max} are maximal values obtained.

We also explored how variable selection affects the models performance. Applying backward selection method based on variable significance, we found that eliminating the least important variable *loan* improves the overall model performance. Figure 3 outlines results. NN models with 41-3-1 architecture show best average accuracy of 90.489%. They outperform the 48-20-15-1 architecture from [Elsalamon & Elsayad, 2013]. There were also certain model instances, which show higher accuracy (ACC_{max} in the table of Figure 3). Models with 41-8-1 architecture have best average AUC of 0.915. Certain model instances achieved AUC=0.939. The variance and instability of results can be explained by insufficient saturation of data for training. The model can't be trained well to discriminate between classes, particularly to recognize the under-presented 'yes' class. Nevertheless, experiments show that given the data saturation, a 41-3-1 neural net can be trained to reach accuracy 93.143%, which significantly outperforms the 48-20-15-1 one proposed in [Elsalamon & Elsayad, 2013] and [Elsalamony, 2014].

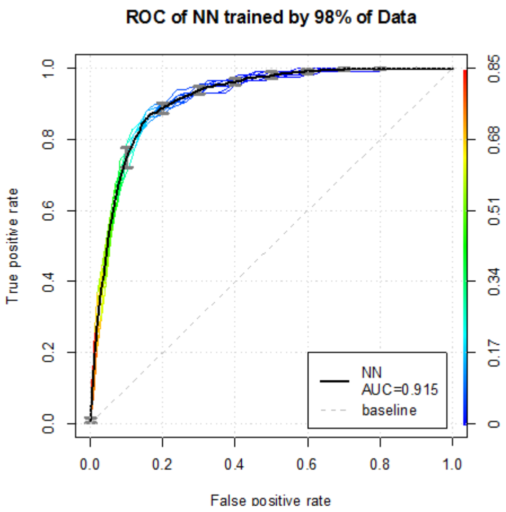


Fig. 4. ROC curves of 10 neural network models with 42-8-1 architecture. Colors sho Fig. 4. ROC curves of 10 neural network models with 42-8-1 architecture. Colors show values of the cutoff points applied. Black line represents average values of the 10 models. Standard deviation bars measure variance.

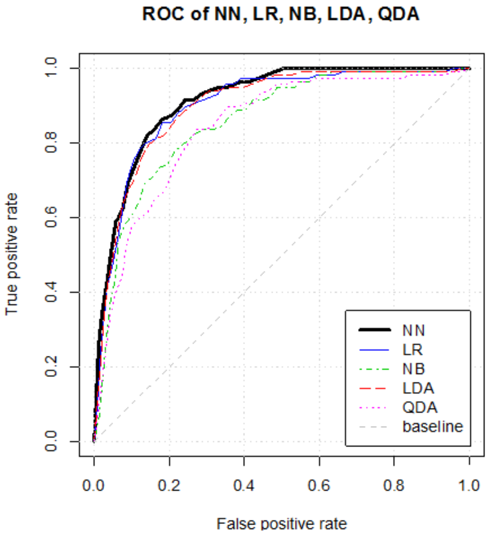


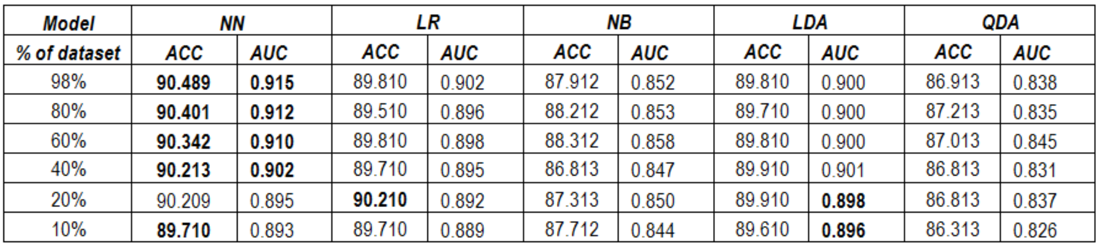
Fig. 5. ROC curve of 5 models: Neural Network, Logistic Regression, Naive Bayes, Linear Discriminant Analysis, and Quadratic Discriminant Analysis. Each model runs with its optimal hyper-parameter values and size of the training dataset

Figure 4 shows ten colored curves, each of which is a plot of a 41-8-1 NN model trained and validated by the 98% dataset. The colors represent different cutoff points with color bar shown on the right side of the box. The black curve is average of the 10 curves. The variance of TPR is depicted by the standard deviation bars.

In order to compare different modeling techniques, we also built models based on logistic regression, naive Bayes, linear discriminant analysis, and quadratic discriminant analysis. Figure 5 shows ROC curves of the models in one plot. Generally, if two ROC curves do not intersect then one model dominates over the other. When ROC curves cross each other, one model is better for some threshold values, and is worse for others. In that situation the AUC measure can lead to biased results and we are not able to select the best model. The figure shows that the curves of NN and LR intersect one another, but in most of the regions NN outperform LR being closer to the top-left corner. This is particularly visible in the most north-west regions, there maximal accuracy is achieved. NN entirely dominate over LDA, NB, and QDA, which performance can be ranked in that order.

Table 2 shows how data saturation affects performance of all 5 models. It can be seen that NN is best performer at nearly all levels of saturation with exception of poorly saturated data (10-20%), where LDA shows better characteristics, particularly measured by AUC.

Table 2: Performance of the five models with different levels of data saturation, ranging from 98% to 10% of the original dataset.



Model	**NN**		**LR**		**NB**		**LDA**		**QDA**	
% of dataset	**ACC**	**AUC**	**ACC**	**AUC**	**ACC**	**AUC**	**ACC**	**AUC**	**ACC**	**AUC**
98%	**90.489**	**0.915**	89.810	0.902	87.912	0.852	89.810	0.900	86.913	0.838
80%	**90.401**	**0.912**	89.510	0.896	88.212	0.853	89.710	0.900	87.213	0.835
60%	**90.342**	**0.910**	89.810	0.898	88.312	0.858	89.810	0.900	87.013	0.845
40%	**90.213**	**0.902**	89.710	0.895	86.813	0.847	89.910	0.901	86.813	0.831
20%	90.209	0.895	**90.210**	0.892	87.313	0.850	89.910	**0.898**	86.813	0.837
10%	**89.710**	0.893	89.710	0.889	87.712	0.844	89.610	**0.896**	86.313	0.826

Conclusion

This paper presents a case study of data mining modeling techniques for direct marketing. We address some issues which we find as gaps in previous studies, namely:

The most common partitioning procedure for training, validation, and test sets uses random sampling. Although, this is a fair way to select a sample, some 'lucky' draws train the model much better than others. Thus, the model instances show variance in behavior and characteristics, influenced by the randomness. In order to address this issue and further to [Moro et al., 2011], [Elsalamon & Elsayad, 2013], [Elsalamony, 2014], we used a methodology, which combines cross-validation (CV), multiple runs over random selection of the folds and initial weights, and multiple runs over random selection of partitions. Each model was tested 300 times involving 3-fold cross-validation, random partitioning and iterations. We applied double-testing with both validation and test sets.

We also explored how NN design affect the model performance in order to find the optimal size of the hidden layer. Given, that the task is a classic binary classification problem without clearly separable feature extraction stages, we found that the two-hidden layers architecture proposed in [Elsalamon & Elsayad, 2013], [Elsalamony, 2014] could be simplified to one hidden layer with structure 41-8-1. The simpler architectures are always preferable as they can be built and trained easily and run faster.

We also did comparatative analysis of neural nets, logistic regression, naive Bayes, linear and quadratic discriminant analysis taking into account their performance at different levels of data saturation. We found that NN is best performer in nearly all levels of saturation with exception of poorly saturated data (10-20%), where LDA shows better characteristics, measured by AUC. We also did comparatative ROC analysis of the models.

Bibliography

[Bache, & Lichman, 2013] Bache, K. & Lichman, M. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2013.

[Broyden , 1970] Broyden, C., The convergence of a class of double rank minimization algorithms: The new algorithm, J. Inst. Math. Appl., 6: 222–231, 1970.

[Chapman et al., 2000] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. CRISP-DM 1.0 - Step-by-step data mining guide, CRISP-DM Consortium, 2000

[Clark & Niblett, 1989] Clark, P. & Niblett, T. The CN2 induction algorithm. Machine Learning, 3, 261-283. 1989.

[Cortez, 2010] Cortez, P. "Data Mining with Neural Networks and Support Vector Machines using the R/rminer Tool". In Proc. of the 10th Industrial Conference on Data Mining (Berlin, Germany, Jul.). Springer, LNAI 6171, 572– 583, 2010.

[Elsalamon & Elsayad, 2013] Elsalamony, H., Elsayad, A., Bank Direct Marketing Based on Neural Network, International Journal of Engineering and Advanced Technology, 2(6):392-400, 2013.

[Elsalamony, 2014] Elsalamony, H. Bank Direct Marketing Analysis of Data Mining Techniques., International Journal of Computer Applications, 85 (7):12-22, 2014.

[Fawcett , 2005] Fawcett, T., An introduction to ROC analysis, Pattern Recognition Letters 27, No.8, 861–874, 2005

[Fisher, 1936] Fisher, R. "The Use of Multiple Measurements in Taxonomic Problems," Annals of Eugenics, 7, 179–188.

[Moro et al., 2011] Moro, S., Laureano, R., Cortez, P., Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimarães, Portugal, October, 2011.

[Fletcher , 1970] Fletcher, R., A new approach to variable metric algorithms, Computer J., 13: 317–322, 1970.

[R Development Core Team, 2009] R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org, 2009.

[Shin & Cho, 2006] Shin, H. J. and Cho, S., Response modeling with support vector machines, Expert Systems with Applications, 30(4): 746-760, 2006.

[Sing et al., 2005] Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T., ROCR: visualizing classifier performance in R., Bioinformatics 21(20):3940-3941, 2005.

[Sing'oei & Wang, 2013] Sing'oei, L., Wang, J., Data Mining Framework for Direct Marketing: A Case Study of Bank Marketing, International Journal of Computer Science Issues (IJCSI), 10(2):198-203, 2013.

[Yu & Cho, 2006] Yu, E. and Cho, S., Constructing response model using ensemble based on feature subset selection, Expert Systems with Applications, 30(2): 352-360, 2006.

Authors' Information

Anatoli Nachev – Business Information Systems, Cairnes Business School, National University of Ireland, Galway, Ireland; e-mail: anatoli.nachev@nuigalway.ie

Major Fields of Scientific Research: data mining, neural networks, support vector machines, adaptive resonance theory.

DISCRIMINATIVE APPROACH TO DISCOVERY IMPLICIT KNOWLEDGE

Sergiy Chalyi, Olga Kalynychenko, Sergiy Shabanov-Kushnarenko, Vira Golyan

Abstract*: The process of finding implicit knowledge based on the well-known DIKW concept ("data - information - knowledge - wisdom") is considered. It has been shown that the transformation of data into information and then into knowledge is carried out by means of the implicit context (implicit knowledge). The DIKW concept is added by elements of implicit knowledge. A general algorithm to identify it is offered. Based on the extended concept of DIKW, a difference approach to identification of implicit knowledge is presented. The approach is based on a comparison of the results of the data mining process for similar processes performed at different times and finding the difference between the graphs obtained. This is considered as formalized implicit dependencies.*

Keywords*: process mining, implicit dependences, explicit dependences, implicit knowledge, explicit knowledge, concept "data - information - knowledge - wisdom".*

ACM Classification Keywords*: I.2.6 Learning - Knowledge acquisition*

Introduction

Representation and use of implicit knowledge currently provides significant opportunities to improve the effectiveness of artificial intelligence systems. The lack of a well-developed concept requires an analysis of approaches to identify and use such knowledge by man directly, and subsequently, the formation of models of representation, allocation of use of implicit knowledge.

The problem of obtaining implicit knowledge (inseparable from a human being) is viewed not only in the field of artificial intelligence, but also in the philosophical, psychological, economic studies. The importance of this problem is connected with the key features of implicit knowledge:

- Inseparability from a human being (knowledge allows to get results without awareness of gaining the result);

- Tight integration with explicit knowledge.

Explicit, formalized knowledge is usually based on a implicit context [Goodman, 2003]. The latter is understood by a person, but its finding is associated with considerable difficulties. Consequently, the completeness of the knowledge description in artificial intelligence systems is achieved by finding of both of explicit and implicit knowledge components that determines the relevance of this study.

The problem of implicit knowledge application in philosophical, psychological, economic studies

Before holding the detailed discussion of the problem and the concept of implicit knowledge finding, it is useful to consider various aspects of the interaction of explicit and implicit components in terms of philosophical, psychological, and economic positions. This consideration makes the case for the importance and the ability to reveal knowledge on the basis of the data sets analysis.

The problem of personal knowledge was discussed in philosophical works by M. Polanyi [Goodman, 2003], [Tandem Computers Inc, 1996]. He singled out this kind of knowledge as man’s inherent advantage over animals. He believed proficiency to be the main characteristic of this knowledge. When finding man’s knowledge we may consider separately "what-knowledge" and "how-knowledge." The first of these is selected and formalized, and the second is implied (in fact, it is implicit) [Goodman, 2003]. In other words, "what-knowledge" usually can be easily explained and documented by man. Knowledge of the second type is characterized by the fact that one can see only the result, and the reasons of its achievement are very difficult to explain. For example, a person can easily solve the problem of face recognition, but he usually finds it difficult to explain his internal recognition algorithm.

Thus, from a philosophical point of view we can highlight the importance of the implicit component, i.e. the "how-knowledge" component (Fig. 1).

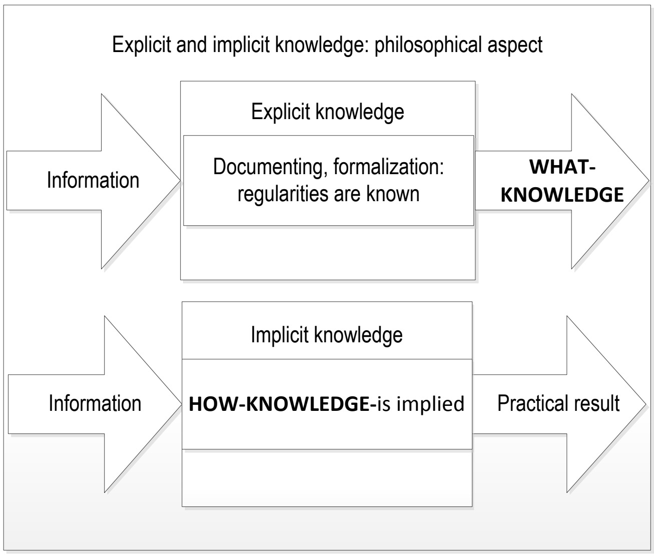


Fig. 1. Philosophical aspect of implicit knowledge

In the field of psychology of intelligence, human intelligence is considered in relation to his knowledge. Simple and hierarchical models are used to describe knowledge in the given area.

In a simple one-layer model intelligence reflects the current level of psychological development of an individual, and it is expressed through a variety of psychological manifestations [Kholodnaya, 2002]. Knowledge in such a model reflects intelligence functionality and can be represented by a set of factors taking into account the types of mental activities, the object with which mental actions have been performed and the final result of the actions [Ngai, 2009]. This simplified model gives an opportunity to get a single set of explicit and implicit dependencies in human knowledge.

Further details of natural intelligence and knowledge are the three-layer hierarchical models. In the papers [Konar, 2000], [Ian, 2011] the following layers are identified as general intelligence; general intelligence capacities; special capacities of a human being. Note that the second layer in this model reflects verbal, numerical, spatial intelligence capacities, and the third layer reflects professional capacities of a human being, namely: algorithmic, technical thinking, math skills, etc. As one can see from the structure of the model, knowledge, in this case, is distributed into layers as follows: the upper layer mostly presents implicit knowledge, the lower layer mostly presents explicit and subject to formalizing knowledge.

The three-layer hierarchical model of natural intelligence [Bondarenko, 2011] offers a different distribution of knowledge at the second layer setting verbal and nonverbal abilities of a human being. At the third layer we distinguish human intellectual capacities which use explicit and implicit knowledge in separate areas of activity. Thus, the current intelligence models specify integral use of explicit and implicit knowledge, the advantage of an explicit (implicit) component being determined by the layer of the model hierarchy, as shown in Fig. 2.

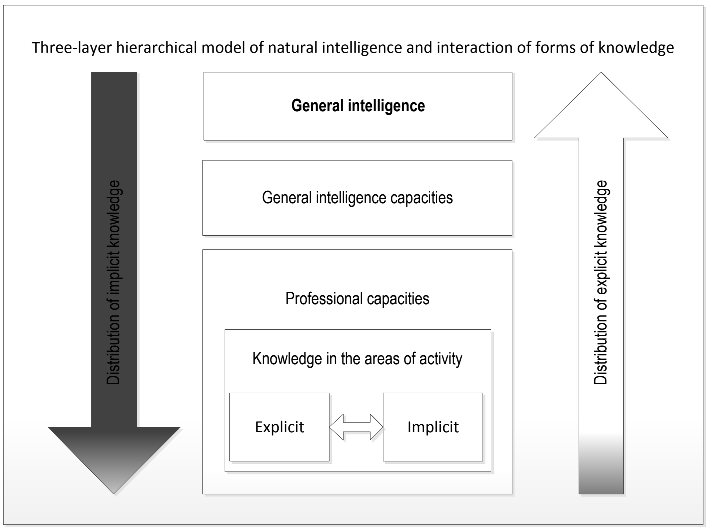

Three-layer hierarchical model of natural intelligence and interaction of forms of knowledge
Distribution of implicit knowledge
General intelligence
General intelligence capacities
Professional capacities
Knowledge in the areas of activity
Explicit
Implicit
Distribution of explicit knowledge

Fig. 2. Distribution of implicit knowledge by using a three-layer model of natural intelligence

In today's economy, knowledge has become more important than the traditional factors of production. The complex of explicit and implicit knowledge forms knowledge of an organization (organizational knowledge). The latter provides enterprise activity both due to document forms of representation, and the knowledge, experience and skills of the employees.

For the first time, the term "organizational knowledge" was proposed by I. Nonaka and H.Takeuchi [Tsuchiya, 1993] in studying the origin and development of innovation in Japanese companies. The authors view organizational knowledge as the knowledge which integrates the totality of knowledge, staff experience at the organizational level as a whole. The knowledge of an organization comprises both formalized corporate knowledge about its functioning, and the implicit knowledge of individual employees (Fig. 3).

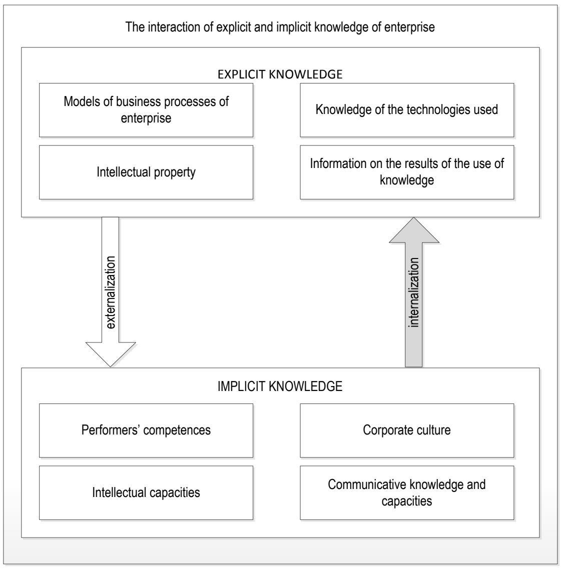

The interaction of explicit and implicit knowledge of enterprise
EXPLICIT KNOWLEDGE
Models of business processes of enterprise
Knowledge of the technologies used
Intellectual property
Information on the results of the use of knowledge
externalization
internalization
IMPLICIT KNOWLEDGE
Performers' competences
Corporate culture
Intellectual capacities
Communicative knowledge and capacities

Fig. 3. The interaction of explicit and implicit knowledge: the economic aspect

Organizational knowledge includes the following components inherent to a human-being:

- Competence of performers which includes education, skills, experience and practical skills to perform employment duties;
- Labor intellectual assets as "total intelligence" of the company employees which are determined by education and qualifications and also depend on prior intellectual activities of employees;
- A common corporate culture that embraces non-formalized set of knowledge on how to perform business processes and interaction between employees;
- Communication resources that cover the knowledge and experience to organize relations with the counterparty of the company.

The explicit component of organizational knowledge in particular includes:

- Formalized description of business processes of the company as well as the general management culture;
- Intellectual property of the company;
- Knowledge resources about managerial, financial, scientific and legal production technologies being used;
- Information resources which record the results of knowledge application in daily activities of the organization.

Thus, the research in the field of economics emphasizes the importance of integrated use of explicit and implicit knowledge, justifying their continuous interaction in the process of economic activity. Moreover, information resources, as components of the explicit knowledge, may reflect the results of using implicit knowledge in the management of the organization.

The conducted analysis of explicit and implicit knowledge in the philosophical, psychological, and economic aspects reveals the following features of their use and interaction (Fig. 4):

- By using explicit knowledge one can highlight a practical result and also to explain, write down and formalize the used regularities;
- By using implicit knowledge one can only get a practical result. To explain the way it is obtained is usually difficult;
- Information about the activity (of an individual, organization) contains "traces" of the application of both explicit and implicit knowledge.

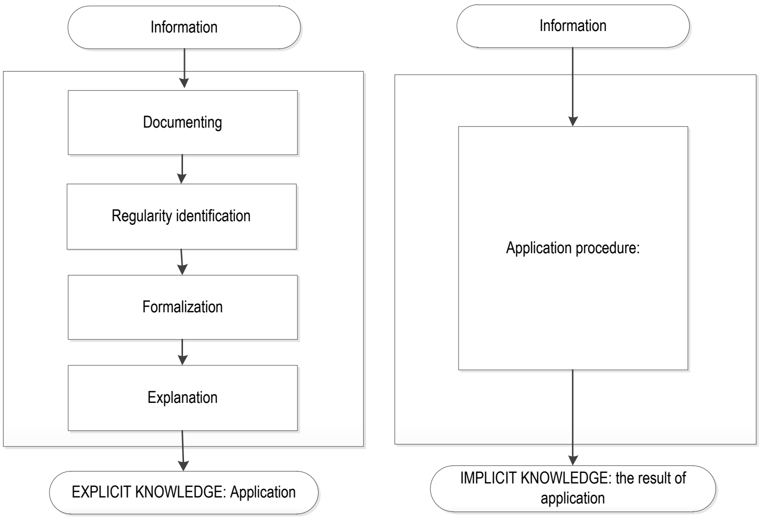

Information
Documenting
Regularity identification
Formalization
Explanation
EXPLICIT KNOWLEDGE: Application
Information
Application procedure:
IMPLICIT KNOWLEDGE: the result of application

Fig. 4. Comparison of the key features of explicit and implicit knowledge application

The given features of implicit knowledge suggest that implicit knowledge can be duplicated traditionally only by means of informal methods (communication, learning combined with the accumulation of human experience), which makes its use difficult in artificial intelligence systems. It is also important to note that the lack of inter-relationships and the presence of only the external manifestations also complicate the formalization of such knowledge.

At the same time it is possible to search for both explicit and implicit knowledge based on information files (data sets), resulting from the application of relevant knowledge.

Separation of explicit knowledge in this case is carried out by data-, process mining techniques. Separation of implicit knowledge by analyzing data sets requires further investigation.

The above mentioned reasoning specifies the relevance of search and formalization of implicit knowledge for future application in artificial intelligence systems.

Problem setting

The above key features of implicit knowledge illustrate the difficulties of its search and formalization in the general case due to the influence of the human factor. At the same time it is possible to extract and formalize the hidden knowledge from the patterned array of data in the problems of data-, process- and web - mining.

This possibility is based on the effect of implicit knowledge on the formation of the dependencies resulting from intelligent analysis. Such dependencies can be displayed in the form of graphs and, in fact, represent explicit knowledge of the processes (structured objects) of the subject domain.

However, the resulting knowledge of identical processes or objects, based on the analysis of data sets during different time intervals, differ in many cases. The reason for these differences is largely determined by the use of formalized, hidden knowledge while executing initial processes and forming corresponding data sets.

These considerations testify that, in principle, it is possible to identify implicit dependencies in the analysis of structured objects (processes) resulting from the data set research.

All this shows the importance of model development problems for finding implicit knowledge based on the analysis of relevant data sets.

DIKW concept as the basic scheme for discovering implicit knowledge

Before further consideration of the problem of discovering implicit knowledge it is necessary to analyze the overall sequence of person’s work with knowledge. This sequence is based on the alternating use of explicit and implicit knowledge. In the learning process there occurs knowledge transformation from one form into another.

Conversion of explicit knowledge in a implicit form is performed by training. In this process, the implicit and explicit knowledge components are initial ones. Explicit knowledge is presented as well-known strategies, technologies and documented materials. Implicit knowledge is owned by people who teach and who are trained. Those who teach help "absorb" presented material; integrate it into the world view existing in student’s mind. As a result, explicit knowledge is converted into skills, abilities and experience. Implicit context is transferred by an individual who teaches and provides (facilitates) learning.

Conversion of knowledge from the explicit form into a implicit one is performed by means of finding and subsequent documenting of the implicit component. At the same time additional explicit knowledge (e.g. formal documentation rules) is used.

Further expansion of the transformation sequence of explicit knowledge into a implicit form and vice versa leads us to the well-known DIKW concept "data - information - knowledge - wisdom [Tsuchiya, 1993]".

In this paper this model is of interest due to the fact that in the process of transformation between the levels of the model implicit knowledge is used. Formalization of such transformations creates conditions for the use of implicit knowledge in artificial intelligence systems.

Let us briefly examine the levels of the given model, adapting them with regard to the peculiarity of the problem to reveal implicit knowledge.

The first level of the model being considered covers the original data sets, which are then converted into information and knowledge. The main features of the given data set are as follows:

- It is a direct result of observations;
- It covers a set of arbitrary symbols whose meaning at this level is not considered;
- It usually has a specific form of representation, so the conversion of this form for further use may be needed.

When considering the problems posed in this paper concerning the use of implicit knowledge in artificial intelligence systems as the first-level data $\{p_1, p_2, ..., p_n\}, i = \overline{1,n}$ of the DIKW model, it is advisable to use databases and structured text files.

The second level – the information level has the following differences from the level of data:

- The links between the data that determine the value of a data set and allow drawing conclusions about the data available are given;
- The possibility of using the data in the current level is not determined.

When using predicate models, the relationship between the data is given in the form of a predicate $I\left(p_1, p_2, ..., p_n\right)$. that determines the structure of information. This predicate can be mapped to a system of binary predicates $I_i, i = \overline{1,m}$, which is represented as a relational network for parallel processing [Mitra, 2003].

The third level - the level of knowledge - has the following features:

- Knowledge integrates information, and it is practically useful.
- Knowledge is presented in the form of individual interconnections, integration of knowledge and the creation of new knowledge at this level is not considered.

Knowledge at this level can be represented as predicates that define the relationship between the elements of previous levels.

The level of wisdom allows us to find fundamentally new understanding of the existing knowledge. From the standpoint of artificial intelligence the level of meta-knowledge can be represented as a second-order predicate (a predicate from predicates) $M(I_1, I_2, ..., I_k)$, given at the set $\{I_i, i = \overline{1,k}\}$.

In some papers meta-knowledge and wisdom are separated; in this case a hierarchy of 5 levels is formed. However, the level of wisdom is inherent to human intellect exclusively. At this level human intelligence operates with abstract values differentiating, for example, between good and evil, bad and kind.

Therefore, in accordance with the problem being solved, we single out 4 levels, and the last level is considered as the level of meta-knowledge.

In the DIKW model we identify a level of understanding as a process of creating new knowledge out of the existing one. The key function of understanding is a function of teaching new knowledge. Understanding allows us not only to create new knowledge, but also to apply this knowledge to perform useful (in the sense of achieving the expected results) actions. As we have seen previously, understanding requires the use of explicit and implicit knowledge. Therefore, understanding is based on the levels of knowledge, information, data, and uses implicit knowledge in the process of transition from one level to another (e.g. the implicit context).

Thus, the DIKW concept expands the previously discussed process of work with explicit and implicit knowledge and, consequently, allows us to pass to the modeling of finding implicit knowledge.

All the above reasoning shows the importance of identifying the role of implicit knowledge in the hierarchy of knowledge, and also requires the formalization of impact of hidden knowledge in the concept of "data - information - knowledge - meta-knowledge". Building such a formal model allows us to justify finding implicit knowledge in general and implicit dependencies in particular based on the analysis of structured data sets.

An important feature of knowledge, which is used in natural intelligence, is to represent knowledge as a process, as opposed to being an object of knowledge in artificial intelligence systems. Interrelation of data, information and knowledge in explicit and implicit forms define a process aspect of knowledge in accordance with the DIKW concept (Fig. 5). Knowledge in the process approach is not only the object of use, but it also can play an active role.

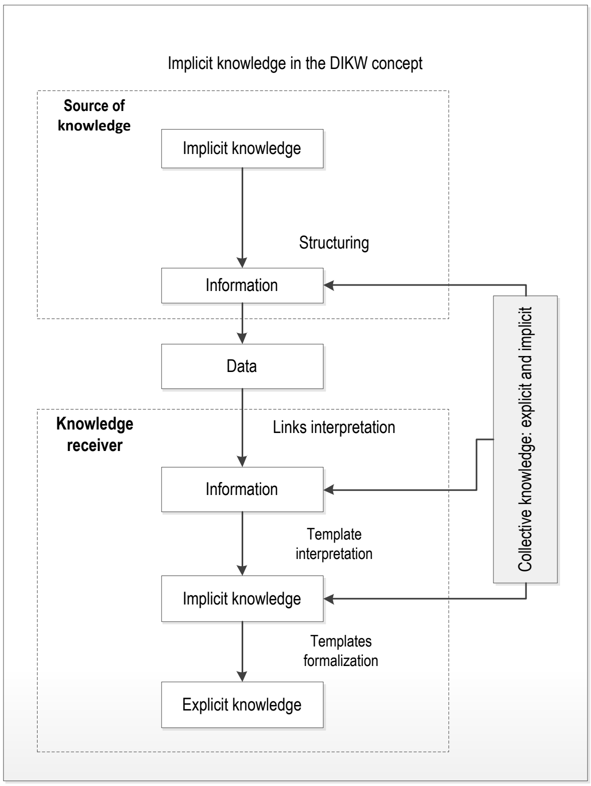


Fig. 5. The DIKW concept and use of implicit knowledge

Thus, the source of knowledge represents its implicit knowledge in a structured way, in the form of information. In the process of structuring, explicit knowledge is used - for example, about the required structure and the form of presenting information. The information obtained contains knowledge in the hidden form. It can be transformed into knowledge only after its interpretation by using additional knowledge, both implicit and explicit.

Then knowledge is transferred as a data set. Indeed, any documented rules, formulas, texts, diagrams, etc. are simply a collection of characters as long as their interpretation is made.

Knowledge receiver performs interpretation of links between data, getting information, then the interpretation of templates obtaining implicit knowledge. In the process of interpretation at this stage a procedural aspect of knowledge is commonly used.

Transformation of knowledge from the implicit into the explicit form is performed by means formalizing previously obtained templates at the last (optional) stage. The given stage completes the transfer of knowledge in natural intelligence. The explicit knowledge obtained can be further used in AI systems.

Note that in accordance with the DIKW concept the patterned array of data (object) corresponds to the information level.

The last key property of implicit knowledge is that it is directly related to performing various actions, such as in technological processes, business processes, design processes, etc. Unlike explicit knowledge, which is separated from an individual and, therefore, can be studied, modified and used at any given time, implicit knowledge manifests itself when fixing operations as part of the above processes, namely in the form of structured data sets.

The important feature of the considered process of knowledge transfer is as follows: for two people to be able to transfer implicit knowledge to each other, they must have collective knowledge (both explicit and implicit). This means that their structuring and interpretation knowledge systems should correspond to each other [Mitra, 2003]. Thus, implicit dependencies can be found by a system of artificial intelligence based on the interpretation of structured data sets under the following conditions:

- Interpretation technique is consistent with the technique, by means of which implicit knowledge has been transformed into a structured data object;
- There is a predefined set of explicit (formalized) dependencies, which were derived from structured data sets.

These features of implicit knowledge lead to the following interim conclusions:

- In natural intelligence explicit and implicit knowledge are complementary. Implicit knowledge provides the formation of explicit knowledge in data and information processing;
- Implicit knowledge is not perceived by man and can be obtained only on the basis of the analysis of actions related to it;
- To discover implicit knowledge it is necessary to use general explicit knowledge concerning the subject domain.

Discriminant model of implicit knowledge finding

The considered pattern of extracting implicit knowledge in terms of the DIKW concept is the basis for the model of finding implicit knowledge from data sets based on process mining results.

The basis of this model is the process representation of knowledge, which describes knowledge as a process and is characterized by the following features:

1) Explicit knowledge representation as a graph that shows a sequence of related activities and events in the subject area.

2) Invariant representation, which means the possibility of its various imaging while maintaining a predetermined interrelation between vertices and arcs.
3) The ability to integrate knowledge presented in the form of individual processes, since a group of similar graphs covers explicit knowledge about a class of similar objects.
4) The ability to verify the completeness and consistency of knowledge.

The third feature among the considered properties of process representation of knowledge provides the ability to discover implicit knowledge according to the DIKW concept by performing the following steps (Fig. 6):

1) Formation of a set of standardized processes for a given subject domain $\{P_j\}$, represented as a set of vertices V_j and arcs E_j. Graph vertices of the process reflect its activity and, therefore, in addition to its N_j identities they also can be characterized by a set of A_j attributes: $P_j = \{V_j, E_j\}, V_j = \{N_j, A_j\}, \mathrm{j} = \overline{1, J}.$ The given set displays explicit knowledge about the P_j processes (or about structured objects in general) and is usually available directly as a result of initial process development. Consequently, this step corresponds to the first three levels of the DIKW concept.

2) Gaining knowledge of this or similar processes based on the analysis of data arrays (process log) by means of process mining techniques in the form of the $\{P_{jk}\}$ set. As a result of executing this step for each process P_j we obtain $k-$graphs reflecting the knowledge of each of its $k-$implementations, at that P_{jk} graphs can "slightly" differ from standardized, specified in the first step. These differences may be caused either by incomplete particular implementation as compared to the original model or by dependencies that are not reflected in the graphs of processes and, therefore, are implicit. In general, such implicit dependencies do not allow us to get isomorphic graphs.

3) Finding implicit dependencies. At this step it is necessary to reveal and formalize the differences between the graphs describing identical processes or objects. These differences may represent a formalized part of implicit dependencies. This step corresponds to the expanded DIKW concept shown in Fig. 5. Finding these differences will be formalized later in the discriminant model.

4) Adding a formalized implicit component obtained in the previous step to the process model of knowledge representation.

Original data in the process of finding implicit knowledge in accordance with the above algorithm are the original model of the process P_j and its log L_j. The process consists of a set of activities whose performance is fixed in the form of time-bound events in the process log. The process log consist of a set of traces of these activities: $L_j = \{\mathrm{S}_{jk}\}$. Each S_{jk} trace corresponds to one-shot execution of the $j-$ process and represents a set of events reflecting the sequence of the performed $s^i_{jk} \in S_{jk}$ activities. Thus, the process log integrates all of its documented implementations. The differences between the original and the final models of the process built on the basis of the analysis of each of the process log traces can be divided according to the following classification criteria:

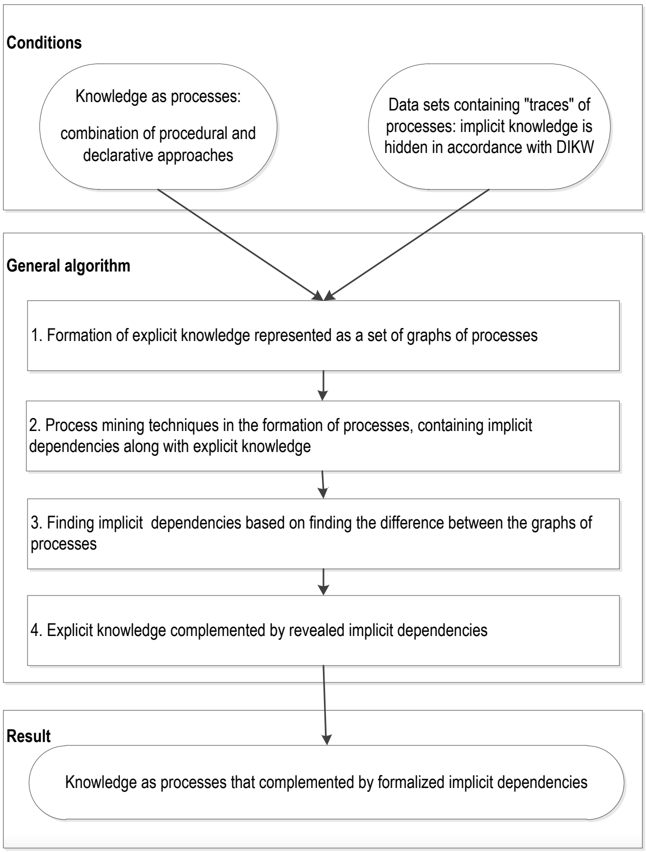

Conditions
Knowledge as processes:
combination of procedural and declarative approaches
Data sets containing "traces" of processes: implicit knowledge is hidden in accordance with DIKW
General algorithm
1. Formation of explicit knowledge represented as a set of graphs of processes
2. Process mining techniques in the formation of processes, containing implicit dependencies along with explicit knowledge
3. Finding implicit dependencies based on finding the difference between the graphs of processes
4. Explicit knowledge complemented by revealed implicit dependencies
Result
Knowledge as processes that complemented by formalized implicit dependencies

Fig. 6. The general pattern of finding implicit knowledge from data sets

the original model of the P_j process contains s^i_{jk} activities that are missing in the final P_{jk} model obtained as a result of the analysis of the $k-$trace:

$$\exists\{s^i_{jk}\}:(s^i_{jk}\in P_j)\wedge(s^i_{jk}\notin P_{jk}), \tag{1}$$

the final P_{jk} model contains activities that are missing in the original P_j model:

$$\exists\{s^i_{jk}\}:(s^i_{jk}\notin P_j)\wedge(s^i_{jk}\in P_{jk}), \tag{2}$$

In the first case, we can say that in the process of specific implementation not all the opportunities available in the original model have been used.

The second case shows the incompleteness of the original model, which may be associated with the use of implicit knowledge in the process implementation.

In accordance with the criterion (2) for the implementation of the considered sequence of steps it is proposed to use the discriminant model of finding implicit knowledge based on pair-wise comparison of P_j and P_{jk} graphs for $k-$execution of the j- process.

Note that the discrepancy between P_j and P_{jk} graphs represents the information level in the DIKW model. Indeed, the differences between the graphs are represented by the subgraph that contains a set of vertices and arcs connecting them which are available in the P_{jk} graph and are missing in the P_j graph. Thus, in the general case, this subgraph contains only a fragment of the original process in the form of structured information. But we are interested in implicit knowledge resulting in such a difference between the graphs. In accordance with the concept discussed above, implicit knowledge is a process that leads to a change in the structure of information. Therefore, to formalize this knowledge it is proposed to find a sequence of transformations of the P_j graph into the P_{jk} graph, if the condition (2) is fulfilled. Then the model of implicit knowledge, which is identified on the basis of differences between the original and final models, is based on the use of the Add operator, which adds the missing vertices to the P_j graph, as well as arcs connected by the given vertices.

Let us define the Add operator as follows:

$$Add(s^i_{jk}) \in P_j \quad \forall s^i_{jk} \left| (s^i_{jk} \notin P_j) \wedge (s^i_{jk} \in P_{jk}) \right. \tag{3}$$

The model of finding implicit knowledge is based on the cyclic elimination of the discrepancy between the original and final graphs of processes.

$$M^{P_{jk}}_{P_j} = \underset{i}{Add}(s^i_{jk}) \left| (s^i_{jk} \notin P_j) \wedge (s^i_{jk} \in P_{jk}) \right. \tag{4}$$

The application result is represented as the process of changing the original P_j graph that reflects the procedural nature of implicit knowledge.

Knowledge representation and the use of implicit dependencies

As it has been shown previously, the use of obtained implicit dependencies is based on the representation of knowledge in the form of graphs of processes. Formalized implicit dependencies complement the graph with new vertices and arcs, reflecting previously hidden cause- and-effect relationships between the individual elements of the process.

As a whole, the process representation of knowledge is the development of script representation. Such representation contains a sequence of frames describing a stereotyped sequence of events taking into account the context and is a way to represent procedural knowledge [Mitra, 2003]. Process representation of knowledge allows us to combine procedural and declarative knowledge and has the following features:

- Knowledge is presented as a set of graphs reflecting possible sequences of tasks (actions), as well as the events that are required to perform these tasks;
- Knowledge integration in the form of processes carried out by means of uniting on the basis of common events;
- Derivation restrictions in the form of additional rules are used.

Thus, in this case, knowledge representation has a two-layer structure combining a flexible scheme of interaction between the fragments of knowledge in the form of processes and a relatively rigid structure of processes (Fig. 7).

At the level of knowledge representation in the form of processes, it is necessary to take into account the starting event of the process, one or more tolerable end events and, probably, the process priority, as well as the rules - restrictions for the process.

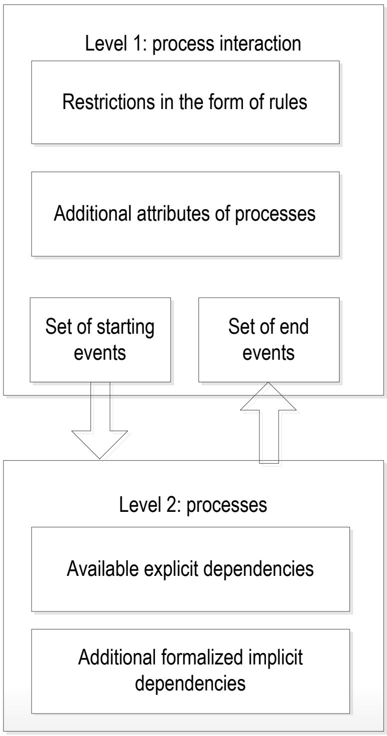

Level 1: process interaction
Restrictions in the form of rules
Additional attributes of processes
Set of starting events
Set of end events
Level 2: processes
Available explicit dependencies
Additional formalized implicit dependencies

Fig. 7. The general structure of knowledge representation in the form of processes and the location of implicit dependencies

At the process level we consider the graph of the process describing the set of its tasks as well as the sequence of their execution. Existence of implicit dependencies is possible at this level of knowledge representation. They are not represented in the process model prior to their identification, which is why there arises a problem of the model adequacy. After finding and formalizing implicit dependencies, they become explicit, form part of the model and do not differ from the previously considered explicit knowledge.

Conclusions

The performed analysis of implicit knowledge has shown that hidden knowledge is characterized only by external manifestations in the process of its application, and, therefore, it is usually identified and copied only with the participation of an individual. This fact complicates its use in artificial intelligence systems.

An approach to finding implicit knowledge based on the DIKW concept is offered. This concept focuses on the use of implicit context when converting data into information and knowledge that has allowed developing a generalized algorithm for finding implicit knowledge using a difference approach to its identification. The approach is based on finding the difference between the original model of the process and the end model, obtained as the result of the process log analysis by means of process mining. If the original model is incomplete compared to the end model, a process that complements the original model is formed, this process can reflect implicit knowledge in a formalized manner.

Bibliography

[Goodman, 2003] Goodman, C.P. The Tacit Dimension. Polanyiana, 2003/1-2. – P. 133-157.

[Tandem Computers Inc, 1996] Knowledge Discovery Through Data Mining: What Is Knowledge Discovery? — Tandem Computers Inc., 1996.

[Kholodnaya, 2002] Kholodnaya M.A. Psyhology of Intelligence. Paradoxes of Research / − SPsB.: Piter, 2002. − 272 c.

[Ngai, 2009] E.W.T. Ngai, Li Xiu, D.C.K. Chau Application of data mining techniques in customer relationship management: A literature review and classification Expert Systems with Applications, Volume 36, Issue 2, Part 2, March 2009, Pages 2592–2602

[Konar, 2000] Konar A., Artificial intelligence and soft computing: behavioral and cognitive modeling of the human brain. – CRC Press LLC. – Boca Raton, Florida/ – 2000

[Ian, 2011]Ian H. Witten, Eibe Frank and Mark A. Hall Data Mining: Practical Machine Learning Tools and Techniques. — 3rd Edition. — Morgan Kaufmann, 2011. — P. 664.

[Han, 2001] Han J., Kamber M., Data mining: Concepts and Techniques. – Morgan Kaufmann Publishers. – 2001.

[Bondarenko, 2011] Bondarenko M.F. About Boolean relational networks / M.F. Bondarenko, I.V. Kameneva, N.Ye. Rusakova, Yu.P. Shabanov-Kushnarenko// Bionics of Intelligence. – 2011. – № 1. – C. 3–7.

[Tsuchiya, 1993] S. Tsuchiya, “Improving knowledge creation ability through organizational learning,” in ISMICK 1993: Proceedings of the International Symposium on the Management of Industrial and Corporate Knowledge, 1993, pp. 87–95.

[Mitra, 2003] Mitra S., Acharya T., Data Mining. Multimedia, Soft Computing, and Bioinformatics. – John Wiley & Sons, Inc. – Hoboken, New Jersey. – 2003

Authors' Information

Sergiy Chalyi – D.Sc, professor of Information Control Systems department, Kharkov National University of Radio Electronics Ukraine; Lenin av., 14, 61166 Kharkov, Ukraine; e-mail: s_chaliy@mail.ru

Sergiy Shabanov-Kushnarenko – D.Sc, professor of Software Engineering department, Kharkov National University of Radio Electronics Ukraine; Lenin av., 14, 61166 Kharkov, Ukraine; e-mail: sergabaev@mail.ru

Olga Kalynychenko - PhD, associate professor of Software Engineering department, Kharkov National University of Radio Electronics Ukraine; Lenin av., 14, 61166 Kharkov, Ukraine; e-mail: okalinichenko@mail.ru

Vira Golyan - PhD, Software Engineering department, Kharkov National University of Radio Electronics Ukraine; Lenin av., 14, 61166 Kharkov, Ukraine; e-mail: veragolyan@yandex.ru

INTELLIGENT ANALYSIS OF MANUFACTURING DATA

Galina Setlak, Monika Piróg-Mazur, Łukasz Paśko

***Abstract**: The article discusses the data analysis conducted for a company of the glass industry. The first part of the paper introduces the examined production process and presents the descriptive statistics used to analyze the manufacturing data. The dataset was collected from measuring points in the process of quality control of products in a given period. In the second part, another dataset is presented. This data was used as a training set for artificial neural networks. Finally, the paper describes the results of research on the possible use of the neural networks for the automatic classification of defects in finished products.*

***Keywords**: artificial neural networks, production process, descriptive statistics.*

***ACM Classification Keywords**: I. Computing Methodologies; I.2 Artificial Intelligence; I.2.6 Learning; Connectionism and neural nets. J. Computer Applications; J.2 Physical Sciences and Engineering; Engineering.*

Introduction

Modern companies operating in a market economy are facing the need to improve product quality, increase productivity and reduce costs, as well as to maximize profits. Apart from physical resources, information have become one of the most important assets of every organization. Therefore every firm that would like to be highly competitive collects data on various aspects of processes it conducts. Because large amount of data collected, in order to improve the quality and reduce costs, the proper methods and tools of data analysis have to be used. Very important task is the extraction of knowledge hidden in the large data sets for supporting business activities. It justifies the efforts to develop new approaches in this area and the use of modern advanced methods and tools of artificial intelligence.

Forecasting plays an important role in the functioning of companies and constitutes an integral part of the production process. It reduces uncertainty and helps to eliminate losses through the decision making process improvements. In the forecasting the mathematical and statistical methods, non-mathematical methods, or artificial neural networks can be used, what facilitates the work, reduces the time-to-market and lowers costs.

Statistical analysis of the data allows the formulation of generalizations based on the obtained results. It also allows the prediction of the events evolution, that is to build forecasts. What is more, it provides tools to organize data about the phenomena, and thus the construction of the overall picture. One of the methods category, one can draw conclusions about the entire population based on the random sample is descriptive statistics [Aczel, 2000]. Descriptive statistics deals with problems of statistical surveys, methods of statistical observation as well as methods of preparation and presentation of statistical properties of the total data set. Using descriptive statistics the biggest set of defects occurring on the production line was analyzed. It is called the "SWA" defect, that gathers defects of deposits and lines on bottles.

The production process in the Glassworks defines the tasks of converting raw materials (blank) into finished product, according to the requirements specified in the project. The development of technological process is a very important stage in the production planning and preparation. It is, however, very difficult automation process due to the large number of engineers involved and the know-how used in the design process.

Technologists' experience significantly affects the technological process and its costs. The technological process together with the auxiliary operations (movements of the material) constitute a manufacturing process in which the final product is obtained. In the manufacturing of glass there are 9 main activities related to the transformation of raw materials into finished products (intended for an external consignee). With such great complexity of production process, the occurrence of defects is possible and therefore it's very important for the company to optimize the whole production process [Dejniak et al., 2011], [Piróg-Mazur et al., 2011].

Major production lines, which include dozens of machines linked together, include the measurement points for quality control purposes. Currently in the Glassworks, data from measuring points are collected using specialized software by the PIC - Production Information Computer. The examination parameters set for the individual elements of bottles consist of: the neck characteristics, the thickness of the wall, contoured body and the bottom. The information generated by the PIC software include: a summary of the losses on the production line, a summary statement for the entire steel mill, a summary of the waste equipment for FP (cold end), the losses in the selected line details, recoil defect data (expressed in percentage terms), kickback data defects in the piece, stationary devices report, a summary of the results of all lines and changing the production line to another [Piróg-Mazur et al., 2011].

The article presents the descriptive statistics of the selected defects and test results concerning the possible use of artificial neural networks for the automatic classification of defects in finished products and their causes. The aim of the study was to create a neural classifier, which task is to classify the defects of the products into three classes, where every class expresses the difficulty degree of the defect elimination (low, medium, high).

Descriptive statistics of data

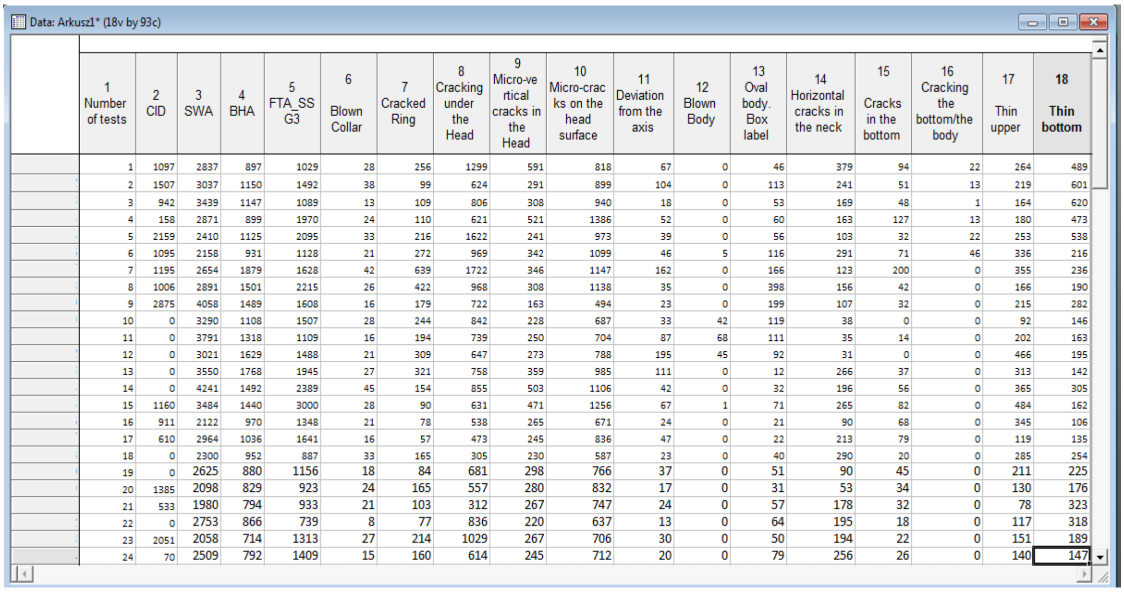
The purpose of descriptive statistics methods is to summarize a set of data and draw some basic conclusions and make generalizations. Descriptive statistics are used as the first and fundamental step in the analysis of the collected data. By entering the analysis of the process we are faced with the choice of an appropriate sample size - a group of representatives. For this reason some statistical concepts can refer to both – the entire population, and the sample (these are called empirical values) [Luszniewicz et al., 2003].

The sample size is 93. It contains data collected during one month, broken down into three shifts. Data describe bottle recoil with characteristics of individual defects (Table 1). Data have been gathered from the measuring points of the production line. In the column headings there are names of defects identified on the production line.

The minimum number of occurrence of defects "SWA" was 952 pieces, and maximally there was made 4241 pieces of recoil in the form of cullet. The arithmetic mean is one of the most intuitive assessment measures of the population. The average of observation set is the sum of all values divided by the number of elements in this set. The average recoil bottles of "SWA" defect was 2168 pieces. The median value in the statistics is a feature value in an ordered set, above and below which there is an equal number of observations values. The median is called the 2-quantile, and the second quartile. The value of median is 2044, what means that half of the observed changes produced no less than 2044 pieces of bottles with "SWA" defect.

Standard deviation is a basic measure of the variability of observed results. It provides information on results of the "change", i.e. whether the spread of results around the mean is small or big. The coefficient of variation shows how strong is the diversity of the data. The standard deviation value is 699, and the coefficient of variation is 32.23 (32%), what shows the moderate variation of bottles production quality at each change. Kurtosis is a measure of the results concentration. Kurtosis informs us about how our observations results are concentrated around the average. This measure tells us how many of our observations results are close to the average. Kurtosis value for the sample is 0.42.

Table 1. Details of the data set for one production line.



Data: Arkusz1* (18v by 93c)

1 Number of tests	2 CID	3 SWA	4 BHA	5 FTA_SSG3	6 Blown Collar	7 Cracked Ring	8 Cracking under the Head	9 Micro-vertical cracks in the Head	10 Micro-cracks on the head surface	11 Deviation from the axis	12 Blown Body	13 Oval body. Box label	14 Horizontal cracks in the neck	15 Cracks in the bottom	16 Cracking the bottom/the body	17 Thin upper	**18 Thin bottom**
1	1097	2837	897	1029	28	256	1299	591	818	67	0	46	379	94	22	264	489
2	1507	3037	1150	1492	38	99	624	291	899	104	0	113	241	51	13	219	601
3	942	3439	1147	1089	13	109	806	308	940	18	0	53	169	48	1	164	620
4	158	2871	899	1970	24	110	621	521	1386	52	0	60	163	127	13	180	473
5	2159	2410	1125	2095	33	216	1622	241	973	39	0	56	103	32	22	253	538
6	1095	2158	931	1128	21	272	969	342	1099	46	5	116	291	71	46	336	216
7	1195	2654	1879	1628	42	639	1722	346	1147	162	0	166	123	200	0	355	236
8	1006	2891	1501	2215	26	422	968	308	1138	35	0	398	156	42	0	166	190
9	2875	4058	1489	1608	16	179	722	163	494	23	0	199	107	32	0	215	282
10	0	3290	1108	1507	28	244	842	228	687	33	42	119	38	0	0	92	146
11	0	3791	1318	1109	16	194	739	250	704	87	68	111	35	14	0	202	163
12	0	3021	1629	1488	21	309	647	273	788	195	45	92	31	0	0	466	195
13	0	3550	1768	1945	27	321	758	359	985	111	0	12	266	37	0	313	142
14	0	4241	1492	2389	45	154	855	503	1106	42	0	32	196	56	0	365	305
15	1160	3484	1440	3000	28	90	631	471	1256	67	1	71	265	82	0	484	162
16	911	2122	970	1348	21	78	538	265	671	24	0	21	90	68	0	345	106
17	610	2964	1036	1641	16	57	473	245	836	47	0	22	213	79	0	119	135
18	0	2300	952	887	33	165	305	230	587	23	0	40	290	20	0	285	254
19	0	2625	880	1156	18	84	681	298	766	37	0	51	90	45	0	211	225
20	1385	2098	829	923	24	165	557	280	832	17	0	31	53	34	0	130	176
21	533	1980	794	933	21	103	312	267	747	24	0	57	178	32	0	78	323
22	0	2753	866	739	8	77	836	220	637	13	0	64	195	18	0	117	318
23	2051	2058	714	1313	27	214	1029	267	706	30	0	50	194	22	0	151	189
24	70	2509	792	1409	15	160	614	245	712	20	0	79	256	26	0	140	147

Below there has been presented a line graph (Figure 1) which is the most common type of statistical graphs. Data are presented with a line, usually broken one. Each point is connected to a line from the first to the last value. This type of chart is used most often for the presentation of the data collected in a given period of time. The x-axis presents a fixed unit of time, while y-axis shows the selected variable – “SWA” defect. With this form, we can determine a variability in the "SWA" in the given period of time. The trend line is always associated with a series of data, but it does not reflect the data in series. The trend line is used to illustrate trends in existing data or to predict future data. An exponential trend line curve is a line used in cases where the values are rising or falling with constantly increasing speed.

One of the basic concepts used in the statistical analysis is stationary variable. Intuitively stationary variable is a variable which properties do not change over time. From the graph in Figure 1, we can see that the graph is non-stationary.

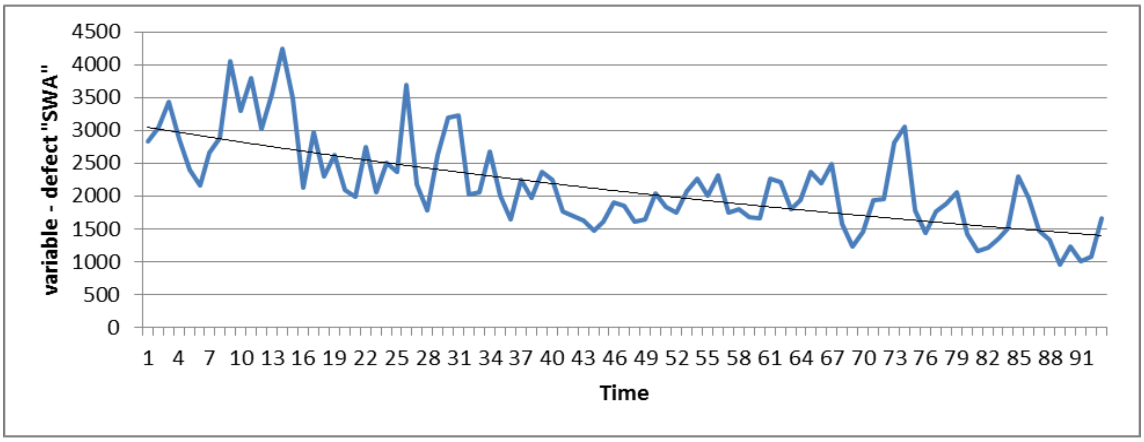
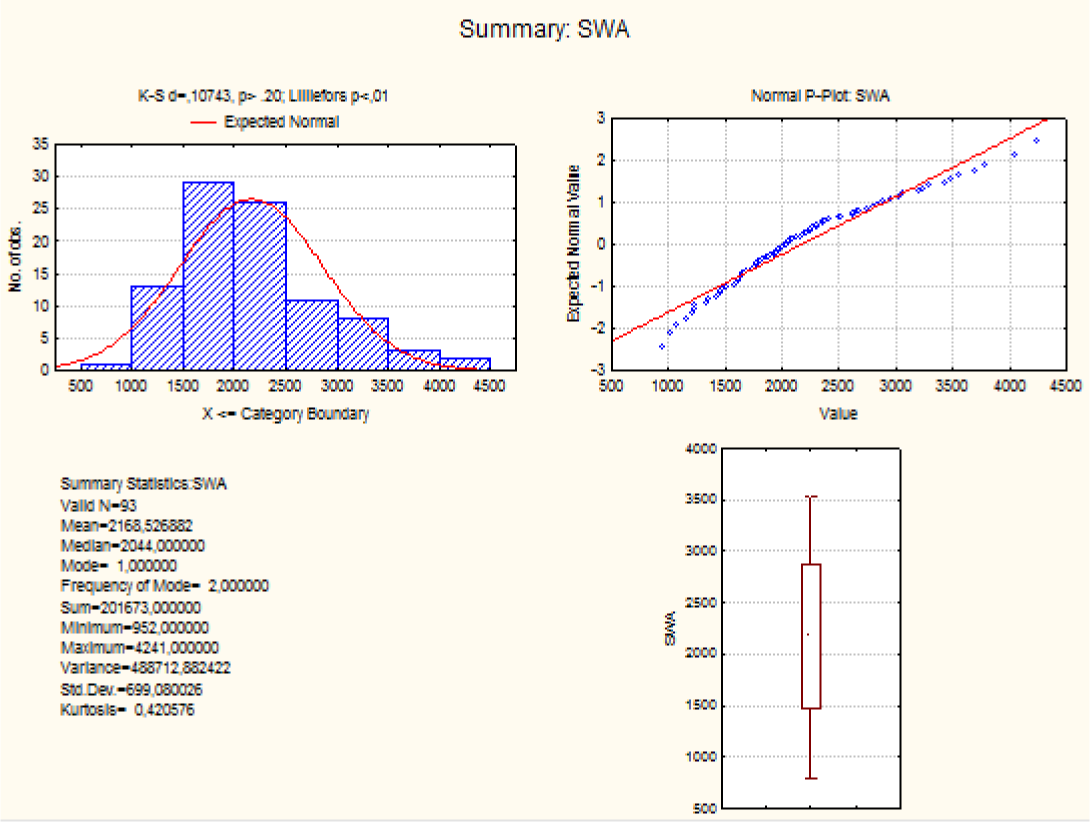


Fig. 1. Line graph of “SWA” defects with the exponential trend line. Source: own work

The histograms (Fig. 2) are the graphic representations of the quantity distribution of the "SWA" variable on which the columns (bars) are plotted over the class intervals, and the height of the column is proportional to the size of classes. The chart facilitates the evaluation of empirical normality (description of the values taken by the characteristic statistical sample using their frequency), because the histogram is applied to fit the curve of the normal distribution. It also allows qualitative evaluation of various aspects of the distribution [Internetowy podręcznik statystyki]. The distribution in this case is unimodal (mode equals 1 - it has one peak), and the frequency of mode equals 2.

The chart of normality for the "SWA" variable from the analyzed data set is given below (Fig. 2). If the points lie close to the straight line graph, and they are uniformly distributed at both sides (e.g., alternately), the data come from a normal distribution.

The box-and-whisker charts (Fig. 2) are developed based on the descriptive statistics, so their use is limited to the numerical characteristics. Most frequently developed charts are those containing median, quartiles, minimum, and maximum. The length of the rectangle represents the middle 50% of observations values. The box is separated by a vertical line (or dot), which determines the value of the median. It divides the quartered section into two areas in which there is 25% of the observations. Whiskers combine a box with the largest and smallest value of the test variable. The first section is 25% of the observations with values below the lower quartile and the second with 25% of the observations with values greater than the upper quartile. The position of the box with respect to the number line shows the position of the distribution. The dot indicates the central tendency of the distribution. The length of the rectangle and the entire chart shows the dispersion of the characteristics in the data set. The proportions on both sides of the vertical line defined by the median value indicate the type of skewness of characteristics distribution (whether it is right- or left sided) [Luszniewicz et al., 2003].



```
Summary Statistics:SWA
Valid N=93
Mean=2168,526882
Median=2044,000000
Mode= 1,000000
Frequency of Mode= 2,000000
Sum=201673,000000
Minimum=952,000000
Maximum=4241,000000
Variance=488712,882422
Std.Dev.=699,080026
Kurtosis= 0,420576
```

Fig. 2. Descriptive statistics of “SWA” defects. Source: own work

In the histogram can be seen numbers of observations (Y- axis, vertical), in the interval (x-axis, horizontal). The size of the compartments is equal. That is, the first pistil describes a number of changes to produce bottles of defect "SWA" from 500 to 1,000. Width of the ranges in the histogram is the same. The difference is the amount of columns (number of observations). In the analyzed example can be seen that most changes have occurred which produced defective bottle in the number of units from 1500 to 2000 and from 2000 to 2500. In chart normality there are shown the minor deviations in the case of points at the top and at the bottom - these points lie further from the straight line than the other points. However, the deviation is so small that the Shapiro-Wilk test does not indicate deviations from a normal distribution. The analyzed box-and-whisker graph shows that the distribution is symmetrical, without outliers.

These statistical graphs are the visualizations of previously conducted statistical analyzes (for example, group data or descriptive statistics).

In the next sections, artificial neural networks are described. The networks have been used to classify defects in finished products in terms of the difficulty degree of the defect elimination.

Artificial neural networks

Artificial neural networks are often used as classifiers, what has been presented in the following papers: [Adamczak, 2001], [Jang et al., 1997], [Moon et al., 1998], [Osowski, 2013], [Setlak, 2000], [Setlak, 2004], [Setlak, Paśko, 2012], [Stąpor, 2011], [Zieliński, 2000]. Neural networks are considered as one of the data mining techniques. The objective of data mining is to find some hidden patterns and relationships that occur in the large sets of data. Discovered patterns can provide a hitherto unknown knowledge, crucial from the point of view of analyzed production process. Data mining is also called intelligent data analysis because it can use the methods and techniques of artificial intelligence [Hand et al., 2005], [Larose, 2006].

In the experiment there have been used only those networks which are fully connected (each neuron of the preceding layer is connected to all the neurons of the next layer). The analyzed networks included one-way connections, without feedback. To train the networks, the supervised learning method has been used.

The classes of artificial neural networks may differ with regard to definition of their activation functions *f(net)*. During the experiment, the following types of neural networks have been examined: linear networks, multilayer perceptrons (MLP) with three kinds of activation function, networks with radial basis functions (RBF), and probabilistic neural networks (PNN).

Linear networks

Linear networks have two layers of neurons: the input and output. The neurons in the input layer serve only to provide input data to the output neurons, without performing any transformations at the same time. Each neuron contained in the output layer, designated as out, comprises a linear activation function, which for the k-th output neuron is described with formulae:

$$f_k^{out}\left(net_k^{out}\right) = net_k^{out}, \tag{1}$$

where net is the sum of neuron's weighted inputs. The function described by the formulae (1) is sometimes called an identity function. Networks only formed from linear neurons have limited capabilities and can be used to solve the simplest problems [Witkowska, 2002].

MLP networks

Multilayer perceptrons (MLP) are the most universal networks commonly used in the classification problems. Such network has at least three layers: input, hidden and output. Like the linear-type network, the input layer does not perform the computational functions. Perceptron can have more than one hidden layer, but in practice it is assumed that most of the problems can be solved using perceptron with one or two hidden layers.

In the performed study it has been analyzed the MLP network, which hidden and output neurons have been equipped with the following activation functions:

- unipolar threshold function:

$$f(net) = \begin{cases} 1 \text{ where } net \geq 0 \\ 0 \text{ where } net < 0 \end{cases}, \tag{2}$$

- logistic function (or unipolar sigmoidal function):

$$f(net) = \frac{1}{1 + \exp(-\beta\, net)}, \tag{3}$$

- hyperbolic tangent function (or bipolar sigmoidal function):

$$f(net) = \frac{\exp(\beta\, net) - \exp(-\beta\, net)}{\exp(\beta\, net) + \exp(-\beta\, net)}. \tag{4}$$

In the formulas (3) and (4) β is a function parameter - the higher the value is the graph of the function is steeper, close to the threshold function [Nałęcz, 2000], [Osowski, 2013], [Tadeusiewicz, 1993].

RBF networks

Radial basis function networks (RBF) are used especially for nonlinear approximation of numerical variables, but they can be also used in classification tasks, where they describe the probability density function of the input variables. RBF networks typically have three layers: input, hidden, and output. Input neurons transfer the signals to hidden neurons that are equipped with radial basis functions. The radial functions are a class of functions which values are decreasing or increasing monotonically with distance from the center of the neuron. Therefore, in contrast to multilayer perceptrons, the RBF's hidden neurons are arranged as centers in the data space. The most frequently used radial function is the Gaussian function. The output layer implements a linear summation of signals from the hidden layer [Nałęcz, 2000].

PNN networks

Probabilistic neural networks (PNN) are used as classifiers. During a training process, PNN networks learn to estimate the probability density function, which is represented by the training data. In the base case, this network contains an input, a radial, and an output layer. The input neurons do not transform the input data. In the radial layer, the number of neurons corresponds to the number of training cases. Each radial neuron is equipped with the Gaussian function centered at the data space over the corresponding training case. The task of the output neurons is to sum the signals coming from the radial layer. After normalization, the calculated sums estimate the probability of belonging of the training case to the output classes [Żurada et al., 1996].

The experiment and the results

The second analyzed data set contained 409 cases. Each case is described using four input variables and one output variable that are presented in Table 2.

Due to the qualitative character of the input variables, each of them has been re-coded by the *one of n* method. Such coding is based on the mapping of a qualitative variable having *n* values, using the *n* input neurons. Each neuron corresponds to a different value of the variable. In the case of occurrence a given value, a corresponding neuron transmits to the network the value 1, and the remaining neurons take the values 0.

The number of neurons in the input layer is determined by the number of variables that are taken into account by the network. The maximum number of neurons occurs when the network uses all four input variables. Taking into account the above-described encoding *one of n*, the number of input neurons in this case is 742.

Table 2. Variables in the second data set. Source: own work

Designation	Description	Type
VAR1	the type of the defect	input
VAR2	the group of defects	input
VAR3	the main cause of the defect	input
VAR4	the way to eliminate the defect	input
VAR5	the difficulty degree of the defect elimination	output

Output layer determines the final outcome of the network; hence the number of neurons in this layer is also closely conditioned by the nature of the problem under consideration. In this experiment, the output variable is the difficulty degree of the defect elimination. The difficulty degree can have one of the following three values:

- low (designation **L**) – the defect is easy to elimination, production line downtime is minimal,
- medium (**M**) – the elimination of the defect requires a longer stop of production process and a greater usage of company's resources,
- high (**H**) – the defect is very difficult to eliminate, the removal of the defect is the most time-consuming.

The difficulty degrees are treated as classes during the classification task. Therefore, each of the examined neural networks includes three output neurons corresponding to three different classes. The object is classified to the class for which the corresponding output neuron reaches the highest value of the activation function.

During the experiment, the following parameters of network architecture were determined:

- the number of hidden layers of networks (none, one or two),
- number of neurons in the hidden layers,
- activation functions of hidden and output neurons.

The above parameters were determined in two ways:

- by using the *Intelligent Problem Solver* options (*IPS*), used for their automatic determination,
- independently modifying parameters until satisfactory results of the network were obtained (only in the case of MLP networks).

IPS option, in addition to creating network architecture, also allows to automatically training the network, and adopting parameters of the learning process selected by the software. In contrast, self-design of MLP network required the selection of an appropriate algorithm and learning parameters that were determined experimentally based on the numerous repetitions of the training process:

- learning method: backpropagation algorithm,

- learning coefficient: 0.1,
- momentum: 0.3,
- number of epochs: 100.

STATISTICA Neural Networks software generated, using the *IPS* option, seven linear types of network, some characteristics of which have been shown in Table 3. Each of them was trained by pseudo-inverse method.

By analyzing the input variables used by the program it can be observed that the most important inputs for the linear network are VAR3 and VAR4. "Knowledge" only of VAR3 variable can correctly classify over 98% of the cases, which shows the LIN1/3 network. The use of only VAR4 results in almost 96% of correct classifications. Using both of these variables in the network LIN2 does not improve the results, even increases the number of misclassifications. To significantly reduce the number of misclassifications, variables VAR1 and VAR2 should be taken into account, what is shown by the results of LIN3 and LIN4 networks.

In addition, Table 3 illustrates another significant correlation. The percentage of correct classifications in LIN3 and LIN4 is very similar, despite the fact that these networks take into consideration the different number of input variables. Comparing these networks it can be concluded that if the variables VAR1, VAR3, VAR4 are available for the linear neural network, the additional use of VAR2 does not significantly affect the outcome of the classification task.

Table 3. Selected parameters of the linear networks. Source: own work

Designation	**The input variables**	**Percentage of correct classifications**	**Number of misclassifications for each class**			
			L	**M**	**H**	**total**
LIN3	VAR1, VAR3, VAR4	99,51%	0	2	0	2
LIN4	VAR1, VAR2, VAR3, VAR4	99,27%	0	3	0	3
LIN2	VAR3, VAR4	94,62%	16	5	1	22
LIN1/1	VAR1	49,14%	41	88	79	208
LIN1/2	VAR2	42,54%	35	106	94	235
LIN1/3	VAR3	98,04%	0	4	4	8
LIN1/4	VAR4	95,60%	1	9	8	18

The second group of examined networks is multilayer perceptrons with one hidden layer. Networks created using *IPS* options differ not only in the number of inputs, but also in the number of hidden neurons. Features of selected perceptrons have been shown in Table 4. The use of perceptrons with one hidden layer allowed increasing the percentage of correct classifications. The best of them make one misclassification. However, as in the case of the linear networks, the largest number of errors concerns the M class.

Looking at the first four networks it can be seen that the number of misclassifications for the perceptrons with three inputs was the same as in the case of the networks with all input variables. This confirms earlier observations noted on the example of the linear networks. Besides, when comparing the MLP 3-42 (three input neurons, 42 hidden neurons) and MLP 3-26 it can be seen that increasing the number of hidden neurons does not change the results of the classification. A similar relationship exists in the network with four inputs: MLP 4-42 and 4-28.

Table 4. Selected parameters of the MLP networks with one hidden layer. Source: own work

Designation	**The input variables**	**Percentage of correct classifications**	**Number of misclassifications for each class**			
			L	**M**	**H**	**total**
MLP 3-42	VAR1, VAR3, VAR4	99,76%	0	1	0	1
MLP 4-42	VAR1, VAR2, VAR3, VAR4	99,76%	0	1	0	1
MLP 3-26	VAR1, VAR3, VAR4	99,76%	0	1	0	1
MLP 4-28	VAR1, VAR2, VAR3, VAR4	99,76%	0	1	0	1
MLP 2-14	VAR3, VAR4	99,27%	1	2	0	3
MLP 1-29	VAR3	98,04%	3	3	2	8
MLP 1-8	VAR1	49,14%	66	80	62	208

Considering the importance of the variables, the program most often selects VAR3 input. All perceptrons (except the last) take into account this variable. Looking at the networks, which use a one variable, it can be seen that the results of the perceptrons are identical as the linear networks. Examples of this are MLP 1-29 and LIN1/3 giving 8 errors, or MLP 1-8 and LIN1/1 with 208 errors. Although the total number of misclassifications is the same, the compared pairs of networks differ in the distribution of errors in the classes.

The third group of the networks consists of perceptrons with two hidden layers of neurons. Characteristics of these perceptrons are described in Table 5. When analyzing the percentage of correct classification it can be seen that the addition of the second hidden layer did not result in reduction in the number of classification errors. Each network uses at least the three inputs committed one error. This error cannot be eliminated by the extension of network architecture. Network MLP 2-6-9 (2 input neurons, 6 in the first hidden layer, 9 in the second hidden layer) can be compared to MLP 2-14. Both perceptrons have variables VAR3 and VAR4. Despite the differences in their architecture, the use of these classifiers gives 3 incorrect classifications.

Comparing the outcomes of MLP 1-4-14 with the results of previous networks using only the VAR1 (MLP 1-8, LIN1/1), it can be concluded that adding more layers and hidden neurons does also not enable to achieve better results. Using only the variable VAR1, results in efficiency of about 50% correct classification. Similarly in the case of networks, which use the VAR3 (MLP 1-19-20, MLP 1-29, LIN1/3) – each of them commits the same number of misclassifications. As is shown, the networks with two hidden layers (Table 5) have identical percentage of correct classifications as the corresponding networks with one hidden layer (Table 4). The difference is only the distribution of errors in each class of the output variable.

Table 5. Selected parameters of the MLP networks with two hidden layers. Source: own work

Designation	**The input variables**	**Percentage of correct classifications**	**Number of misclassifications for each class**			
			L	**M**	**H**	**total**
MLP 3-25-20	VAR1, VAR3, VAR4	99,76%	0	1	0	1
MLP 4-43-36	VAR1, VAR2, VAR3, VAR4	99,76%	0	1	0	1
MLP 3-11-14	VAR1, VAR3, VAR4	99,76%	1	0	0	1
MLP 4-21-21	VAR1, VAR2, VAR3, VAR4	99,76%	1	0	0	1
MLP 2-6-9	VAR3, VAR4	99,27%	0	3	0	3
MLP 1-19-20	VAR3	98,04%	2	4	2	8
MLP 1-4-14	VAR1	49,14%	57	75	76	208

All MLP networks generated by the *Intelligent Problem Solver* contained neurons with logistic activation function. To compare the effects of other activation function, there has been used the opportunity of independent design and training of the networks.

Firstly, 12 perceptrons have been created, which characteristics and results of classification are summarized in Table 6. For analyzing four different network architecture have been selected that varied in the number of input variables, hidden layers, and hidden neurons. Next, in each of these architectures three activation functions have been used: logistic, hyperbolic tangent and threshold. These functions are applied in all hidden and output neurons.

Table 6. Selected parameters of the self-created MLP networks. Source: own work

Designation	**Number of neurons in hidden layers**		**The input variables**	**The activation function**	**Number of misclassifications**
	first	**second**			
MLP 3-42	42	-	VAR1, VAR3, VAR4	Logistic	1
MLP 3-42	42	-	VAR1, VAR3, VAR4	Hyperbolic	1
MLP 3-42	42	-	VAR1, VAR3, VAR4	Step	7
MLP 4-42	42	-	VAR1, VAR2, VAR3, VAR4	Logistic	1
MLP 4-42	42	-	VAR1, VAR2, VAR3, VAR4	Hyperbolic	1
MLP 4-42	42	-	VAR1, VAR2, VAR3, VAR4	Step	205
MLP 3-25-20	25	20	VAR1, VAR3, VAR4	Logistic	1
MLP 3-25-20	25	20	VAR1, VAR3, VAR4	Hyperbolic	1
MLP 3-25-20	25	20	VAR1, VAR3, VAR4	Step	272
MLP 4-43-36	43	36	VAR1, VAR2, VAR3, VAR4	Logistic	1
MLP 4-43-36	43	36	VAR1, VAR2, VAR3, VAR4	Hyperbolic	1
MLP 4-43-36	43	36	VAR1, VAR2, VAR3, VAR4	Step	272

Analyzing the results expressed with number of misclassifications it can be concluded that the used activation function was primarily responsible for the results. All networks using the logistic function and hyperbolic tangent function committed one error. The network with threshold activation function behaved differently. Perceptron MLP 3-42 incorrectly classified 7 cases, however, expanding the network architecture, it also increased the number of errors, which reached up to 272 false classifications of networks with two hidden layers (MLP 3-25-20).

Comparing the number of inputs, which were taken into consideration during learning process, a different relationship can be observed. All perceptrons classify cases identically, regardless of whether they used all of inputs or only three of them. The exception was the mentioned perceptrons with the threshold activation function. Using the three input variables results in 7 classification errors (MLP 3-42), while taking into account all inputs, the number of errors increases to 205 (MLP 4-42).

During training the networks there was also measured the duration of this process. Of course, the results of the measurements also depend on hardware capabilities, and moreover the duration of training with considered amount of data is not a significant parameter, therefore it is not taken into consideration in the analysis. Before starting learning it was assumed that the maximum number of epochs of backpropagation algorithm would amount to 100. For each perceptron algorithm ended learning before it reached that number. However, it is hard to find any relationship between the number of epochs and other network parameters considered in this study, thus the analysis of the number of epochs has also been omitted.

The last group of networks consists of RBF and PNN, whose characteristics are shown in Table 7. By using the IPS option, several RBF networks with different numbers of hidden neurons have been generated (from 68 to 393). Each of these RBF used all input variables to classify. Three selected networks are presented in Table 7.

In case of the RBF networks it can be observed the following relationship: when the number of hidden neurons increases, the number of misclassifications decreases. Such situations did not exist for MLP networks – an increase in the number of neurons did not result in a decrease in classification errors. In addition, the best of established RBF networks committed 8 errors. This result is worse in comparison with the linear and MLP networks.

Table 7. Selected parameters of the RBF and PNN networks. Source: own work

Designation	The input variables	Percentage of correct classifications	Number of misclassifications for each class			
			L	M	H	total
RBF 4-393	VAR1, VAR2, VAR3, VAR4	98,04%	0	8	0	8
RBF 4-237	VAR1, VAR2, VAR3, VAR4	86,06%	6	38	13	57
RBF 4-68	VAR1, VAR2, VAR3, VAR4	55,99%	60	62	58	180
PNN 4-409	VAR1, VAR2, VAR3, VAR4	99,76%	0	1	0	1
PNN 3-409	VAR1, VAR3, VAR4	99,76%	0	1	0	1
PNN 2-409	VAR3, VAR4	99,27%	0	3	0	3
PNN 1-409	VAR3	98,04%	0	5	3	8

In contrast, each PNN network is equipped with 409 hidden neurons, which is consistent with the number of training cases. The *IPS* option created four PNN networks with different number of inputs. The results of PNN are identical to the MLP with one or two hidden layers. The only difference is the distribution of errors in the output classes.

To better compare the behavior of the neural networks, confusion matrices have been prepared (Table 8).

Table 8. Confusion matrices of the selected networks. Source: own work

		Predicted class								
		L	M	H	L	M	H	L	M	H
Actual class	L	159	0	0	159	0	0	159	0	0
	M	2	111	0	1	112	0	8	105	0
	H	0	0	137	0	0	137	0	0	137
Network's designation		LIN3			MLP 3-42 MLP 3-25-20 PNN 4-409			RBF 4-393		

The matrices were created for all tested types of networks, taking into account the classifiers committing the least number of errors. Rows of the matrices correspond to correct classes, while the columns represent classes predicted by the classifier. The matrices have been established for five neural networks. For three of them (MLP 3-42, MLP 3-25-20, PNN 4-409), the matrices are the same. The results show that the misclassifications refer only the cases that belong to the class M. These cases are incorrectly assigned to the class L.

Conclusion

In a manufacturing company the decision-making process related to quality control constitutes an essential and crucial part of the whole production process. Therefore the data collected in a manufacturing process that can support decision-making, must be properly interpreted. At this point, the data mining techniques can be used, which are designed to extract information from the data.

The analysis of data is the starting point for daily discussions about the number of defects generated during production process, causes of the defects, methods to eliminate the defects, and the difficulty degree of the defect elimination.

In the paper a part of the data has been presented, and the descriptive statistics on sample dataset have been calculated, which is the starting point for further research. Finally, artificial neural networks have been used to classify the causes of defects in finished products in terms of the difficulty degree of the defect elimination.

The purpose of the descriptive statistics was to investigate the distribution of the features as well as to estimate the characteristics of this distribution. For the statistical analyzes, the *STATISTICA* software (module *Basic Statistics and Tables*) has been used. To perform the neural networks experiments, the *STATISTICA Neural Networks* software has been applied.

The dataset used for training the neural networks has not be extended with new cases from the future. Therefore, the study did not analyze the ability of generalization of the networks. The analysis focuses only on testing the ability to learn the dataset with qualitative variables.

The best of the established networks has made one misclassification. This error could not be avoided by changing the type of the network, the use of different activation functions, and by modifying the network's architecture.

Comparing types of neural networks used it can be concluded that the problem was solved the best by MLP and PNN networks. Trying to select the network with the lowest number of misclassification and the simplest architecture, the MLP 3-26 should be chosen. Among the examined activation functions in MLP, the logistic function and hyperbolic tangent function allowed to achieve good classification results. The threshold function is not suitable to solve the considered problem.

The obtained results can be treated as analysis of significance of the variables, which indicates the most important inputs for the classification and allows determining the optimal subset of input variables. The most significant inputs for the networks are VAR3 and VAR4 variables. However, the use of only these inputs does not give the best classification. The best results of neural networks have been achieved by taking into account all input variables as well as using only three variables: VAR1, VAR3, and VAR4. These observations are confirmed in the case of networks created with automated *IPS* option and networks generated and learned independently using experimental method. Thus, having only three mentioned inputs, the neural networks are able to classify with almost absolute certainty.

In the further research the attempts to explore knowledge from data using another data mining techniques, such as neuro-fuzzy systems, decision trees and association rules will be undertaken.

Acknowledgements

Research financed by the European Social Fund under the project entitled "Podkarpacki scholarship fund for graduate students".

Bibliography

[Adamczak, 2001] Adamczak R.: Zastosowanie sieci neuronowych do klasyfikacji danych doświadczalnych, PhD thesis, Nicolaus Copernicus University, Toruń, 2001.

[Aczel, 2000] Aczel A.D.: Statystyka w zarządzaniu, PWN, Warszawa, 2000.

[Dejniak et al., 2011] Dejniak D., Piróg-Mazur M.: Elements of forecasting and time series analysis in the process of the glass industry, 6th National Scientific Conference INFORMATION SYSTEMS IN MANAGEMENT, 2011.

[Hand et al., 2005] Hand D., Mannila H., Smyth P.: Eksploracja danych, WNT, Warszawa, 2005.

[Internetowy podręcznik statystyki] Internetowy podręcznik statystyki, www.statsoft.pl/textbook/, © Copyright StatSoft, Inc., 1984-2011.

[Larose, 2006] Larose D.T.: Odkrywanie wiedzy z danych, PWN, Warszawa, 2006.

[Luszniewicz et al., 2003] Luszniewicz A., Słaby T.: Statystyka. Teoria i zastosowania, Wydawnictwo C.H.Beck, Warszawa, 2003.

[Moon et al., 1998] Moon Y.B., Divers C.K., Kim H.-J.: AEWS: An Integrated Knowledge-based System with Neural Network for Reliability Prediction. Computers in Industry, Vol.35, No. 2, 1998, pp. 312-344.

[Nałęcz, 2000] Nałęcz T. (Ed.): Biocybernetyka i Inżynieria Biomedyczna 2000, Tom 6: Sieci neuronowe, Warszawa, 2000.

[Osowski, 2013] Osowski S.: Metody i narzędzia eksploracji danych, Wyd. BTC, Legionowo, 2013.

[Piróg-Mazur et al., 2011] Piróg-Mazur M., Setlak G.: Budowa bazy danych oraz bazy wiedzy dla przedsiębiorstwa produkcyjnego w przemyśle szklarskim, VII Krajowa Konferencja Bazy Danych: Aplikacje i Systemy BDAS'2011, Studia Informatica, Ustroń 2011.

[Piróg-Mazur et al., 2011] Piróg-Mazur M., Setlak G.: Database and Knowledge Base as Integral Part of the Intelligent Decision Support System, Created for Manufacturing Companies. In: Business and Engineering Applications of Intelligent and Information Systems, G. Setlak, K. Markov (Eds.), pp. 202-210, Rzeszów, 2011, ITHEA.

[Setlak, 2004] Setlak G.: Intelligent Decision Support System, LOGOS, Kiev, 2004, (in Rus.).

[Setlak, 2000] Setlak G.: Neural Networks in the Intelligent Information Systems of Production Control, Journal of Automation and Information Sciences, Begell House Inc. Publishers, ISSN Print: 1064-2315, Vol. 32, 2000 Is. 2, pp. 88-93.

[Setlak, Paśko, 2012] Setlak G., Paśko Ł.: Intelligent Analysis of Marketing Data. In: Artificial Intelligence Methods and Techniques for Business and Engineering Applications, Galina Setlak, Mikhail Alexandrov, Krassimir Markov (Eds.), ITHEA®, Rzeszów, Sofia, 2012, pp. 254-275.

[Stąpor, 2011] Stąpor K.: Metody klasyfikacji obiektów w wizji komputerowej, PWN, Warszawa, 2011.

[Zieliński, 2000] Zieliński J.: Inteligentne systemy w zarządzaniu – teoria i praktyka, PWN, Warszawa, 2000.

[Witkowska, 2002] Witkowska D.: Sztuczne sieci neuronowe i metody statystyczne, C.H. Beck Press, Warszawa, 2002.

[Tadeusiewicz, 1993] Tadeusiewicz R.: Sieci neuronowe, Academic Publishing House, Warszawa, 1993.

[Żurada et al., 1996] Żurada J., Barski M., Jędruch W.: Sztuczne sieci neuronowe, PWN, Warszawa, 1996.

Authors' Information

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, Jarosław, Poland, e-mail: 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.*

Monika Piróg-Mazur, **M.Phil.**, 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*

Major Fields of Scientific Research: knowledge representation, decision support system, project management.

Łukasz Paśko, **M.Phil., Eng.** – Rzeszow University of Technology, Department of Computer Science, The Faculty of Mechanical Engineering and Aeronautics, Powstancow Warszawy ave. 8, 35-959 Rzeszow, Poland; e-mail: *lukasz.pasko48@gmail.com*

Major Fields of Scientific Research: artificial intelligence, decision support systems, data mining.

OBJECT-ORIENTED DYNAMIC NETWORKS

Dmytro Terletskyi, Alexandr Provotar

***Abstract**: This paper contains description of such knowledge representation model as Object-Oriented Dynamic Network (OODN), which gives us an opportunity to represent knowledge, which can be modified in time, to build new relations between objects and classes of objects and to represent results of their modifications. The model is based on representation of objects via their properties and methods. It gives us a possibility to classify the objects and, in a sense, to build hierarchy of their types. Furthermore, it enables to represent relation of modification between concepts, to build new classes of objects based on existing classes and to create sets and multisets of concepts. OODN can be represented as a connected and directed graph, where nodes are concepts and edges are relations between them. Using such model of knowledge representation, we can consider modifications of knowledge and movement through the graph of network as a process of logical reasoning or finding the right solutions or creativity, etc. The proposed approach gives us an opportunity to model some aspects of human knowledge system and main mechanisms of human thought, in particular getting a new experience and knowledge.*

***Keywords**: class of objects, inhomogeneous class of objects, sets of objects.*

***ACM Classification Keywords**: E.2 Data Storage Representations — Object representation, D.3.3 Language Constructs and Features — Abstract data types, Classes and objects, Data types and structures, D.1.5 Object-oriented Programming, F.4.1 Mathematical Logic — Set theory.*

Introduction

Modern AI includes many directions, which differ from each other, but at the same time they have something in common. One of the main targets for researchers of all these directions is the development of intelligent information systems (IIS) for solving particular practical problems in corresponding areas. Nowadays, development of IIS is often reduced to heuristic programming. Nevertheless, there is a variety of IIS, which are based on knowledge representation model (KRM). The most famous and common are Semantic Nets, Conceptual Dependency, Frames, Scripts, Logical and Production Models, Ontologies, etc. Each of these models has own specifics and is useful in the particular domain. However, we need to implement a KRM while developing certain IIS. That is why our IIS is going to have at least two levels: the level of KR model and the level of its practical implementation. Sometimes, implementation of certain KRM can cause additional problems and difficulties, which are connected with interaction between two abstraction levels of IIS. This situation led to the development of logical programing and such programming language as Prolog, where KRM is integrated within the language. Such approach gives us an opportunity to represent knowledge using just programming language, because corresponding KRM is integrated within it.

Clearly, the development of software for solving particular tasks, management by some process, etc. is easier than the development of software, which has some level of individuality and intellectuality and can do more complicated tasks than just some computations. That is why questions about useful and powerful tool for such development appear. Modern programming includes many different paradigms, approaches, techniques and programming languages. Object-Oriented Programming (OOP) is one of the famous, useful and powerful programming paradigms nowadays. Indeed, according to [Langpop; Tiobe; Sourseforge] the most popular programming languages in 2013 were languages, which support OOP. Furthermore, many of programming languages have been extended to object-oriented languages: C to C++; Prolog to Object Prolog; COBOL to Object COBOL; SQL to OQL; and LISP to COOL [Sowa, 2000]. That is why questions about usefulness of OOP approach for development IIS, which are based on some KRM, appear.

Object-Oriented Knowledge Representation Models

As it was mentioned before, there are many different types of KR models. However, as in the case with Prolog, if we want to use an OOP language, as a tool for implementation of IIS, we need to use certain object-oriented KR models as a base for them. Nowadays, the most known and developed object-oriented KR models are *frames* and *scripts*, that is why let us consider some their basic and important aspects.

Frames. Generally, frame consists of set of slots, where the corresponding value is assigned for each slot. Every slot has some filler for itself and in such a way provides information about one of the frame's attributes. Furthermore, some of fillers can be frames. According to [Brachman, Levesque, 2004], there are two types of frames: *individual frames,* which are used for representation of single objects, and *generic frames,* which are used for representation of classes of objects. There are special slots as *instance-of*, *is-a*, *a-kind-of*, etc., which help to organize the relations between different frames and types of frames, in particular between individual frames and generic frames, and in such a way to build up the frame system.

In addition, frames have some methods associated with them, which are called *procedures* or *procedural attachments*. Each procedure is a set of some instructions, associated with a frame that can be executed on request. Particular examples of procedural attachments are *slot-reader*, *slot-writer*, etc. Other important procedures are instance constructors, which create instances of classes. Such procedures called when-needed or if-needed procedures and can execute only when they are really required.

Frames have a powerful tool for creating new knowledge called inheritance. It means that frames can inherit the attributes of other frames in the hierarchical structure. Such kinds of slots as instance-of and is-a play an important part in this process, in particular they fill individual frames using other individual or generic frames. Furthermore, frames have such tool as multiple inheritance, which give an opportunity for the frame to inherit properties from more than one other frame.

Analyzing structure of the frames, we can conclude about such their advantages as ability to be represented in the form of a table; to store and use default values in the reasoning process; ability to be structured hierarchically and thus allow easy classification of knowledge; to combine procedural and declarative knowledge using one knowledge representation scheme; to constrain allowed values, or make values be entered within a specific range [Kendal, Creen, 2007]. In addition, frames make semantic nets more powerful by allowing complex objects to be represented as a single frame, rather than as a large network structure. It also provides a common way to represent stereotypic entities, classes, inheritance, and default values [Luger, 2008]. Frame systems tend to have a centralized, conventional control regime, whereas OOP systems have objects acting as small, independent agents sending each other messages, that is why, there can be some applications for which a frame-based system can provide some advantages over a more generic OOP system [Brachman, Levesque, 2004]. Likewise, frames can be used as a data structure for Expert Systems [Coppin, 2004].

However, frames also have some disadvantages. They do not provide the most efficient method to store data for a computer; can lead to "procedural fever"; require care in the design stage to ensure that suitable taxonomies [Kendal, Creen, 2007]. In addition, frames can be considered as a simplified version of a semantic network where only is-a relationships are applied. That is why, the inheritance of default properties of frames in a hierarchy leads to problems with *exceptions* and *multiple inheritance*. First problem arises when a property of a supertype applies to most but not all of its subtypes. The second one arises when a particular subtype may have more than one supertype from which it can inherit properties, which may conflict [Way, 1991; Coppin, 2004].

Scripts. The scripts are structured representation describing stereotyped sequences of events in a particular context. They are frame-like structures for organizing conceptual dependency structures into descriptions of typical situations. Scripts consist of *entry conditions* that must be true for the script to be called; *results* that are true once the script has terminated; *props* that support the content of the script; *roles* that the individual participants perform; *scenes* that presents a temporal aspect of the script.

Frames and scripts are particularly appealing as means for knowledge representation because psychological studies have shown that people tend to rely on knowledge from previous experience whenever possible, and they use this knowledge and adapt it to handle new or slightly different situations. Therefore, instead of analyzing and building descriptions of each new situation as it occurs, people draw on a large collection of structures, which represents their previous experience with objects, people, and situations, and use these past expectations to guide them in analyzing and representing new experiences [Way, 1991].

Nevertheless, scripts, like other KRM have certain problems, in particular the *script match* problem and *between-the-lines* problem. The first problem is that not exists algorithm, which can guarantee correct choices of script in particular situation. The second problem is that not possible to know beforehand the possible occurrences that can break a script. These problems are not unique to script technology but are inherent in the problem of modelling semantic meaning [Graham, Barrett, 1997].

Despite all advantages and disadvantages of frames, which were mentioned above, we can conclude, that frames help us to describe and somehow to represent relations between objects and classes, which are represented through relations with other objects and classes, using inheritance and special slots as *instance-of*, *is-a*, *a-kind-of*, etc. However, such representation of objects and classes does not describe their properties and types without additional information as links with other objects and classes. It means, for representation of some objects we need to represent a lot of other objects and classes, which have higher level in the particular hierarchy of frames. Thus, logical reasoning within particular structure of frames reduces to manipulations with its hierarchy. Nevertheless, questions about what is a starting point in the logical reasoning, how many frames we need for such reasoning, what is superclass in the hierarchy, etc. appear.

Concerning scripts, we can conclude that they have similar to frames nature, but at the same time, they are used for representation of sequences of actions in particular locations. The main feature of scripts is a representation of possible scenarios in the certain locations. However, they do not pay enough attention to features of certain location, in particular objects, which form it. Nevertheless, question about how many related scripts we need for managing by different situations in certain location also appears.

Objects and Classes

We can represent different objects and classes, using frames and scripts. The difference between them is that they represent different types of objects and classes. Frames represent relations between some objects and classes, creating a hierarchy in such a way. Scripts represent actions and relations between them, creating some scenarios. However, both of them do not describe features of their objects and classes and thus do not express fully their semantics. In contrast to frames and scripts, OOP pays more attention to description of features of objects and classes, herewith also creating some hierarchy. That is why, let us consider some features of OOP and try to figure out what an object and a class is.

In OOP object and class are the main concepts. Objects are the building blocks for object-oriented programs. We associate these blocks with the objects of real world, during developing programs. Every object is defined by two terms: attributes and behaviors. Attributes are properties of object, which describe it, and behaviors are procedures, functions (methods) which we can apply to this object and change its state, form and so on [Weisfeld, 2008]. Real world consists of objects, and OOP is the approach for description and simulation of this world or some its particular parts [Pecinovský, 2013].

Let consider such object as "natural number". It is clear that every natural number must be integer and positive. These are characteristic properties of natural numbers. It is obvious, that 12 is really a natural number, but −1 and 7.32, for example, are not natural numbers. While analyzing this fact, we can conclude that each object has certain properties, which define it as some essence. In contrast to OOP, generally properties of objects can be divided into two types – quantitative and qualitative. We are going to define these two types of object's properties formally, but their semantics has intuitive nature.

Definition 1. *Quantitative property of object A is a tuple $p_i(A)=(v(p_i(A)),u(p_i(A)))$ where $i=\overline{1,n}$, $v(p_i(A))$ is an quantitative value of $p_i(A)$ and $u(p_i(A))$ are units of measure of quantitative value of $p_i(A)$.*

Example 1. Suppose we have a car and one of its properties is speed. We can present this property as $p_s(Car)=(v(p_s(Car)),u(p_s(Car)))$ and if our car has speed 150 km/hour, then property is the following $p_s(Car)=(150,km/hour)$.

Definition 2. *Two quantitative properties $p_i(A)$ and $p_j(B)$ where $i=\overline{1,n}$, $j=\overline{1,m}$ are equivalent, i.e. $Eq(p_i(A),p_j(B))=1$, if and only if $u(p_i(A))=u(p_j(B))$.*

Definition 3. *Qualitative property of object A is a verification function $p_i(A)=vf_i(A)$, $i=\overline{1,n}$ which is defined as a mapping $vf_i(A):p_i(A)\to[0,1]$.*

Example 2. Suppose we have a natural number n, and one of its properties is positivity. We can present this property as follows $p_{pos}(n)=vf_{pos}(n)$, where $vf_{pos}(n)$ is verification function of property $p_{pos}(n)$. In this case, function *is defined as a mapping* $vf_{pos}(n):p_{pos}(n)\to\{0,1\}$, and it is a particular case of verification function – predicate or *Boolean-valued function.*

Definition 4. *Two qualitative properties $p_i(A)$ and $p_j(B)$ where $i=\overline{1,n}$, $j=\overline{1,m}$ are equivalent, i.e. $Eq(p_i(A),p_j(B))=1$, if and only if $(vf_i(A)=vf_j(A))\wedge(vf_i(B)=vf_j(B))$.*

Definition 5. *Specification of object A is a vector $P(A)=(p_1(A),...,p_n(A))$ where $p_i(A)$, $i=\overline{1,n}$ is quantitative or qualitative property of object A.*

Definition 6. *Object is a pair $A/P(A)$, where A is object's identifier and $P(A)$ is a specification of object.*

Essentially, object is a carrier of some properties, which define it as some essence, and help us recognize it among other objects.

Definition 7. *Two objects A and B are similar, if and only if they have the same properties and behavior, i.e. $P(A)=P(B)$ and $F(A)=F(B)$.*

Besides properties of objects, we should allocate operations (methods) which we can apply to objects, considering the features of their specifications. That is why, it will be useful to define concept of object's operation (method).

Definition 8. *Operation (method) of object A is a function $f(A)$, which we can apply to object A considering the features of its specification.*

Definition of method of object is similar to corresponding definition in OOP. However, there is a difference between them. Usually methods in OOP are functions, which we can execute for objects. In contrast to OOP, we divide methods of objects on two types, depending on character of their action: *modifiers* and *exploiters*. Modifiers are functions, which can change objects, in particular some fields of objects. Exploiters are functions, which use objects as arguments and cannot change them.

Example 3. Let us consider such objects as natural numbers n, m. Operations of sum and multiplication, i.e. $f_1(n,m) = n + m$ and $f_2(n,m) = nm$ are the simplest examples of exploiters for them.

Example 4. Let us consider integer number k. Incrementation operation $f_1(k) = k + w$ is the simplest example of modifier.

Example 5. Other simple examples of modifiers and exploiters are $get(Object)$ and $set(Object,Value)$ functions, which are common in many OO languages.

Definition 9. *Signature of object A is a vector $F(A) = (f_1(A),...,f_m(A))$, where $f_i(A)$, $i = \overline{1,m}$ is an operation (method) of object A.*

Generally, we can divide objects on concrete and abstract, and does not matter when or how someone has created each particular object. It is implementation of its abstract image – a prototype, which is essentially an abstract specification for creating the future particular objects. In other words, classes are blueprints, which we use as the basis for objects building [Weisfeld, 2008]. In OOP class consists of fields and methods. Fields form specification of class and methods are functions, which we can apply to objects of this class. We will define concept of class of objects using corresponding idea of OO class.

Definition 10. *Class of objects T is a tuple $T = (P(T),F(T))$, where $P(T)$ is abstract specification of some quantity of objects, and $F(T)$ is their signature.*

When we talk about class of objects, we mean properties of these objects and methods, which we can apply to them. In other words, class of objects is a generalized form of consideration of objects and operations on them, without these objects. In OOP every particular object has the same fields and behavior as its class, i.e. it has the same specification and signature. According to this, we can define concept of homogeneous class of objects.

Definition 13. *Homogeneous class of objects T is a class of objects, which contains only similar objects.*

Example 6. The simplest examples of homogeneous classes of objects are class of natural numbers, class of letters of English alphabet, class of colors of the rainbow, etc.

There are many different objects of real world, which belong to different classes, and if we need to work with them, we can describe them, using new separate homogeneous class for each new type of objects. Especially, if we work with not very big quantity of different types of objects, we can do it without any fears. However, if we need to work, for example, with a few thousands of different types or more, just a process of description of such types is very complex and time-consuming not to mention size of code and performance of such programs. Nevertheless, besides homogeneous classes there are inhomogeneous classes of objects, which describe objects of different types within one class. It means that each object of such class can have different properties and methods.

Definition 14. *Inhomogeneous class of objects* T *is a tuple* $T = (Core(T), pr_1(A_1), ..., pr_n(A_n))$, *where* $Core(T) = (P(T), F(T))$ *is the core of class of objects* T, *which includes only properties and methods similar to corresponding properties of specifications* $P(A_1), ..., P(A_n)$ *and corresponding methods of signatures* $F(A_1), ..., F(A_n)$ *respectively, and where* $pr_i(A_i) = (P(A_i), F(A_i))$, $i = \overline{1, n}$ *are projections of objects* $A_1, ..., A_n$, *which consist of properties and methods typical only for these objects.*

Example 7. Let us consider such class of objects as natural numbers N. Clearly that it is a member of such classes as integer numbers, rational numbers and real numbers simultaneously, i.e. $N \in Z \in Q \in R$. As we can see, class R is the biggest in this case. Furthermore, it consists of objects of different types that contradicts concept of OO class. Of course, in programing languages such types are basic and are built in language. However, in OOP we need to use separate class for description of each such class, because different objects from one OO class cannot have different specifications and signatures. That is why we cannot describe such classes of objects using only one OO class.

Operations on Objects and Classes of Objects

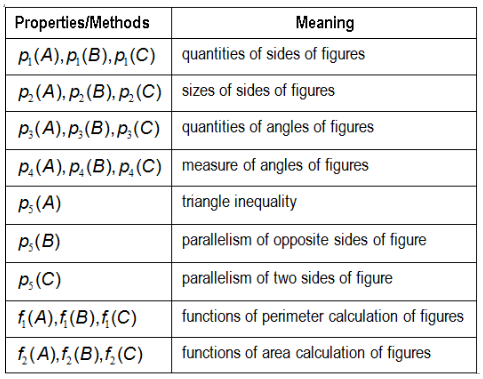
As it was mentioned above, in OOP objects have methods, however, majority of them are local with respect to objects, and cannot be applied to objects of different types, i.e. they are not polymorphic. Of course, there are some methods, which we can apply to objects of different types, but usually we need to use overloading operators for it [Stroustrup, 2013]. Nevertheless, union, intersection, difference, symmetric difference and cloning operations on objects and classes of objects, were proposed in [Terletskyi, Provotar, 2014]. These operations have set-theoretic nature and they are universal in this sense as they can be applied to any objects and classes of objects regardless of their features. We will not concentrate much attention on these operations and just will consider some examples of their using.

Example 8. Let us consider such geometrical objects as triangle, square and trapeze. It is obvious, that these objects belong to different classes of convex polygons. Let us denote triangle as A square as B, trapeze as C and describe their classes as follows

$$T(A) = (P(A), F(A)) = ((p_1(A), ..., p_5(A)), (f_1(A), f_2(A))),$$
$$T(B) = (P(B), F(B)) = ((p_1(B), ..., p_5(B)), (f_1(B), f_2(B))),$$
$$T(C) = (P(C), F(C)) = ((p_1(C), ..., p_5(C)), (f_1(C), f_2(C))).$$

The meaning of each property and method is defined by the Table 1.

Table 1. Meaning of properties and methods of figures A, B, C



Properties/Methods	Meaning
$p_1(A), p_1(B), p_1(C)$	quantities of sides of figures
$p_2(A), p_2(B), p_2(C)$	sizes of sides of figures
$p_3(A), p_3(B), p_3(C)$	quantities of angles of figures
$p_4(A), p_4(B), p_4(C)$	measure of angles of figures
$p_5(A)$	triangle inequality
$p_5(B)$	parallelism of opposite sides of figure
$p_5(C)$	parallelism of two sides of figure
$f_1(A), f_1(B), f_1(C)$	functions of perimeter calculation of figures
$f_2(A), f_2(B), f_2(C)$	functions of area calculation of figures

Specifications and signatures of objects A,B,C can include more or less properties and methods, than we have presented in Table 1, everything depends on the level of detail.

Union. Let us apply the union operation to objects A,B,C and create new set of objects.

$$S = A/T(A) \cup B/T(B) \cup C/T(C) = \{A,B,C\}/T(S)$$

We have obtained a new set of objects S and a new class of objects

$$T(S) = (Core(T(S)), pr_1(A), pr_2(B), pr_3(C)),$$

where $Core(T(S)) = (p_1(T(S)), p_2(T(S)), p_3(T(S)), p_4(T(S)), f_1(T(S)))$, property $p_1(T(S))$ is a quantity of sides of figures, property $p_2(T(S))$ is size of sides of figures, $p_3(T(S))$ is a quantity of angles of figures, $p_4(T(S))$ is measure of angles of figures, method $f_1(T(S))$ is a function of figures' perimeter calculation and $pr_1(A) = (p_5(A), f_2(A))$, $pr_2(B) = (p_5(B), f_2(B))$, $pr_3(C) = (p_5(C), f_2(C))$. Essentially, a set of objects S is the set of triangles of class A, squares of class B and trapezes of class C. Class of objects $T(S)$ describes three types of geometrical figures $T(A)$, $T(B)$ and $T(C)$.

Intersection. Let us calculate intersection of triangle A and square B.

$$A/T(A) \cap B/T(B) = T(A \cap B)$$

As the result, we have obtained new class of objects $T(A \cap B)$, which does not contain any projections of objects, i.e. $T(A \cap B) = (Core(T(A \cap B)))$, where

$$Core(T(A \cap B)) = (p_1(T(A \cap B)), p_2(T(A \cap B)), p_3(T(A \cap B)), p_4(T(A \cap B)), f_1(T(A \cap B))).$$

Meaning of all properties and methods of $Core(T(A \cap B))$ is definitely the same as in the case of union. Class of objects $T(A \cap B)$ describes some type of geometrical figures. However, we do not know exactly which one, but it consists of properties and methods which are simultaneously common for triangle A and square B. Moreover, this class of objects is a part of description of any convex polygon, because each polygon has sides and angles.

Difference. Let us calculate difference of triangle A and trapeze C.

$$A/T(A) \setminus C/T(C) = T(A \setminus C)$$

As the result, we have obtained new class of objects $T(A \setminus C)$, which does not contain core, i.e. $T(A \setminus C) = (pr_1(A))$, where $pr_1(A) = (p_5(A), f_2(A))$. The obtained new class $T(A \setminus C)$, describes, unlike the previous case, the concrete geometric figure - triangle, but using less specification, than given in the Table 1. **Symmetric Difference.** Let us calculate symmetrical difference using the same figures, i.e. triangle A and trapeze C.

$$A/T(A) \div C/T(C) = T(A \div C)$$

As the result, we have obtained a new class of objects $T(A \div C)$, which as in the previous case, does not contain core, i.e. $T(A \div C) = (pr_1(A), pr_2(C))$, where $pr_1(A) = (p_5(A), f_2(A))$ and $pr_2(C) = (p_5(C), f_2(C))$. The class of objects $T(A \div C)$, describes two types of geometrical figures, one of them is a triangle, another one is ambiguously defined.

Cloning. Let us apply cloning operation to triangle A.

$$Clone_1(A) = A_1 / T(A)$$

As the result, we have obtained a new indexed copy of triangle A.

Clearly, that these five operations are exploiters, by using which we can create new classes of objects and sets of objects. It is directly connected to creation of the inhomogeneous classes of objects that means obtaining of new knowledge, in particular classes of these sets. The Example 8 illustrates the basic ideas of these operations in a context of their application to classes of objects. In addition, they can be extended for objects [Terletskyi, 2013].

Creation of new classes of objects, using proposed operations, is directly connected to process of Runtime Class Generation (RCG) in OOP. The main idea of RCG is an opportunity to obtain new classes of objects during program execution. Nowadays, there are few approaches for implementation of this task for some OOP languages, in particular [CGLib] for Java and [CodeDOM] for C#. However, these tools are implemented for such platforms of programming as Java and .NET and based on manipulating with bytecode, which limits their application.

One of the main tools for class creation in OOP is inheritance of basic classes. Furthermore, some of the OOP languages provide multiple inheritance, in particular C++ [Stroustrup, 2013]. However, such approach leads to the same problems, which we have mentioned about frames. That is why we propose another approach to generation of classes of objects. Operations, which were considered in the Example 8, are parts of it. In addition, we will propose another kind of operations, which are modifiers.

Definition 15. *Modification function* $m(p_i(A))$ *is a function, which can change somehow a property* $p_i(A)$.

Definition 16. *Modifier of object* A *is a vector* $M(A)=(m_1(p_1(A)),...,m_n(p_n(A)))$*, where* $m_i(p_i(A))$*,* $i=\overline{1,n}$ *is a modification function of* i*-th property of object* A.

Definition 17. *Modifier of class of objects* T *is a vector*

$$M(T)=(m_1(p_1(T)),...,m_n(p_n(T)),m_1(f_1(T)),...,m_k(f_k(T))),$$

where $m_i(p_i(A))$*,* $i=\overline{1,n}$ *is a modification function of* i*-th property of class of objects* T *and* $m_j(f_j(T))$*,* $j=\overline{1,k}$ *is a modification function of* i*-th methods of class of objects* T

We can divide modifiers on *complete*, *partial*, *generating*, *destroying* and *commutable*, depending on the character of changes of objects or their classes.

Definition 18. *Full modifier of object (class) is an object's (class's) modifier, which can simultaneously modify all properties (properties and methods) of particular object (class).*

Definition 20. *Partial modifier of object (class) is an object's (class's) modifier, which can simultaneously modify some part of properties (properties and methods) of particular object (class).*

Definition 21. *Generating modifier of object (class) is an object's (class's) modifier, which can add some new properties (properties and methods) to specification (specification and signature) of particular object (class).*

Definition 22. *Destroying modifier of object (class) is an object's (class's) modifier, which can destroy some properties (properties and methods) of particular object (class).*

Definition 23. *Commutable modifier of object (class) is an object's (class's) modifier, which can exchange some properties (properties and methods) of particular object (class) to other properties (properties and methods).*

In addition, there are combined modifiers that simultaneously merge a few different types of modification. We propose a way of creating these combinations (look Table 2).

Table 2. Combination table of modifiers.

▨	*F*	*P*	*G*	*D*	*C*
F	▨				
P		▨			
G			▨		
D				▨	
C					▨

Using Table 2, we can create modifiers that have more complex structure and allow us to describe more difficult transformations of objects and classes of objects.

Generally, modifications of objects and classes of objects give us opportunities to create new classes of objects and thus extend OOP paradigm in this direction.

Object-Oriented Dynamic Networks

Taking into account advantages and disadvantages of previously considered object-oriented KR models, we tried to propose new object-oriented KRM, which is based on more detail description of objects and classes of objects than frames or scripts, in a sense it is closer to OOP. Let us define it.

Definition 25. *Object-Oriented Dynamic Network is a 5-tuple* $OODN = (O, C, R, E, M)$*, where:*

- O *– a set of objects;*
- C *– a set of classes of objects, which describe objects from set* O*;*
- R *– a set of relations, which are defined on set* O *and* C*;*
- E *– a set of exploiters, which are defined on set* O *and* C*;*
- M *– a set of modifiers, which are defined on set* O *and* C*.*

As you can see, this model uses concepts of object, object's class and operations on them, which were formulated previously. Concerning set of relations, it can contain any relations between objects, classes, including such as *is-a*, *a-kind-of*, *instance-of*, etc., which are common for semantic nets, frames and scripts. Let us consider some examples for more detail explanation of main principals of OODN.

Example 9. Suppose we have classes of objects $T(P)$, $T(R)$ and $T(S)$ which describe class of polygons, class of rhombuses, class of squares respectively. Let us define these classes as follows

$$T(P) = (P(P), F(P)) = ((p_1(P), \dots, p_4(P)), (f_1(P))),$$
$$T(R) = (P(R), F(R)) = ((p_1(R), \dots, p_5(R)), (f_1(R), f_2(R))),$$
$$T(S) = (P(S), F(S)) = ((p_1(S), \dots, p_6(S)), (f_1(S), f_2(S))).$$

The meaning of each property and method is defined by the Table 3.

Table 3. Meaning of properties and methods of classes of objects $T(P)$, $T(R)$, $T(S)$

Properties/Methods	Meaning
$p_1(P), p_1(R), p_1(S)$	quantities of sides of figures
$p_2(P), p_2(R), p_2(S)$	sizes of sides of figures
$p_3(P), p_3(R), p_3(S)$	quantities of angles of figures
$p_4(P), p_4(R), p_4(S)$	measure of angles of figures
$p_5(R), p_5(S)$	equality of all sides of figure
$p_6(S)$	equality of all angles of figure
$f_1(P), f_1(R), f_1(S)$	functions of perimeter calculation of figures
$f_2(R), f_2(S)$	functions of area calculation of figures

In addition, let us consider particular objects of these classes of objects, i.e. rhombus R_1 and square S_1. The meaning of each property and method is defined by the Table 4 and Table 5 respectively.

Table 4. Specifications of objects R_1, S_1

Rhombus R_1		Square S_1	
Properties	Values	Properties	Values
$p_1(R_1)$	4	$p_1(S_1)$	4
$p_2(R_1)$	2cm, 2cm, 2cm, 2cm	$p_2(S_1)$	3cm, 3cm, 3cm, 3cm
$p_3(R_1)$	4	$p_3(S_1)$	4
$p_4(R_1)$	$70°,110°,70°,110°,$	$p_4(S_1)$	$90°,90°,90°,90°$
$p_5(R_1)$	1	$p_5(S_1)$	1
×	×	$p_6(S_1)$	1

Table 5. Signatures of objects R_1, S_1

Methods	Values	Methods	Values
$f_1(R_1)$	$P(R_1) = 4a$	$f_1(S_1)$	$P(S_1) = 4a$
$f_2(R_1)$	$S(R_1) = d_1 d_2 / 2$	$f_2(S_1)$	$S(S_1) = a^2$

Let us build object-oriented dynamic network for these objects and classes of objects. Clearly, that set of objects is $O = \{R_1, S_1\}$ and set of classes of objects is the following $C = \{T(P), T(R), T(S)\}$. It is obvious, that classes $T(R)$ and $T(S)$ are kinds of class $T(P)$. It is basically know, that the square is a rhombus. According to these facts, we can conclude that sets of relations are the following

$$R = \{R_1 \xrightarrow{\text{instance-of}} T(R), S_1 \xrightarrow{\text{instance-of}} T(S),$$

$$T(R) \xrightarrow{a-kind-of} T(P), T(S) \xrightarrow{a-kind-of} T(P), T(S) \xrightarrow{is-a} T(R)\}.$$

We also can rewrite these relations in the following way

$$R = \{R_1 \in T(R), S_1 \in T(S), T(P) \subseteq T(R), T(P) \subseteq T(S), T(R) \subseteq T(S)\}.$$

Let us define the next set of modifiers $M = \{M_1(T(S)), M_1(T(R)), M_2(T(R)), M_1(T(P)), M_1(R_1)\}$, where: $M_1(T(S))$, $M_1(T(R))$, $M_2(T(R))$, $M_1(T(P))$, transform classes of objects $T(S)$, $T(R)$, $T(R)$, $T(P)$, into classes of objects $T(R)$, $T(L_1)$, $T(S)$, $T(L)$ respectively, $M_1(R_1)$ transforms object R_1 to object L_{1_1}. Definitions of classes $T(L)$ and $T(L_1)$ will be given later. The parts of OODN for objects R_1, S_1 and classes of objects $T(P)$, $T(R)$, $T(S)$ are shown on Figure 1 and Figure 2.

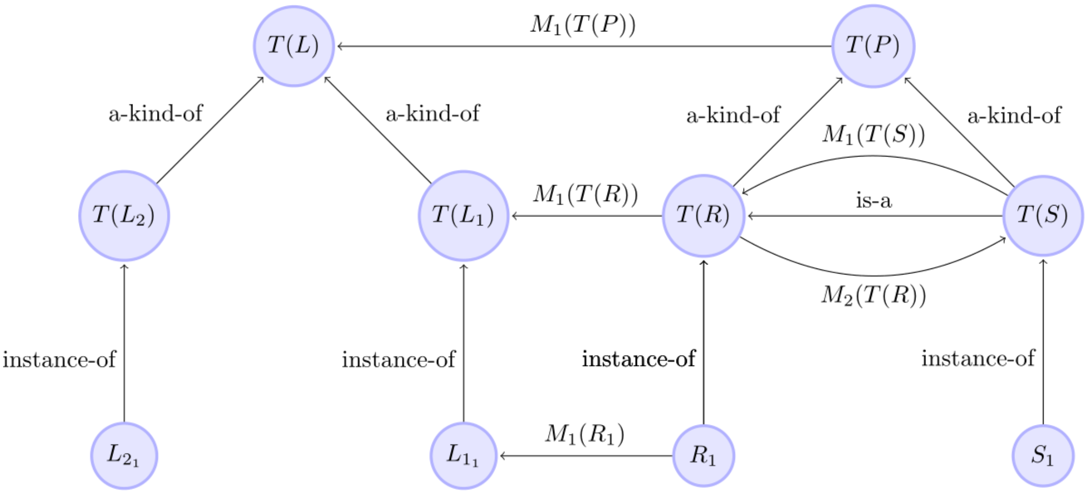


Fig. 1. Part of OODN of polygons: objects, classes, relations and modifiers.

On the Figure 1 the part of OODN's graph, which graphically illustrates the structure of the OODN, is drawn. We can divide this graph on left and right parts. Second of them shows the dependencies between classes of polygons and their modifiers, which transform them into other classes. Let us consider modifiers $M_1(T(S))$ and $M_2(T(R))$ in more detail. The first one modifies class of objects $T(S)$ to class of objects $T(R)$, deleting property $p_6(S)$. The second one modifies class of objects $T(R)$ to class of objects $T(S)$, adding property $p_6(S)$. Clearly that they are inverse to each other. These two modifiers illustrate the process of redrawing of geometrical figures and they are examples of partial-deleting modifier.

Now, let us consider modifiers $M_1(T(R))$, $M_1(T(P))$ and $M_1(R_1)$ in more detail. The first one transforms class of objects $T(R)$ to class of objects $T(L_1)$, which is a class of polylines of type L_1. This transformation occurs, changing property $p_1(R)$, namely $M_1(T(R)) = m_1(p_1(R)) = m_1(4, sizes) = (3, sizes)$. This modifier is an example of partial modifier. Modifier $M_1(T(P))$ is similar to $M_1(T(R))$. It transforms class of objects $T(P)$ to class of objects $T(L)$ – a class of all polylines. The last one – $M_1(R_1)$ modifies object R_1 to object L_{1_1} just as $M_1(T(R))$ modifies $T(R)$ to $T(L_1)$. Analyzing Figure 1, we can conclude that modifiers are kind of transitions between different objects and classes of objects. In such a way, we can model knowledge, which can be modified in time. Furthermore, modifiers form new kind of relations between objects and set of objects.

Generally, these kinds of relations can be represented as *modification-of*. Modifiers can also be considered in temporal context, in particular as future results of modifications.

On the Figure 2 the set of exploiters, which create new objects and classes of objects, using sets O and C, without any changing of their elements, is drawn. Analyzing this figure, we can see that it illustrates whole operations on objects and classes of objects from Example 8.

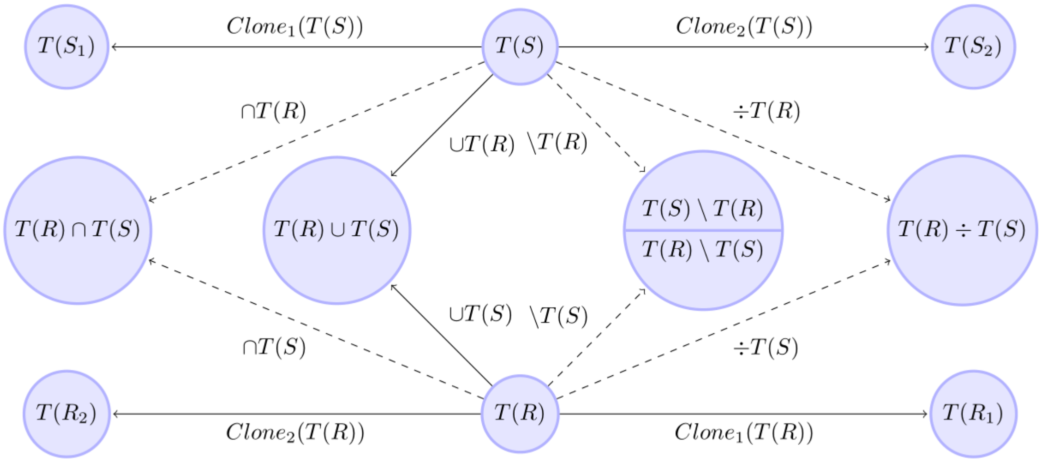


Fig. 2. Part of OODN of polygons: exploiters.

There are some edges, which are drawn by dotted line. It is because results of corresponding operations do not always exist [Terletskyi, Provotar, 2014]. In contrast to modifiers, which act as transitions, exploiters create new classes of objects based on basic classes from set C. In other words, in such a way we can create new classes, which are non-obvious at first glance. Clearly, that there are other types of exploiters, which can be applicable in this example, however there is a question about their generality in comparison with those, which were considered in the Example 8. Analyzing Figure 2, we can conclude that proposed kind of exploiters extend basic set of objects O and classes of objects C. In such a way, they increase description's concentration of particular domain.

Conclusions

This paper contains analysis of such common object-oriented KRM as frames and scripts, which advantages and disadvantages were considered. Furthermore, concepts of object and class were considered from different sides, in particular from OOP's one. The concepts of object and class of objects, which differ from OOP's version, and operations on them, which give us an opportunity to create new sets of objects and new classes of objects, in particular inhomogeneous, were proposed. The operations have set-theoretical nature and are quite general, that gives us a possibility to apply them to any object and class of objects.

The main result of this paper is new object-oriented KRM – Object-Oriented Dynamic Network. It gives us an opportunity to represent knowledge, which can be modified in time, to build new relations between objects and classes of objects and to represent results of their modifications. OODN is based on representation of objects and classes of objects via their properties and methods. It allows us to classify the objects and, in a sense, to build hierarchy of their types. Furthermore, it enables to represent relation of modification between concepts, to build new classes of objects based on existing classes and to create sets and multisets of concepts. Using such model of knowledge representation, we can consider modifications of knowledge and movement through the graph of model as a process of logical reasoning or finding the right solutions or creativity, etc. The proposed approach gives us a possibility to model some aspects of human knowledge system and main mechanisms of human thought, in particular getting a new experience and knowledge. The OODN, in a sense, is similar to OOP languages, but at the same time, it extends classical OOP paradigm, forming new view on the creation of classes of objects. However, despite all advantages, proposed KRM requires further research.

Bibliography

[Langpop] Programming Language Popularity, http://langpop.com.

[Tiobe] TIOBE Programming Community Index, http://www.tiobe.com/index.php/content/paperinfo/tpci/index.html.

[Sourseforge] The Transparent Language Popularity Index, http://lang-index.sourceforge.net.

[Sowa, 2000] J.F. Sowa. Knowledge Representation: Logical, Philosophical and Computational Foundations. – Brooks/Cole, 2000.

[Brachman, Levesque, 2004] R.J. Brachman, H.J. Levesque. Knowledge Representation and Reasoning. – Morgan Kaufmann Publishers, 2004.

[Kendal, Creen, 2007] S.L. Kendal, M. Creen. An Introduction to Knowledge Engineering. – Springer Verlag, 2007.

[Luger, 2008] G.F. Luger. Artificial Intelligence. Structures and Strategies for Complex Problem Solving: 6-th edition. – Addison-Wesley, 2008.

[Coppin, 2004] B. Coppin. Artificial intelligence illuminated. Jones and Bartlett Publishers, Inc. 2004.

[Way, 1991] E.C. Way. Knowledge Representation and Metaphor. – Springer Science + Business Media, B.V., 1991.

[Graham, Barrett, 1997] D. Graham, A. Barrett. Knowledge-Based Image Processing Systems. – Springer-Verlag, 1997.

[Weisfeld, 2008] M. Weisfeld. The Object-Oriented Thought Process. Third Edition. – Addison-Wesley Professional, 2008.

[Pecinovský, 2013] R. Pecinovský. OOP – Learn Object Oriented Thinking and Programming. – Tomáš Bruckner, Řepín-Živonín, 2013.

[Stroustrup, 2013] B. Stroustrup. The C+ + Programming Language: Fourth Edition. – Addison-Wesley Professional, 2013.

[Terletskyi, Provotar, 2014] D.O. Terletskyi, A.I. Provotar. Mathematical Foundations for Designing and Development of Intelligent Systems of Information Analysis. – Scientific Journal "Problems in Programing", 2014, N 2-3, 233-241 pp.

[Terletskyi, 2013] D.O. Terletskyi. The System of Set Theory for Operating with Essences in the Objects Intellectual Environment. In proceedings of the 6-th International Conference "Advanced Computer Systems and Networks: Design and Application" (ACSN'2013), Lviv, Ukraine, September 16-18, 2013, pp. 226-229.

[CGLib] Code Generation Library CGLib, https://github.com/cglib/cglib.

[CodeDOM] Code Document Object Model, http://msdn.microsoft.com/en-us/library/650ax5cx.aspx.

Authors' Information

Dmytro Terletskyi – 3rd year postgraduate student, Cybernetics Faculty, Taras Shevchenko National University of Kyiv, Kyiv-680, Ukraine; e-mail: dmytro.terletskyi@gmail.com

Major Fields of Scientific Research: Artificial Intelligence, Discrete Mathematics, Programming, Software Engineering.

Alexandr Provotar – Full Professor, Faculty of Mathematics and Natural Sciences, University of Rzeszów, 35 - 310 Rzeszów, Poland; e-mail: aprowata1@bigmir.net

Major Fields of Scientific Research: Artificial Intelligence, Category theory, Bioinformatics, Nanopharmacology.

Natural Language Processing Models

ONTOLOGY BUILDING AND ANNOTATION OF DESTABILIZING EVENTS IN NEWS FEEDS[1]

Vera Danilova

***Abstract**: This paper presents an attempt of elaborating a domain-specific knowledge resource to create semantic annotations of destabilizing events (civil unrest) within the framework of the socio-political event extraction task. The final objective is to reduce the manual effort of sociological researchers by automatically generating structured information on the progress of an event (protest as a verbal expression or an action), its participants, origins and the aftermath, as well as other reported details that can contribute to the analysis. The previous experience of destabilizing event ontology building (SPEED project), as well as several state-of-the-art works on the analysis of protest behaviour, are addressed. Ontology development in Protégé-4.3 and mapping using the GATE Developer 8.0 are described.*

***Keywords**: ontology construction, sociological application, event indexing*

***ACM Classification Keywords**: H.3.1 Content Analysis and Indexing*

Introduction

Protest activity manifests itself in a wide variety of events from verbal expressions, relatively peaceful demonstrations and strikes to civil wars, revolutions, coup d'état, etc. and is being explained by diverse social, economic, political and environmental causes. The main task addressed in the present paper deals with improving the quality of social protest events extracted from news streams by enriching them with semantic data. Russian researchers of social protest point out that the analysis of protest activity causes, motives and factors is insufficient in most state-of-the-art sociological papers within the framework of national studies. They address the formation of social movements (ethnic, feminist, ecological) as a conflict type of collective action, job actions, regional protest activity, modeling of public opinion dynamics, etc.. Although certain phenomena are covered, there is still a room for further research [Dementieva, 2013]. The scope of the international studies ranges from the examination of the intra-state protest expression to the global phenomena. In [Braha, 2012], newspaper reports on civil unrest events that took place in 170 countries during the period 1919-2008 are analyzed in order to model the mechanisms of social unrest contagion, which have proved to be similar to those of natural hazards and epidemics. The long-term dataset includes such action types as anti-government demonstrations, riots, and general strikes.

Ontology is an essential application for knowledge management and research. It is used for semantic search, annotation and linking, etc.. The most common representation of an ontology is a graph, where concepts are stored in the nodes and the edges represent the corresponding relations. The main advantage of this structure is that it can be easily understood by the machine due to the explicit and accurate definition of hierarchical and nonhierarchical relations between concepts. The largest ontology of destabilizing events

[1] Work done under partial support of the Catholic University of San Pablo (grant FINCyT-PERU)

(Societal Stability Protocol) that covers many kinds of human-initiated protests, politically motivated attacks by non-governmental initiators, as well as the reaction of the government was built within the framework of the SPEED project at the Cline Institute for Democracy of Illinois and will be addressed in detail in Section 2. It is a hierarchical domain ontology that generates event data collected across all countries after the Second World War [Hayes and Nardulli, 2011]. The data for the period from 2006 till present was crawled from over 5.000 news feeds in 120 countries several times each day.

The main objective of our work at the present stage is to develop a conceptual representation of protest activity in Russia and other countries (viewpoint of Russian news media), basing on the data crawled from Russian news streams. Concept mapping will allow the ability to automatically register the relationships between the observed characteristics of protest activity, e.g.: participation of certain actors (governmental/non-governmental) depending on the origins of the protest, association between the direct or indirect motivation to/not to participate and the actual involvement in the action, etc..

The rest of the paper is organized as follows. Section 2 describes the Societal Stability Protocol as a detailed conceptualization of the civil unrest domain, the structure of which was taken into account in the course of ontology building. Also, we mention the most common techniques in the field of ontology building. In Section 3, input data acquisition and tools are outlined. In Section 4, we describe the construction of the ontology, lexicons and annotation rules. Section 5 discusses the results and concludes the paper.

Related Work

An ontology can be constructed in a manual, semi-automatic or automatic way using natural language processing, unsupervised machine learning, and other techniques and resources. Ontologies can be generated either directly from text or from dictionaries, thesauri, knowledge bases, semi-structured and relational schemata [Lieto, 2008]. Learning pipeline includes term extraction, disambiguation, concept identification, concept hierarchy construction, identification of relations and rules within the ontology. Term extraction uses either indexing mechanisms from the information retrieval domain or natural language processing. For the context-based term disambiguation, clustering along with the use of association measures to detect statistically correlated pairs is applied. Thesauri and dictionaries are also employed to group terms with similar meanings. For concept identification, unsupervised machine learning techniques are widely used. In some approaches to concept hierarchy construction, lexical relations of hyponyms are extracted from corpus using automatically acquired context-based lexico-syntactic patterns. Also, an approach based on Harris's Distributional Semantics Hypothesis [Harris, 1970] has been applied. It takes into account the correlation between a word and a context. Contexts are encoded in term vectors, clustering is performed and, finally, a distance measure (TF-IDF or chi-square) is applied to separate term senses. The identification of nonhierarchical relations within an ontology involves the use of text mining techniques and linguistic analysis. The automatic rule identification (a rule looks like "X caused Y", "Y is triggered by X", etc.) is a less developed area, where there are no common and well-established approaches [Toledo-Alvarado et al., 2012]. The interaction with the user or expert at different stages, as well as the comparison to the existing ontologies and term hierarchies, contribute to the ontology refinement.

Societal Stability Protocol created within the framework of the SPEED Project (Social, Political, and Economic Event Database) of the Cline Institute for Democracy of Illinois contains a well-developed ontology of destabilizing events. A destabilizing event is a happening that unsettles the routines and expectations of citizens, causes them to be fearful, and raises societal anxiety about the future [Nardulli et al., 2013]. The database encompasses texts collected from the newspapers issued in 165 countries in the Post WWII era. All the articles are translated into English. The final protocol design iteration includes eleven sections responsible for data processing. The sixth sections deals with the domain ontology of event types consisting of three main Tier 1 categories (political expression events, politically motivated attacks, disruptive state acts). Political expression involves an obligatory presence of such parameters as public articulation, non-governmental actor, threatening or unwelcome political message. The main expression modes are a) verbal or written message, b) symbolic act, c) forming an association and d) mass demonstration or strike with the subsequent subcategories. Politically motivated attacks are violent actions or attempts by non-governmental initiators. Political motives in this case are defined as hatred toward socio-cultural groups or revenge for their prior actions, desire to change or control the government, follow or oppose a political ideology, advance a social cause etc.. Disruptive state acts include extraordinary or repressive acts by governmental initiators.

SPEED personnel constructed the ontology by exploring the literature on political violence, terrorism, political instability, and social movements in search of event categories. Secondly, real data was analyzed and the respective classification was refined. Event-specific information that is relevant to the study of the origins and development of civil unrest was defined. Event attributes, such as geospatial and temporal information (event coverage, latitude/longitude, precise/estimated time and date), participants (governmental/non-governmental initiators and their traits (number, weapons used), messengers, rioters, reactors), consequences (negative/positive to initiators and actual participants), targets and effects (*what* happened to *whom*: damage, injuries etc.), origins (*why* it happened) and event linking, are distributed between the other sections [Hayes and Nardulli, 2011; Nardulli et al., 2013]. The access to SSP is limited.

Our study focuses on the real-time dynamics of the daily happenings and their attributes reported in the news of a specific country. These smaller events are surrounded by various factors that under certain circumstances may affect the status quo. At this stage, a gold-standard ontology of social protest events is created using the Protégé-4.3 tool and verified by an expert in order to provide quality semantic annotation to the data via the GATE 8.0 framework.

Input Data and Tools

Input Data. News titles are selected as the input data for the ontology building and annotation, because they contain short descriptions of a wide variety of events related to the protest activity, most of which are reflected in the SSP ontology. The relevant data on the participants of events, their attributes and triggers can be extracted from the titles. It was also experimentally proved that the noun in the headline is the main argument of an event in 80% of cases [Wunderwald, 2011].

The dataset includes 2000 news titles, extracted from Russian and Ukranian news portals that provide data on socio-political situation in Russia and abroad on a daily basis (ria.ru, lenta.ru, fontanka.ru, forbes.ru, livejournal.com, gazeta.ru, news.rambler.ru, newsru.com, interfax.ru, news.yandex.ru, news.bigmir.net, kommersant.ru, hopesandfears.com/news, kp.ru, mk.ru, ng.ru, gazeta.ua/ru/, pravda.ru, trud.ru). The test set and gold standard for the present experiments include 553 titles each.

The crawlers are created within the Scrapy web crawling framework (http://scrapy.org). They extract news titles and, optionally, the text body, date/time and source. The crawler relies on the mutual presence of several keywords from two predefined keyword lists. The keyword lists are based on the previous manual analysis of civil unrest-related titles that mention conceptual components (trigger, actors, time, location, purpose, etc.) and their combinations. A separate module deals with the large amount of duplicates and partial duplicates in the dataset. They are completely removed using embedded python libraries. Levenstein distance algorithm (NLTK package) for string similarity is efficient, but very slow for big datasets. An example of the resulting collection is presented in the Tab. 1.

Tools. GATE Developer 8.0 (General Architecture for Text Engineering: http://gate.ac.uk) is a powerful open source annotation tool. Multiple plugins for the processing of various natural languages can be uploaded or created manually within the framework. In our experiments it is used, firstly, for preprocessing the collection, which includes such steps as tokenizing, sentence splitting, gazetteer lookup (an embedded Russian Gazetteer, a manually populated OntoGazetteer and a Flexible Gazetteer) and morphological analysis (an embedded version of Yandex API Mystem 2.1 (http://api.yandex.ru/mystem/doc/)). Flexible gazetteer matches words in any morphological variant, while standard GATE gazetteers (ANNIE, embedded language-specific gazetteers) provide only exact string match. OntoGazetteer main functionality consists in lexicon mapping to the ontology classes.

Table 1. An example of crawled titles with translation

Митинг в защиту детдома №2 состоится 16 марта.	An action in defence of the orphanage No2 will be held on 16th of March.
В Чите прошли два митинга в поддержку народов Украины.	Two actions were held in Chita in support of the peoples of Ukraine.
Митинг националистов на Марсовом поле завершился без происшествий.	The nationalist protest at the Marsovo field ended without consequences.
В Симферополе прошел митинг против "евромайдана".	In Simferopol a protest was held against "Euromaidan".
Пассажиры задержанного рейса устроили митинг в аэропорту.	The passengers of the delayed flight staged a protest in the airport.

Secondly, it is applied for the gazetteer population, grammar building and ontology mapping. The concept hierarchy itself is represented formally using Protégé-4.3 software of the Stanford University (http://protege.stanford.edu).

Ontology Construction and Text Annotation

Ontology Construction. The gold standard ontology of social protest is constructed manually on the basis of the SSP domain ontology and real data from news feeds, and it is formalized within the Protégé-4.3 framework. It has been revised by one domain expert, and it will be subject to control assessment focused on revealing and highlighting the less studied aspects. The current version spans action types, participant classification and different attributes that are included in the corresponding sections of the SSP: geographical and temporal characteristics, motives, consequences, origins, event nature (pacific/violent), etc.. All these data have been considered within the same ontology in order to visualize the dependencies, if any, between the attributes, participants and events, which can be a powerful application for sociologists.

Classes. Class hierarchy is based on the analysis of 2000 unique news headlines that were crawled using combinations of keywords from two sets. The first set contains words like "protest", "demonstration", "piqueting", "boycott", "march" etc., the second - words like "against", "contra", "in support of", etc.. SSP classes are taken into account, as well as the wide variety of resources on the Web, including the interactive access to DBpedia ontology classes and relations (gFacet tool: http://visualdataweb.org). Firstly, lists of events constituting protest activity, as well as those preceding and following it, have been built, analyzed and organized into a structure. Secondly, other parameters, such as geospatial, temporal data and event status have been added .

Protest activity is divided into *verbal expression* and actual *action*, which can be a *mass gathering* or a *symbolic act*. These classes are not disjoint, because such events may anticipate or follow one another. Also, an *action* can be *pacific* or *violent* (or it turns out to be violent (*ViolentConsequence*), we have not encountered any example of the opposite). A violent action may involve the use of weapons (*WeaponType* class). *ActionReason* divides actions into the expressions of protest, support, requirement, conmemoration or other. All of these reasons (support, conmemoration, requirement) can imply a protest. The scale of the action is measured by the amount of actions (single/multiple), amount of participants (one/group/many), location coverage (town/province/country/world). The status of the action is planned/in_progess/finished/never_took_place. *ActionType* includes *Strike, HungerStrike, BusinessStrike, March, Concert, Picketing, MassDemonstration, Riot, Rebellion, Revolution,* symbolic acts, etc.. *Motivation/Demotivation* classes are based on the reported data on event support (financing or other) or rejection by governmental or non-governmental actors. The events that precede and follow the action are put in the corresponding classes that describe threats, warnings, authority interventions into the planning process, different kinds of consequences (financial, property damage or other) and reactions of governmental and non-governmental actors. Participants include individuals, unnamed and named groups of people, governing authorities, political parties, church representatives, enterprises, law enforcement, etc., that can be initiators, targets, victims, support or participants of a protest event. We present a screenshot of the ontology that does not cover the entire structure, because the latter is subject to a control revision. The current version includes 13 first-level classes, 71 second-level classes and 102 third-level classes. A screenshot of the ontology visualization in the Protégé-4.3 framework is presented in Fig. 1.

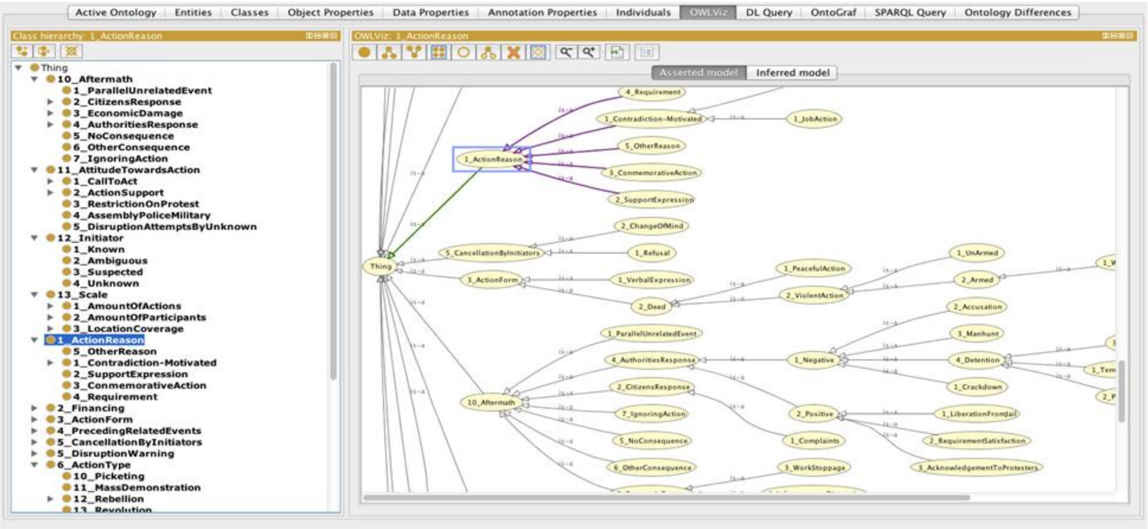
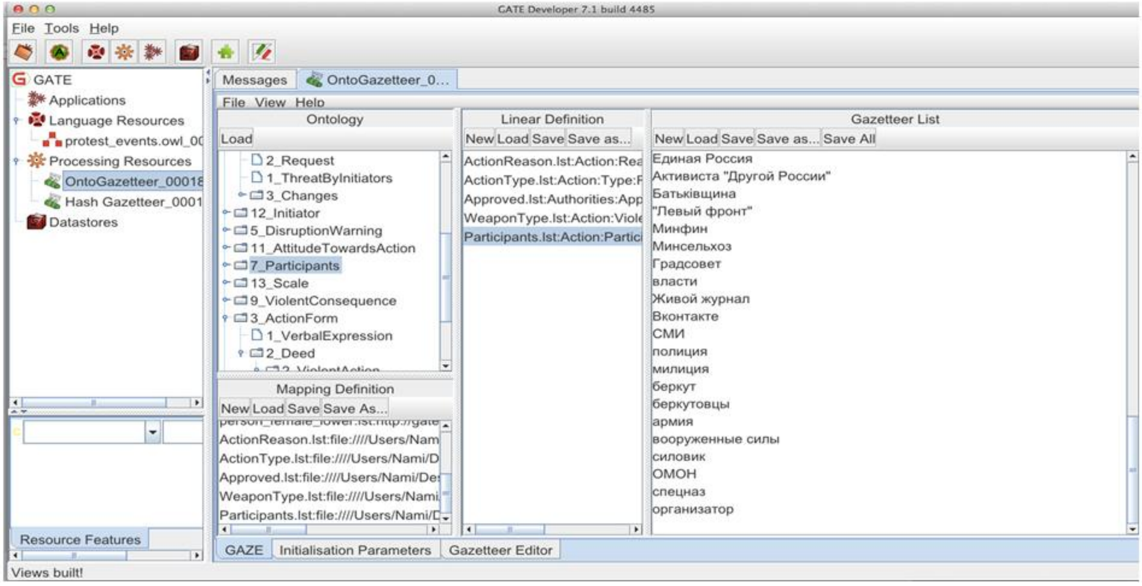


Fig. 1. Class hierarchy implemented in Protégé-4.3

Properties. Most ontology classes are characterized by the mutual influence, which can be manually defined in Protégé by means of specific rules (object property hierarchy). It currently includes a basic set of rules, such as HasParticipant, InitiatedBy, HasProperty, HasMotivation, HasReason, HasConsequence, IsFollowedBy, IsPrecededBy, etc.

Text Annotation. Annotation is the first step to transforming the unstructured text into quantitative event data. In our study we focus on ontology-based annotation and, as the future work, we consider the ontology population from the annotated data. At the current stage we have a manually constructed social protest gazetteer that is mapped to the ontology within the GATE Developer 8.0 (Fig. 2).

We also use embedded Russian lexicons (Lang_Russian package) that contain lists of geographical terms, date/time expressions, named entities, such as person names, person titles, organization names (government, companies, etc.). The annotation tool relies on JAPE (Java Annotation Pattern Engine) rules. JAPE finite state transducers consist of a left-hand side (LHS) that sets pattern constraints and a right-hand side (RHS) that contains annotation commands. In our experiments we apply cascaded grammars over annotations. Rules take into account the preprocessing results, OntoGazetteer, Russian gazetteer and Flexible gazetteer lookups, PoS tagging, discourse structure, etc.



2_Request
1_ThreatByInitiators
3_Changes
12_Initiator
5_DisruptionWarning
11_AttitudeTowardsAction
7_Participants
13_Scale
9_ViolentConsequence
3_ActionForm
1_VerbalExpression
2_Deed

ActionReason.lst:Action:Rea
ActionType.lst:Action:Type:F
Approved.lst:Authorities:App
WeaponType.lst:Action:Viole
Participants.lst:Action:Partic

ActionReason.lst:file:////Users/Nam
ActionType.lst:file:////Users/Nami/D
Approved.lst:file:////Users/Nami/De
WeaponType.lst:file:////Users/Nami
Participants.lst:file:////Users/Nami/D

Единая Россия
Активиста "Другой России"
Батьківщина
"Левый фронт"
Минфин
Минсельхоз
Градсовет
власти
Живой журнал
Вконтакте
СМИ
полиция
милиция
беркут
беркутовцы
армия
вооруженные силы
силовик
ОМОН
спецназ
организатор

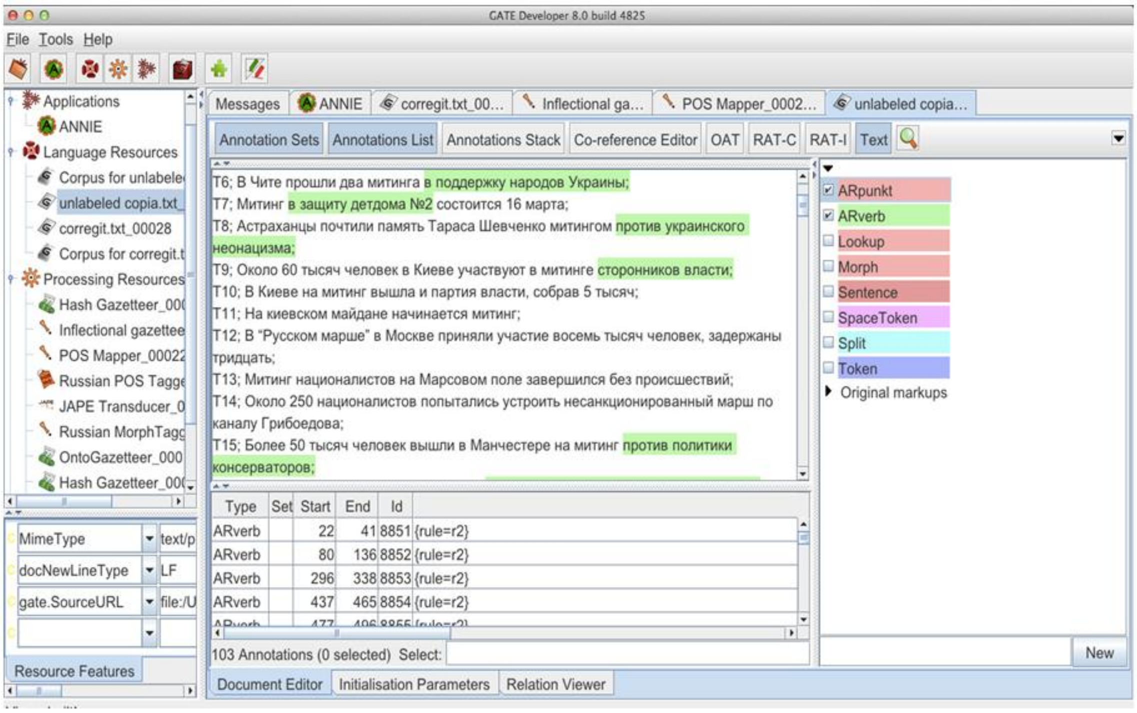
Fig. 2. OntoGazetteer

Annotation Experiments

Annotation rules. Within the present work, we have carried out several experiments on annotating word sequences that characterize the origins of protest activity. The *ActionReason* class representation commonly consists of two components: OntoGazetteer component that defines the action nature (protest, support, conmemoration, requirement or other) and a word sequence of variable length that contains new information, which needs to be categorized for automatic ontology population. Also, the data on the origins of protest activity can be represented as a prepositional or postpositional adjective to the protest type substantive: "антивоенный протест" ("*a protest against war*"), "митинг неонационалистов" ("demonstration of neonationalists"), etc..

Within the framework of the present experiments, we performed the annotation using two sets of rules and OntoGazetteer lookup. As it turned out, *ActionReason* occupies the final position in the headline in a half of the dataset: <*ActionReason*><End Point> - 46%, that is why the first set includes simple and robust rules that rely on the positional properties of this information block. The main pattern constraints are as follows: a sequence is annotated if it is preceded by [ActionReason Lookup], contains any number of tokens and is followed by the sentence end, a verb in indicative mood or a coma. [ActionReason Lookup] includes words denoting protest, support, conmemoration, requirement: "в поддержку" ("in support of"), "в защиту" ("in defence of"), "против" ("against"), etc.. A screenshot of the annotation based on the first set of rules (ARverb and ARpunkt) is presented in Fig. 3.

The second set are rules represented in cascaded grammar phases. The first phase filters the tokenizer results. The subsequent phases include pattern/rule pairs for sequential processing of the following: [ActionType Lookup] + [Participant rule]/[Random token sequence rule] + [ActionReason Lookup] + [Noun Phrase rule]. [ActionType Lookup] includes words like "протест" ("protest"), "марш" ("march"), "пикет" ("piqueting"), etc.. [Participant rule] defines the annotation of event participants and relies on the OntoGazetteer lookup and morphological analyzer results. [Random token sequence rule] extracts a random word sequence between ActionType and ActionReason Lookup annotations. [Noun Phrase rule] extracts the sought-for data on event origins that is commonly represented as a complex noun phrase. These rules do not take into account the position of information blocks and rely solely on the gazetteer lookups, tokenizer and PoS tagger results.



T6; В Чите прошли два митинга в поддержку народов Украины;
T7; Митинг в защиту детдома №2 состоится 16 марта;
T8; Астраханцы почтили память Тараса Шевченко митингом против украинского неонацизма;
T9; Около 60 тысяч человек в Киеве участвуют в митинге сторонников власти;
T10; В Киеве на митинг вышла и партия власти, собрав 5 тысяч;
T11; На киевском майдане начинается митинг;
T12; В "Русском марше" в Москве приняли участие восемь тысяч человек, задержаны тридцать;
T13; Митинг националистов на Марсовом поле завершился без происшествий;
T14; Около 250 националистов попытались устроить несанкционированный марш по каналу Грибоедова;
T15; Более 50 тысяч человек вышли в Манчестере на митинг против политики консерваторов;

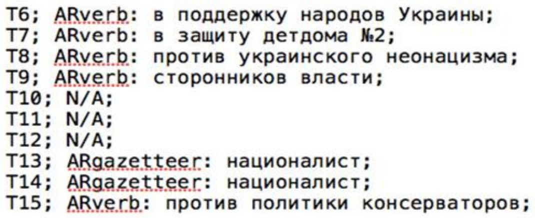
Type	Set	Start	End	Id	
ARverb		22	41	8851	{rule=r2}
ARverb		80	136	8852	{rule=r2}
ARverb		296	338	8853	{rule=r2}
ARverb		437	465	8854	{rule=r2}

103 Annotations (0 selected) Select:

MimeType text/p
docNewLineType LF
gate.SourceURL file:/U

Fig. 3. "ActionReason" class annotation

Gold standard. A gold standard (553 headlines) is built to check the performance of the *ActionReason* class annotation rules at the present stage. In the Fig. 4 an example of the gold standard is given, where T[1...*n*] denotes the number of the sentence, ARverb and ARgazetteer are rules, which must trigger the class annotation, N/A is set if no rule is applicable. The meaning of the annotated strings is as follows: "в поддержку народов Украины" ("*in support of Ukranian peoples*"); "в защиту детдома №2" ("*in defense of the orphanage No2*"); "против украинского неонацизма" ("*against Ukranian neonationalism*"); "сторонников власти" ("*of the pro-governmental activists*"); "националист" ("*nationalists*"); "против политики консерваторов" ("*against the conservative politics*").



```
T6; ARverb: в поддержку народов Украины;
T7; ARverb: в защиту детдома №2;
T8; ARverb: против украинского неонацизма;
T9; ARverb: сторонников власти;
T10; N/A;
T11; N/A;
T12; N/A;
T13; ARgazetteer: националист;
T14; ARgazetteer: националист;
T15; ARverb: против политики консерваторов;
```

Fig. 4. Gold standard

Evaluation. Within the framework of the present experiments, F_1 score (standard harmonic mean of Precision and Recall) of *ActionReason* class annotation has been calculated for two sets of rules. The test set (553 headlines), as well as the gold standard, are divided into five subsets.

$$Precision = \frac{|G \cap C|}{|G|},$$

where G is the number of sequences that were extracted from all the headlines for the *ActionReason* class, C is the amount of strings that coincide with the expert annotation of the same slot.

$$Recall = \frac{|G \cap C|}{|E|},$$

where E is the total amount of sequences that are relevant to the *ActionReason* class within a given test set, according to the expert annotation.

$$F_1 = 2\frac{PR}{P+R},$$

where P is the resulting Precision value for a given test set, and R is the resulting Recall value for a given test set.

Results and Future Work

Results. The annotation of 553 headlines divided into 5 test sets with the information on protest origins (*ActionReason* ontology class) has been performed on the basis of two rule sets. The first set uses OntoGazetteer and Flexible Gazetteer Lookups and position-related rule together with few punctuation and morphological constraints. The second set uses more sophisticated rules taking into account data from all available gazetteers, as well as tokenizer, PoS tagger and NP-chunker output. The results are presented in the Tables 2 and 3. The number of headlines per test set is shown in square brackets. A sequence annotation is considered relevant if it corresponds exactly to the expert annotation in the gold standard.

Table 2. Annotation results for the rule set 1

RuleSet_1	TestSet_1 [100]	TestSet_2 [100]	TestSet_3 [100]	TestSet_4 [100]	TestSet_5 [153]	Total [553]
Retrieved & Relevant	69	60	63	56	87	335
All Relevant	76	71	68	67	97	379
All Retrieved	80	78	77	70	96	401

Table 3. Annotation results for the rule set 2

RuleSet_2	TestSet_1 [100]	TestSet_2 [100]	TestSet_3 [100]	TestSet_4 [100]	TestSet_5 [153]	Total [553]
Retrieved & Relevant	71	66	66	65	91	359
All Relevant	76	71	68	67	97	379
All Retrieved	75	69	67	66	96	373

The results show that RuleSet_1 annotates more irrelevant sequences, while the number of correctly labeled instances is rather high. RuleSet_2 uses many constraints, which allows to annotate more exact sequences, however, in terms of the runtime it performs slightly slower. Manual checking shows that in case of RuleSet_1 most "incorrect" sequences are noisy, but relevant. Evaluation of annotation results is presented in Tab. 4.

The obtained results suggest finding a compromise between RuleSet_1 and RuleSet_2, so that we can reduce the number of constraints and increase the number of correctly annotated sequences. Also, the annotations of other classes that overlap with the *ActionReason* class annotation should be taken into account.

Table 4. Evaluation

Measure/Rule Set	RuleSet_1	RuleSet_2
Precision	0.83	0.96
Recall	0.88	0.94
F1 score	0.85	0.94

Future Work. The present paper describes the creation of knowledge resources for the automatic annotation of events related to the protest activity. The following tasks are proposed as future work: 1) control checking of the ontology by a domain expert; 2) automatic ontology building on the same data; 3) improvement and evaluation of patterns for the corresponding ontology classes; 4) automatic annotation-based ontology population; 5) language coverage improvement for comparison issues: Spanish.

Bibliography

[Braha, 2012] D.Braha. A Universal Model of Global Civil Unrest, PLoS ONE 7(10): e48596, 2012.

[Dementieva, 2013] I.N.Dementieva. Theory and methodology of social protest study, Journal of Public Opinion Monitoring, Vol. 4 (116), 3-12, 2013.

[Harris, 1970] Z. Harris. Papers in Structural and Transformational Linguistics, Dordrecht/ Holland: D. Reidel., x, 850 pp., 1970

[Hayes and Nardulli, 1949] M. Hayes, P. F. Nardulli. SPEEDs Societal Stability Protocol and the Study of Civil Unrest: an Overview and Comparison with Other Event Data Projects (white paper). Cline Center for Democracy, University of Illinois at Urbana-Champaign, 2011.

[Lieto, 2008] A. Lieto. Manual and semi-automatic domain-specific ontology building (master thesis), Università degli studi di Salerno, 2008.

[Nardulli et al., 2013] P. F. Nardulli, M. Hayes, J. Bajjalieh. The SPEED Projects Societal Stability Protocol: An Overview (white paper), Cline Center for Democracy, University of Illinois at Urbana-Champaign, 2013.

[Toledo-Alvarado et al., 2012] J. I. Toledo-Alvarado, A. Guzmán-Arenas, G. L. Martínez-Luna. Automatic Building of an Ontology from a Corpus of Text Documents Using Data Mining Tools, Journal of Applied Research and Technology, Vol.10, No. 3, 398-404, 2012.

[Wunderwald, 2011] M.Wunderwald. *Event Extraction from News Articles (Diploma Thesis)*, Dresden University of Technology. Dept. of Computer Science, 2011.

Authors' Information

Vera Danilova – PhD student at the Autonomous University of Barcelona, Dept. of Romance Languages; Junior research fellow at the Russian Presidential Academy of National Economy and Public Administration. E-mail: maolve@gmail.com

Major Fields of Scientific Research: Multilingual Event Extraction, Ontology Building, Sociological Applications

SEMANTIC AND ONTOLOGICAL RELATIONS IN AIIRE NATURAL LANGUAGE PROCESSOR

Alexey Dobrov

Abstract: *AIIRE is a free open source natural language processor, developed by a team of researchers in Saint-Petersburg, Russia. AIIRE is an implementation of full-scale NLU process, based on the method of inter-level interaction and rule-based disambiguation. Semantic graphs that are built by AIIRE are based on the involved ontology. The rules that concern correspondence between semantic relations (used in semantic graphs by AIIRE) and conceptual relations (used in ontology) is a matter of discussion. Semantic graphs are evaluated from syntactic trees, and, in general, although word-independent syntactic constituent classes tend to denote rather abstract relations (cf. genitive construction in general), the instances of those classes (specific phrases) in theory may denote any subclasses of those relations. The developed algorithm of choosing a relation subclass in each case is also a matter of discussion.*

Keywords: *ontological semantics, lexical disambiguation, NLU, conceptual relations, semantic graphs.*

ACM Classification Keywords: *1.2.7 – Natural Language Processing: Language Parsing and Understanding, Text Analysis*

Introduction

AIIRE is a free open source natural language processor, developed by a team of researchers in Saint-Petersburg, Russia. The acronym 'AIIRE' stands for 'Artificial Intelligence-based Information Retrieval Engine', that was the first production-level application of the developed NLU-kernel. The team was formed in 2003-2005, in the Laboratory for Informational Linguistic Technologies of the Institute of Linguistic Studies in Saint-Petersburg, Russia, and continues its work as the RnD department of Geline company.

The method of inter-level interaction

AIIRE natural language processor is an implementation of full-scale NLU process, based on the method of inter-level interaction and rule-based disambiguation. The method of inter-level interaction was first proposed by G.S. Tseitin in 1985 [Tseitin, 1985], but was not implemented since then because of its complexity, and because of lack of well-developed linguistic software and high-performance hardware. The basic idea of the method was to get rid of the artificial separation of levels of linguistic analysis and to analyze morphology, syntax, and semantics in the same time. This way of analysis allows to perform disambiguation on lower levels using the upper-level rules immediately after the ambiguity arises on the lower levels, rather than after the whole text (or sentence) is analyzed on these levels. E.g., morphological disambiguation can be performed using results of syntactic binding immediately after the first two terms of the syntactic tree are bound (or not bound) according to restrictions of the involved grammar. Furthermore, if two combinations of morphological analysis hypotheses still remain after syntactic binding, i.e., if both combinations give grammatical syntactic trees, then these trees are analyzed in terms of semantics immediately after the trees are produced, semantic restrictions acting as filters to reduce ambiguity on the level of syntax.

The idea of inter-level interaction helps to prevent (or, at least, to reduce drastically) combinatorial explosions, which is crucial for NLU performance. Formal grammars produce a plenty of ambiguities during the analysis, especially when ellipsis is allowed. Moreover, if each textual wordform has just two interpretations, then traditional separation of the levels of analysis leads to 1024 combinations having to be considered just to perform syntactic analysis of a 10-word sentence, each combination having a set of hypotheses of syntactic binding. Dozens of thousands of grammatical syntactic trees are produced, and just a few of them remain after the semantic analysis. The idea of inter-level interaction allows to apply the restrictions of the highest levels of analysis much sooner, reducing the maximum amount of combinations to tens or rarely to hundreds.

Rule-based and Machine-learning approaches to disambiguation

Nowadays, statistical heuristics are much more popular than rule-based methods of disambiguation. Corpus-based approaches had a great success in morphological disambiguation (95% quality, cf. [Hajic et al. 2001]), which seemed a breakthrough in the field. Corpus-based (so-called machine-learning, ML) approaches seem to be essentially more simple and objective than rule-based methods, and are quite efficient for many tasks. The problem remains, however, that statistical heuristics never guarantee the absence of false-negative results (i.e., correct hypotheses being merely culled), which can have catastrophic consequences even when morphological analysis is followed only by syntactic parsing, as the whole syntactic trees are lost. Formal grammars are based on strict rules of grammatical coordination and government, of ellipsis and word-order, and loss of a correct hypothesis of morphological analysis often leads to a complete failure yet on the level of syntax. That is why AIIRE does not use any statistical heuristics to perform disambigution, although further research may help to develop some corpus-based mechanisms that guarantee absence of false-negatives. The same is true, however, not only for NLU-systems like AIIRE, but also for efficient implementations of spell-checkers and punctuation checkers (cf. [Petkevič, 2006]), because of the same inadmissibility of false-negatives.

The main source for disambiguation in AIIRE is its ontology. Grammatical restrictions help to reduce (but not to eliminate) ambiguities on the level of morphology, but grammar just very rarely helps to get rid of homonymy (it happens in a few cases when homonyms have partially different paradigms, e.g., Russian word '*лист*' has plural '*листы*' to denote 'sheets' and '*листья*' to denote 'leaves'), and never helps to choose between word meanings (more precisely, it is a convention in the AIIRE project, that if two different meanings can be distinguished only by means of grammatical context, then these meanings belong to different homonyms of the same word).

AIIRE ontology, semantic dictionaries, thesauri and knowledge bases

The database that contains lexical meanings is called ontology in AIIRE project, but it also can be called (because it provides functionality of) semantic dictionary, as it was in [Leontyeva, 2006], or thesaurus, because it represents main inter-meaning relations, which are normally registered in thesauri, or even a knowledge base, because it provides knowledge not only on conceptual classes, but also on their instances, at least on those registered in Wikipedia. Nevertheless, the term 'ontology' was chosen for the following reasons:

- Semantic dictionaries are principally language-dependent, whereas ontologies are not. AIIRE ontology contains lexical meanings, but it also contains even more concepts that can not be bound to specific lexical entities, sometimes, even to expressions of any languages. These concepts are parts of meanings or superclasses like 'Object that is localizable in any-dimensional linear space, and therefore has dimensions, and therefore has size', which are necessary for restricting semantic valencies of meanings of lexical entities like '*big*', '*small*', '*at*', '*in*', '*within*' and many others to a very abstract class of objects, which contains physical objects, geometric figures, parts of images, of texts, and even of melodies. If object has size, then it can be *small* or *big*, no matter whether this object is physical or virtual, and no matter how many dimensions it has. If object is localizable in any linear space, then it can be located *in* another object in the same space, and they both have size, no matter which kind of space is in question — three (or four-) dimensional physical space or one-dimensional space of text. Abstract concepts like these are never registered in semantic dictionaries, but certainly are registered in ontologies under the code names like 'Localizable' and 'Sized', which are not to be confused with similar lexical entities.

- Thesauri like WordNet (cf. [Miller, 1995], [Fellbaum 1998]) are certainly kinds of semantic dictionaries, although they are restricted to inter-meaning relations like *hyponymy* or *synonymy* and a set of others. Thesauri do not always distinguish between class-to-class and instance-to-class types of inheritance, and therefore sometimes contain encyclopedic knowledge like specific instances (hyponyms) of abstract classes (hypernyms). However, thesauri never contain conceptual relations like 'can perform action' (relation between meanings of nouns and verbs, that reflects ability of instances of class denoted by noun to perform actions that are instances of class denoted by verb). These relations are mostly driven by extralinguistic knowledge (e.g., ability of physical objects to move in physical space, as disability of, e.g., collections of data to do so, can never be deduced merely from language), and, in the same time, are crucial for lexical disambiguation (e.g., sentences like *The table was moved to the corner of the room*, where the word *table* can mean both an object of furniture (a physical object), and a collection of data (e.g., *database table*) can be disambiguated only because of the knowledge that only physical objects can move in physical space).
- Knowledge base is, probably, more proper term to denote AIIRE ontology than dictionary or thesaurus, but still is not specific enough to reflect the fact that the items stored in the ontology are models of concepts, which form not only an inheritance hierarchy, but also an extensive network due to the above mentioned relations. Ontologies are sometimes even treated as kinds of knowledge bases [Knowledge Base, 2014], which, in contrast to other kinds of knowledge bases, have hierarchical structure.

Probably, the most significant difference between AIIRE ontology and semantic dictionary, thesaurus or knowledge base of any kind is that relations that constitute this ontology are concepts themselves and form their own inheritance hierarchy. The same is true for some other ontologies, e.g. OpenCyc, but is not an obligatory requirement to the structure of ontology (e.g., SUMO does not follow this practice, cf. [Zouaq et al. 2009], [Sheffczyk et al. 2006]). This peculiarity of AIIRE ontology is, however, just a side-effect of one of the conventions adopted in AIIRE project, which states that every entity used in AIIRE ontology must be a concept defined in this ontology. This convention also leads to lexical entities (even of different languages) being treated as concepts that are connected with their meanings with the so-called 'to denote object'[2] relation. More than that, each concept, except for those corresponding to lexical entities themselves, must be bound to a lexical entity (which may be rather a large expression), even if natural language doesn't have

[2] Relations are stored as meanings of non-idiomatic natural language expressions, that are formed according to the following convention: an infinitive verb phrase that describes the relation + a noun phrase that describes its object class. Subject classes are described in braces within the description of the relation concept itself.

words or idioms that denote this concept. This requirement allows to present ontology to its editors in form of a dictionary, which is much more common form of presentation for linguists.

The above-mentioned hierarchy of conceptual relations is the main source of restrictions for lexical disambiguation. This hierarchy is linked with other hierarchies: each relation is linked with its subject and object classes (there are relations named 'to have subject' and 'to have object', that mark these links), and with its so-called 'refraction' (inverse relation) thru 'to be refracted with relation' relation. Relations are a priori coventionally divided into direct and reverse ones, according to the order of evaluations of natural language expressions that may lead to these relations. The choice for each relation is sometimes unobvious, because sometimes both direct relation and its refraction can be expressed in natural language without inversions or passivizations, e.g., both 'to have a pet' and 'to belong as a pet to somebody' relations can be expressed with genitive constructions: cf. *the dog of Mary* and *the owner of the dog*. In such cases, inheritance hierarchy is involved: in the above-mentioned example, 'to have a pet' relation is a subclass of 'to have an object', which, in turn, has a subclass that is one of the meanings of preposition *with* (cf. *a girl with a dog*, *a girl with long hair*, *a girl with high IQ* etc.) and therefore is direct. Thus, 'to have a pet' is considered also direct, because directness of the superclass is inherited.

Rules of relation inheritance and overriding

In AIIRE ontology, model of concept is a set of attributes, each attribute being a relation-object pair. Only direct relations of concept are stored in the ontology, their refractions being evaluated and presented to users on a par with direct ones. Directness of relations is also used in conceptual graph normalization procedure: reverse relations are converted into direct ones. This is the only reason why relations have to be a priori conventionally divided into direct and reverse: it is obvious from the definition of relation, that both a relation and its refraction can be chosen to be treated as direct or reverse.

It is also necessary to mention, that there are strict rules of inheritance and overriding of relations in AIIRE ontology, that are crucial for the mechanism of semantic restrictions. If a concept inherits another concept (i.e., it has an attribute with 'to inherit concept' as relation and the latter concept as object), then every attribute of inherited concept is primarily treated as an attribute of of the inheriting one. In the same time, each attribute of the inherited concept can be overridden by one or more attributes of the inheriting one. One attribute overrides another one if and only if relation of the overriding attribute inherits or coincides with relation of the overridden one, and the same is true for the objects of these relation. E.g., concept 'span' has an attribute <'to have size', 'length'> (i.e., size of a span is its length), whereas concept 'time span', although it inherits 'span', has an attribute <'to have size', 'duration'>, which means that size or length of a time span is duration. Overriding attributes can be acquired thru overridden relation-object pair, but, due to overriding, the range of possible attribute values is restricted on each step of inheritance. As relation between instances and their classes is also a kind of inheritance, overriding rules apply to the instances. Furthermore, there is a significant limitation for instances, which states, that all attributes of an instance must override at least one of the attributes of its class. Thus, the sentence *The size of time span was two minutes* does meet these semantic restrictions, whereas *The size of time span was two meters* violates these restrictions.

Conceptual and semantic relations. What is conceptual relation?

As an NLU system, AIIRE produces semantic graphs as final representations of textual semantics. These graphs are completely built from the concepts defined in the ontology. The term 'semantic graph' is used here in a very wide sense, not as a synonym of 'conceptual graph' J. Sowa had proposed in [Sowa 1999]. It is maybe even better not to call this representation a graph, because its edges can act as vertices and have their own edges, which is, however, not forbidden in terms of mathematics. Both vertices and edges of semantic graphs are either concepts from the ontology, or, more frequently, textual concepts that must be subclasses of the ontological ones and have relations with each other. Textual concepts must follow inheritance and overriding rules of ontology. Edges of semantic graphs must be relations, vertices can be concepts of any kind. Not all ontological relations are involved in semantic graphs. For this particular reason, it seems expedient to distinguish between conceptual relations (any relations between concepts) and semantic relations (conceptual relations that can be edges in semantic graphs). In order to define the term 'conceptual relation' more precisely, a mathematical definition of relation can be used: relation is a subset of direct product of several sets. In the case of AIIRE ontology, relations connect conceptual classes or instances, and are binary, thus, it is better to define conceptual relations between classes as classes of pairs of instances of two classes, relation instances being particular pairs. Some relations can be deduced from other relations as their superpositions, in which cases they are called implicational. It is also possible to bind relations with predicates, and thru predicates even with algorithms that evaluate these predicates (e.g., the relation named 'to occur before situation' between two situations can be deduced from time points of these situations and from comparison of their numeric representations in seconds since epoch), but this way of modeling conceptual relations is still in the earliest stages of development.

Inheritance hierarchy of relations. Relations of concept 'concept'

As hierarchy of relations is linked with their subjects and objects, relations can be classified by their participants. Some relations are linked with the root of the ontology (concept 'concept') as subject, which means that any concept can be subject of these relations. E.g., there are two existence-marking relations between any concept and any reality ('to exist in a reality' and 'not to exist in a reality'), two kinds of equivalence relations between any concepts ('to be / not to be equal to concept' and 'to be like / unlike concept'), quantifier-binding relation 'to be bound to quantifier', the refractions of some previously mentioned relations ('to be / not to be subject of relation', 'to be / not to be object of relation'), and markers of their implications ('to be / not to be subject of situation', 'to be / not to be object of situation'). These relations are rather abstract, and in the majority of cases only their subclasses are actually involved into evaluation of semantics, e.g., 'to be subject of action', 'to be object of relation' etc. These relations are necessary for disambiguation of subject-predicate and verb-object constructions, prepositional phrases and their combinations with nominal and verbal phrases, and some other kinds of constructions.

Variable and constant objects and their relations

Some relations are linked with a bit less abstract concept 'variable object' as possible subject. This class corresponds to any object that can change in time (and, therefore, appear or disappear), in contrast with constant objects. Variable objects can have states and are linked with their states with corresponding relations ('to have a historical state', 'to have a contemporary state', etc.), they can be created (invented (but not dicovered), produced, built, born etc.) by someone ('to be created by someone'), at some time ('to be created at year/century/millenium'), with some instrument ('to be created using object'), etc. The majority of nominal meanings correspond to variable objects, nevertheless, in some cases it is important to distinguish between constants and variables to perform lexical disambiguation. E.g., in the phrase *The square was built in eighteenth century* the word *square* is ambiguous: it can denote both an area in an inhabited locality and an equilateral quadrangle. The latter is an abstract geometric figure which is not variable (quadrangle is a geometric set of points, and sets are constants), and thus can not be created or built. Therefore, the ambiguity is resolved only by means of the constant-variable distinction. It is worth mentioning, however, that in a very special context the word *square* can mean not the quadrangle itself, but an image of quadrangle depicted by someone; images are variables and therefore can be created. Furthermore, the verb *to invent* can denote (again, in a very special context) 'to discover'. These contextual meanings are produced by the mechanism of metonymy, which is enabled in AIIRE only in cases when none of the word meanings meet the semantic requirements of the context. This mechanism is described lower; it is based on the so-called 'backbone relations', e.g., geometric figures have a backbone relation 'to be depicted on a picture' with physical pictures, so that, e.g., the sentence *The triangle was built in eighteenth century* has an interpretation which assumes that the word *triangle* metonymically means a picture of triangle. Practically each constant concept has backbone relations with variable concepts, therefore constant-variable distinction rarely reduces ambiguity, but helps choosing between immanent word meanings and meanings of their metonyms.

Physical objects, physical places, and their relations

One of the most significant base classes of AIIRE ontology is 'physical object'. Physical objects are classified as variable triple-dimensional bodies (bodies are non-empty localizable objects) and inherit relations of these superclasses, but these relations are overridden by immanent attributes of physical objects themselves. In particular, physical objects can be parts of other physical objects ('to be / not to be part of physical object'), they can move in physical space in some directions ('to move in physical space in direction'); physical objects can be located or situated in physical space ('to be located in physical place'), inside (outside, behind etc.) other physical objects; physical objects have physical attributes, such as taste, odor, color, form, temperature, viscosity, toughness, weight etc. All these attributes are essential for lexical disambiguation, because lexical ambiguity very often appears as distinction of physical and non-physical meanings. E.g., the famous phrase *colorless green ideas* can be interpreted only as boring (*colorless*) ideas of environmentalists (*green*), whereas the words *colorless* and *green* certainly have meanings that correspond to physical characteristics of color, but the word *ideas* has no meanings that correspond to any physical object. The same mechanism helps to perform disambiguation in genitive constructions or (in languages without inflectional genitive) in constructions with prepositions like *of* (cf. *the foundation of the theory / the foundation of the building*), in constructions with locative prepositions (cf. *a chair in the room / a chair in physics*), and in any constructions referring to motion (cf. *oncoming train / oncoming event*), etc.

Physical objects are not to be confused with physical places, although physical objects have backbone relations with the places where they are located and therefore produce metonymy. Like physical objects, physical places are triple-dimensional and localizable, but they are not necessarily variable, and, what is the most important, they are not bodies. Places themselves do not have physical characteristics such as color or weight, but due to reverse backbone relations with physical objects contained in them, they are sort of pretending to have temperature, color and some other physical attributes (cf. *warm green place*, which is to be interpreted as 'a place where the majority of physical objects are warm and green'). Countries and other inhabited localities are variable places, which have (at each moment of time) backbone relations to constant places (e.g., Germany is a variable place, which can change its borders, but its contemporary borders correspond to a constant place, which had existed even before Germany arose as a state and even before mankind arose as a biological species). Inhabited localities also have backbone relations with their societies (these relations are likely to come from more abstract backbone relations with contained physical objects mentioned above, but it is not necessarily true), so that metonymy also appears in phrases like *Germany has voted for Angela Merkel*. Physical places have a few immanent relations, most of them overriding those of abstract places (parts of any-dimensional spaces) and more abstract concepts. Physical places can be situated near other physical places (but not, e.g., near places in melodies), they can be scenes of some events or processes, they may have other physical places as parts (but not physical objects), and therefore can be parts of other physical places themselves.

Processes, actions, states, activities, subject domains and their relations

Another base-class concept which is very important for semantic analysis is 'process'. Processes are treated as contiguous sequences of situations. As sequences, processes have beginnings and endings, and therefore can begin or end. Because processes are contiguous, processes can last or break. Sequences are subclasses of aggregates, and, because inclusion relation 'to include concept' of aggregates are backbone, processes always produce metonymy with the situations that constitute them. Situations, in turn, can not last like processes (so that the expression *situation lasts* is interpreted only thru reverse metonymy with process), but they are linked with time spans (again with a backbone relation), with their participants (subjects and objects) and with relations (states or actions) between them. Situations can be atomic, in which case they are constants, they can be real or hypothetical, and they are not immanently localizable in space, but have an implicational relation 'to take place' with physical places, which correspond to localizations of situation participants. As the relation 'to correspond to relation' of situations is backbone, situations also produce metonymy with their relations, which allows expressions like *signing of the contract by the committee on Monday in Warsaw*. Situations may have reasons and purposes, and the same applies to processes because of metonymy. Processes may correspond to activities ('to correspond to activity', 'to have an activity as a purpose'); because of metonymy with time spans, processes can take place before, after or simultaneously with other processes; because processes are sequences, they can be parts of other processes and have subprocesses; furthermore, they can correspond (or not correspond) to processual patterns (so that they can bind with adjectival meanings like 'correct' or 'usual' and their opposites).

The hierarchy of processes in linked with the hierarchy of subclasses of the concept 'to perform action / to be in a state', which is subdivided into imperfective and perfective sub hierarchies (but only for actions), so that there are three (for actions —four) isomorphic hierarchies bound to nominal and verbal meanings. Russian verbs have aspect category (each verb can be either perfective, or imperfective), and English verbs do not, so that Russian verbal meanings always correspond to subclasses of concepts that are denoted by English verb. E.g., English verb *to draw* (a picture) does not have any direct equivalent in Russian, but can be translated either with *рисовать* (to be in the process of drawing), or with *нарисовать* (to have drawn). In AIIRE ontology this fact is treated so that English verbal meaning should correspond to superclass ('to draw'), and Russian verbal meanings correspond to its subclasses. Perfective meanings are linked with imperfective meanings with 'to finalize action' relation, and possibility (or impossibility) to express perfection of action or state without any additional components of meaning (in Russian) is one of the criteria to choose whether a verb means action or state. E.g., Russian verb *спать* (to be in the process of sleeping) does have perfective derivates (*поспать* (to have been in the process of sleeping for some time), *проспать* (to have been in the process of oversleeping), *доспать* (to have brought the process of sleeping to some point), etc.), but none of these verbs expresses precisely the meaning 'to have slept', which means that 'to sleep' is not an action, but rather a state. It is worth saying that even in English there is a strong difference between *to have slept* and *to have drawn*: e.g., it is much better to say *I have slept for hours* (121000 literal results in Google, which is synonymic to *I was sleeping for hours* — 76000 results) than *I have drawn for hours* (120 literal results in Google; *I have drawn* is not synonymic to *I was drawing*; *I was drawing for hours* has 14000 literal results in Google). The main difference between actions and states is that actions are processes that obligatorily lead to some changes, whereas states do not necessarily have any results or consequences. Each action has a backbone relation with a state of performing this action; in Russian this distinction is expressed lexically: cf. *идти* (to be in the process of going somewhere) and *ходить* (to be in the state of going, usually ‘hither and thither’ or multiple times to the same place).

All four above-mentioned hierarchies are classified according to characteristics of compatibility. Thus, transitive verb meanings and meanings of their nominal derivates correspond to subclasses of concepts ‘to perform a directional action / to be in a directional state’ and ‘directional action / state’, whereas intransitive verbs and their nominal derivates denote subclasses of nondirectional actions and states. Directional states are aliased as attitudes. Directional actions and states have objects which do not coincide with subjects, whereas nondirectional actions and states are limited to their subjects and therefore are expressed in natural languages as intransitive, or even reflexive verbs. Directional actions and states are classified according to the classes of their objects: e.g., the verb to *approve* means a state (an attitude), which is directed at a thought or idea (or, metonymically, an action of expressing this attitude), whereas the verb to *move* means an action, which is directed to a localizable (usually physical) object, but it can also mean a nondirectional action (e.g., *We moved to another apartment*). Subclasses of attitudes and directional actions are linked to the classes of their objects, and these links follow the rules of attribute overriding, so that lexical disambiguation can be performed. E.g., the verb *to take* can mean an action, which is directed at a physical object (*she took my hand*), and also (among other meanings) an action, which is directed at time span (*she took two hours to find me*).

The same classifications are done according to subject, addressee, and instrument classes of actions and states, regardless of whether they are directional or not. These classifications are vital for lexical disambiguation when it comes to verbs and processual nouns, but, unfortunately, they just rarely help to disambiguate the surrounding context. The reason is that verbs and processual nouns are very often polysemic and produce a significant amount of metonymy, so that if subject, object, instrument or addressee is denoted by polysemic nominal phrases, their meanings combine well with side-meanings of the verb or processual noun. E.g., the above-mentioned verb *to take* is highly polysemic, so the expression *I took the medicine* can be interpreted both as ‘I swallowed the pill’ and, e.g., as ‘I approved the proposed way of healthcare’. That is why it is a very important task (which is still not fulfilled) to deduce as much verbal meanings as possible from other (basic) meanings as metonymic ones, so that they are enabled only in case the basic meanings are exhausted. E.g., the meaning ‘to approve’ of the verb *to take* seems to be metonymic (maybe, partially, metaphoric), whereas the meaning ‘to hold’ seems to be one of the basic ones.

Sometimes, however, lexical disambiguation works efficiently enough to reduce or even to get rid of ambiguity, especially when verbal meanings are restricted to animate subjects or other participants (more precisely, subclasses of ‘someone’ concept, because in terms of semantic restrictions, e.g., plants are much less animate than, e.g. computers or robots ore any other human-like objects). ‘Someone’ is rather a wide class, as its subclasses are not only human-beings, other animals, and human-like subjects, but also societies and organizations that can act as a single person. Organizations have many attributes, some of them coming from metonymy with their members (persons), and some of them being immanent for organizations. Both persons and organizations can possess objects as private property, which is marked with the relation ‘to possess object’. This relation overrides more abstract relation ‘to have object’, so (for persons) other types of possession (also overriding the abstract relation) have to be introduced: e.g., ‘to have a relative’, ‘to have a part of body’, etc. Unlike persons, organizations can belong as property to other organizations and persons (because organizations are treated as ownership resources, and persons are not), and be parts of other organizations (as they are aggregates). Organizations can be registered in political units of an inhabited locality, and there are different ways of their creation, marked with relations that override basic variable object creation: organizations can be established, they can be formed or recruited, etc. Persons, as they are organisms with sexual reproduction, have three other possibilities of creation: they can be born by female persons, created by God or created by bioengineering. Both persons and organizations can perform different type of activities that correspond (again, with a backbone relation) to subject domains.

Subject domains are not to be confused neither with topics, nor with activities. Formally, in set theory, subject domain is domain of definition of a predicate, i.e., a union of sets in which a predicate is specified. In AIIRE ontology, subject domains are treated as aggregates of all real and possible situations and their participants, that belong to a specific class of someone's activity. There is a specific relation called 'to belong to subject domain' that links concepts with their subject domains. All classes of actions that correspond to activities that form subject domains can be performed only by subclasses of 'someone'. E.g., there is subject domain named 'science', which is formed by activity named 'cognition', which corresponds to action named 'to cognize'. This action can be performed only by a subclass of 'someone'. Subject domain hierarchy reflects the hierarchies of actions and activities, and thus is built according to very strict criteria, unlike hierarchies of topics that are usually created for the purposes of text classification.

Classification on the basis of subject domains and conceptual relations

Verbal and prepositional meanings are classified not only according to their semantic compatibility (classes of possible subjects and objects), but also according to their subject domains. E.g., there is a class of verbal meanings named 'to perform or be in state of physical motion', that is a superclass for all verbal meanings that somehow refer to physical motion. This class is further subdivided into subclasses (the general compatibility-based classification is reproduced), but it has its immanent attributes, as any motion has a direction and source point. Relations 'to be directed at physical place' and 'to start from physical place' provide compatibility of verbal meanings with corresponding meanings of prepositions. Some other spatial prepositional meanings are also linked to verbs of motion with some special relations (e.g., meanings of prepositions *through* and *past*). As for the locative prepositional meanings that do not refer to directions, these meanings are linked to all verbal meanings on the upper level, because every situation can be located in physical space.

Typical representatives, relation narrowing, and semantic analysis

It is, unfortunately, impossible to outline all the relations used and defined in AIIRE ontology in this paper, but it seems to be important to mention some large classes of relations or frequently used relations that were not mentioned before. Conceptual classes can have typical representatives (relation is called 'to have typical representative'), which are subclasses that act as metonyms. E.g., conceptual class 'place' has subclass 'physical place', which is a typical representative of place, therefore, the word *place* can denote 'physical place', although its meaning is much wider (places can be in texts, images, melodies, etc.) The same is true for the word *animal*, which tends to denote 'non-human animal'. Typical representatives often have specific compatibility because of their own relations (e.g., as it was previously mentioned, physical places have backbone relations with physical objects and, therefore, can be warm, green, etc.)

Relations, as other concepts, are created and refined in AIIRE ontology rather often, so it is quite difficult and, maybe, unnecessary to describe them all. At the moment, 312 different concepts are used as relations in the ontology, some of them being very specific, e.g., there is relation named 'to be given to employee' which binds any resource of professional occupation (including salary, other employees, vacations, etc.) with employees that receive these resources from their employers. Relations like this one are never directly denoted by constituent classes, but are rather often involved in semantic graphs, because of the underlying algorithm of semantic analysis.

This algorithm simply repeats the rules of relation inheritance and overriding, so that, e.g., genitive constructions (or similar constructions with preposition *of* in languages without morphological genitive) are initially provided with a pattern of semantic graph, which consists of two positions to fulfill (these come from syntactic head and specifier meanings) and an abstract relation ‘to belong to object’ between them. If any concept that is denoted by syntactic constituents is not able to participate in this relation according to relation-overriding rules, then the whole construction is treated as semantically uninterpretable and is culled. If the subject position is fulfilled, then the relation can be substituted with any of its subclasses that override this relation within the subject concept. The same is true for the object. Furthermore, relation substitution can cause both subject and object concepts substitution, if its backward links to subject or object are overridden. E.g., during evaluation of the expression Peter’s vacations the upper-mentioned relation ‘to belong to object’ is substituted with ‘to belong as resource to a person’ (because vacations are a resource), then with ‘to be given to employee’, because vacations are resource of professional occupation, and then *Peter*, who was treated as an instance of ‘person’, is substituted with an isomorphic instance of ‘employee’, because ‘to be given to employee’ has ‘person’ as subject overridden by ‘employee’. This is the basic scheme implemented in AIIRE, the real algorithm is much more complex, as it has to consider typical representatives, backbone relations, polysemy of each syntactic constituent and many other upper-mentioned peculiarities of the ontology.

It is worth emphasizing, that there are many problems concerning the ontology now. Semantic restrictions are very strict, so very often it is the case that they lead to impossibility of binding and need to be weakened somehow. Sometimes, but much more rarely, they are too weak, and ambiguity remains unresolved. However, ontology provides an interface to manipulate semantic restrictions directly, and, as the whole system is free and open-source, and, more than that, available for download from CentOS and NauLinux repositories, the project develops rather rapidly. It seems also worth mentioning, that the author of this paper has defended a PhD thesis in computational linguistics, devoted to automatic classification of news messages using syntactic semantics analysis performed by AIIRE

Bibliography

[Fellbaum, 1998] Christiane Fellbaum (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

[Hajic et al., 2001] J. Hajic, P. Krbec, P. Kveton, K. Oliva, V. Petkevic (2001). Serial Combination of Rules and Statistics: A Case Study in Czech Tagging. In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001), — Toulouse, France, 2001

[Knowledege base, 2014] Knowledge base. (2014, June 12). In Wikipedia, The Free Encyclopedia. Retrieved 21:23, July 3, 2014, from

http://en.wikipedia.org/w/index.php?title=Knowledge_base&oldid=612589599

[Leontyeva, 2006] Н.Н. Леонтьева (2006). Автоматическое понимание текстов. Системы, модели, ресурсы. Москва: Academia, 2006 — 303 с.

[Miller, 1995] George A. Miller (1995). WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.

[Petkevič, 2006] Petkevič V. (2006): Reliable Morphological Disambiguation of Czech: Rule-Based Approach is Necessary. In Insight into the Slovak and Czech Corpus Linguistics (Šimková M. ed.). Veda (Publishing House of the Slovak Academy of Sciences & Ludovít Štúr Institute of Linguistics of the Slovak Academy of Sciences), Bratislava, pp. 26–44, ISBN 80-224-0880-8.

[Sheffczyk et al. 2006] Scheffczyk J., Pease A., Ellsworth M. (2006). Linking FrameNet to the Suggested Upper Merged Ontology. Proceedings of the International Conference on Formal Ontology in Information Systems (FOIS 2006), Baltimore, Maryland, pp. 289-300 — November, 2006

[Sowa, 1999] J.F. Sowa (1999). Conceptual Graphs: Draft Proposed American National Standard. In International Conference on Conceptual Structures ICCS-99, Lecture Notes in Artificial Intelligence 1640 — Berlin, New York: Springer Verlag, 1999 — pp. 1-65

[Tseitin, 1985] Г.С. Цейтин (1985) Программирование на ассоциативных сетях // ЭВМ в проектировании и производстве. -Л.: Машиностроение, 1985. Вып. 2.

[Zouaq et al. 2009] A. Zouaq, M. Gagnon, B. Ozell (2009). A SUMO-based Semantic Analysis for Knowledge Extraction. In Proceedings of the 4th Language & Technology Conference — Poznań, 2009

Authors' Information

Alexey Dobrov – Saint-Petersburg State University, Assistant Professor; Saint-Petersburg National University of Informational Technologies, Mechanics and Optics, tutor; Geline Company, head of the RnD dept.; Ramax International, software developer; P.O. Box: 199226, Novosmolenskaya emb., 4, apt. 71, St. Petersburg, Russia; e-mail: adobrov@aiire.org

Major Fields of Scientific Research: Artificial Intelligence, Natural Language Understanding, Syntax, Semantics, Ontologies

COMPARING METHODS OF AUTOMATIC VERB-NOUN COLLOCATION EXTRACTION

Svetlana Koshcheeva, Victor Zakharov

***Abstract**: Automatic verb-noun collocation extraction is an important natural language processing task, the results of which can be applied in various spheres including machine translation, language teaching, summarization, information extraction, disambiguation, etc. The paper describes a set of experiments the aim of which is to compare several approaches to automatic verb-noun collocation extraction. The main subjects under observation are the impact of span size and POS-filtering on the quality of collocation extraction. The experiments have shown that collocations lists extracted by means of POS-filtering are significantly more precise than those obtained without POS-filtering, whereas the extension of a span size has an ambiguous effect. On the one hand, it enables the extraction of distant collocates, but on the other hand it results in erroneous collocates, which leads therefore to consider the use of syntax-based approach for verb-noun collocation extraction.*

***Keywords**: Corpora, statistical methods, collocations, automatic collocation extraction.*

***ACM Classification Keywords**: I.2.7 Natural Language Processing*

Introduction

The study of collocations is a state-of-the-art approach to the analysis of compatibility of lexical units. Collocation extraction is applied in many fields of linguistic technologies such as semantic and lexicographic research (including the creation of dictionaries and grammars of a new type), language teaching, machine translation, automatic analysis and disambiguation.

Statistical methods of collocation extraction have become widely used in modern corpus linguistics. The simplest way to detect collocations in texts is making up frequency lists of words, which appear to the left or to the right of a node word within the indicated span. The size of this span is usually 5 words to the left or to the right of a node. Statistical association measures (log-likelihood, MI, t-score etc.) have become extremely widespread in modern linguistic research. These measures are based on cooccurrence frequencies of word pairs and frequencies of each constituent (e.g., see [Peina, 2009]), which can as well be calculated within a certain span.

In our research as an association and ranking measure we use Mutual Information (MI) [Church, 1991] which can be understood as a coefficient of syntagmatic strength between collocation constituents [Evert, 2004]. MI for a bigram is calculated by the following formula:

$$MI = log_2 \frac{f(n,c) * N}{f(n) * f(c)} \tag{1}$$

where *MI* is mutual information measure; *f (n, c)*, *f(n)*, *f(c)* are absolute frequencies in a corpus of cooccurrence of *n* (node) and *c* (collocate) and words *n, c* respectively; *N* is a corpus size.

However, it is necessary to take into account the fact that words are syntactically related, they do not occur in texts haphazardly or by chance. Therefore, collocation extraction requires not only statistical methods but also syntax-based approaches, which take into consideration morphological and syntactic properties of words in a corpus.

The aim of this paper is to examine and compare methods of automatic collocation extraction. As a tool for our investigation we have chosen IntelliText, a system developed by the Centre for Translation Studies (CTS) at the University of Leeds (http://corpus.leeds.ac.uk/it/). IntelliText offers ample opportunities for linguistic research and has representative corpora, including morphologically annotated Russian corpora.

Methodology

The subject of our investigation are verb-noun collocations of the type *"verb + noun* (in the accusative without a preposition)". As a tool for verb-noun collocation extraction we used the system IntelliText. The corpus RNC2010-MOCKY, a 2010 version of the Russian National Corpus of 116 mln words, served as a material for our research.

For our experiments we have chosen the following verbs: *выполнять* ('to carry out'), *нарушать* ('to violate'), *принимать* ('to accept'). On the basis of the Dictionary of Collocations in the Russian Language [Slovar' sochetaemosti slov russkogo jazyka, 1983], the Dictionary of Russian Verb-Noun Collocations [Deribas, 1983] and Small Academic Dictionary [Slovar' russkogo jazyka, 1981 – 1984] we made up a list of verb-noun collocations for the chosen verbs with which we compared word-combinations extracted by IntelliText. Below one can see the list of verb-noun collocations for the verb *выполнять*:

1. выполнять директивы
2. выполнять долг
3. выполнять желание
4. выполнять задание
5. выполнять задачу
6. выполнять заказ
7. выполнять заявку
8. выполнять инструкцию
9. выполнять каприз
10. выполнять команду

11.выполнять нагрузку

12. выполнять наказ
13. выполнять норму
14. выполнять обещание
15. выполнять обязанности
16. выполнять обязательства
17. выполнять план
18. выполнять поручение
19. выполнять правила
20. выполнять приказ
21. выполнять приказание
22. выполнять программу
23. выполнять просьбу
24. выполнять работу
25. выполнять распоряжение
26. выполнять решение
27. выполнять роль
28. выполнять совет
29. выполнять требование
30. выполнять указание
31. выполнять упражнение
32. выполнять условие
33. выполнять установку
34. выполнять функцию

The aim of the first experiment was to study the impact of part of speech (POS) filtering on the quality of collocation extraction (specifying the POS of a candidate collocate). The spans [-1, 1], [-2, 2], [-3, 3] were under observation.

The second experiment consisted in investigating the impact of a span size on the quality of verb-noun collocation extraction with POS-filtering. Spans up to 5 words to the right of the verb were studied.

In the course of these experiments it was discovered that among phrases extracted by IntelliText there are those which are not fixed in the dictionaries mentioned above but which nevertheless can be considered as set expressions, i.e. collocations. In order to decide which phrases extracted by IntelliText for the verb *выполнять* are collocations an expert evaluation was carried out. According to this evaluation our list of verb-noun collocations was expanded by the following phrases: *выполнять волю, выполнять движение, выполнять действие, выполнять контракт, выполнять маневр, выполнять миссию, выполнять обработку, выполнять операцию, выполнять пожелание, выполнять предписание, выполнять прыжок, выполнять расчёт, выполнять рекомендацию, выполнять соглашение, выполнять трюк.*

To estimate the quality of verb-noun collocation extraction we calculated precision, recall and F-measure using the following formulas:

$$Precision = \frac{|D_{rel} \cap D_{retr}|}{|D_{retr}|} \tag{2}$$

where D_{rel} is a set of relevant expressions from the list of verb-noun collocations; D_{retr} – a set of phrases extracted by IntelliText.

$$Recall = \frac{|D_{rel} \cap D_{retr}|}{|D_{rel}|} \tag{3}$$

where D_{rel} is a set of relevant expressions from the list of verb-noun collocations; D_{retr} – a set of phrases extracted by IntelliText.

$$F = (\beta^2{+}1)\,\frac{Precision \times Recall}{\beta^2 Precision + Recall} \tag{4}$$

for β = 2 (the priority is given to recall).

Experiments

In Table 1 one can see results of the first experiment (the impact of POS-filtering on verb-noun collocation extraction) for the verb выполнять ('to carry out'). Ranks of collocate candidates are indicated in the first column, other columns contain collocate candidates extracted by IntelliText for different spans, with and without POS-filtering. Words highlighted in white are collocates for the verb выполнять ('to carry out') according to our list of verb-noun collocations.

Table 1. Verb-noun collocation extraction by IntelliText (with and without POS-filtering) for the verb выполнять ('to carry out')

rank	[-1, 1]		[-2, 2]		[-3, 3]	
	without POS-filtering	**noun in the accusative**	**without POS-filtering**	**noun in the accusative**	**without POS-filtering**	**noun in the accusative**
1.	организационные	приказание	организационные	приказание	перевыполнять	приказание
2.	пунктуально	функция	перевыполнять	функция	беспрекословно	функция
3.	беспрекословно	наказ	беспрекословно	поручение	неукоснительно	поручение
4.	неукоснительно	поручение	неукоснительно	наказ	функция	наказ
5.	добросовестно	задание	добросовестно	задание	добросовестн	задание

					о	
6.	приказание	приказ	приказание	обязанность	приказание	обязанность
7.	функция	упражнение	функция	предписание	поручение	предписание
8.	скрупулёзно	указание	поручение	приказ	задание	обещание
9.	безукоризнено	заказ	наказ	обещание	обязанность	заказ
10.	наказ	предписание	задание	заказ	упражнение	приказ
11.	безропотно	обязанность	упражнение	упражнение	обещание	упражнение
12.	поручение	роль	обещание	обязательство	заказ	обязательство
13.	приказ	директива	обязанность	указание	предписание	миссия
14.	задание	обещание	приказ	директива	приказ	указание
15.	исправно	рейс	заказ	миссия	обязательство	директива
16.	упражнение	инструкция	предписание	трюк	указание	предназначение
17.	возложить	завет	обязательство	роль	миссия	долг
18.	послушно	обязательство	указание	долг	возложить	трюк
19.	интернациональный	распоряжение	возложить	распоряжение	долг	роль
20.	заказ	задача	послушно	предназначение	защитный	распоряжение
21.	указание	маневр	миссия	завет	роль	задача
22.	предписание	трюк	трюк	задача	задача	завет
23.	управленческий	требование	защитный	пожелание	распоряжение	пожелание
24.	обязанность	приветствие	долг	приветствие	рейс	каприз
25.	обещание	миссия	роль	рейс	инструкция	рейс
26.	качественно	пожелание	распоряжение	инструкция	воинский	инструкция
27.	роль	полёт	задача	каприз	обязанный	приветствие
28.	защитный	заповедь	рейс	маневр	требование	просьба
29.	призванный	норма	эффективно	требование	служебный	требование
30.	эффективно	рекомендация	инструкция	рекомендация	успешно	заповедь
31.	обязательство	обряд	воинский	просьба	просьба	рекомендация
32.	успешно	план	обязанный	заповедь	работа	бросок
33.	рейс	воля	требование	обряд	строго	маневр
34.	старательно	просьба	успешно	воля	норма	обряд
35.	инструкция	долг	служебный	полёт	операция	норма
36.	обязанный	блок	рекомендация	норма	полёт	воля
37.	распоряжение	работа	строго	ритуал	боевой	работа
38.	строго	стрельба	просьба	работа	способный	ритуал

39.	задача	посадка	полёт	план	честно	полёт
40.	ежедневно	команда	работа	стрельба	план	план
41.	чётко	операция	способный	перевозка	воля	посадка
42.	полёт	удар	чётко	посадка	команда	перевозка
43.	честно	готовность	норма	операция	объём	монтаж
44.	способный	условие	боевой	команда	позволять	классификация
45.	требование	танец	честно	прыжок	одновременно	стрельба
46.	самостоятельно	поворот	воля	заявка	удар	операция
47.	боевой	множество	операция	блок	определённый	назначение
48.	служебный	программа	план	спектр	сложный	команда
49.	важнейший	фигура	позволять	назначение	приходиться	процедура
50.	отказываться	желание	одновременно	программа	правило	прыжок
51.	броситься	решение	команда	процедура	действие	заявка
52.	норма	ряд	объём	ремонт	важный	спектр
53.	одновременно	правило	удар	установка	различный	блок
54.	фактически	способность	определённый	желание	движение	желание
55.	позволять	расчёт	приходиться	удар	предприятие	нагрузка
56.	план	закон	сложный	условие	свой	программа
57.	приходиться	контроль	прямой	фигура	любой	установка
58.	работа	возможность	важный	цикл	программа	действие
59.	определённый	движение	различный	обслуживание	социальный	удар
60.	воля	шаг	свой	правило	организация	ремонт
61.	просьба	действие	правило	действие	должен	обслуживание
62.	команда	срок	любой	расчёт	условие	правило
63.	долг	качество	движение	контракт	решение	комплекс
64.	способность	сила	программа	готовность	который	условие
65.	объём	-	должен	решение	лишь	фигура
66.	свой	-	социальный	объём	также	решение
67.	операция	-	действие	танец	всегда	цикл
68.	различный	-	предприятие	поворот	они	расчёт
69.	сложный	-	условие	множество	весь	объём
70.	удар	-	организация	контроль	надо	контракт
71.	любой	-	решение	комплекс	каждый	готовность
72.	специальный	-	основной	приём	по	движение
73.	должен	-	государственный	ряд	этот	контроль
74.	важный	-	закон	услуга	или	танец
75.	поставить	-	также	движение	другой	приём

76.	продолжать	-	всегда	закон	мочь	поворот
77.	условие	-	лишь	соглашение	только	услуга
78.	прийтись	-	который	способность	его	множество
79.	точно	-	она	исследование	тот	способность
80.	следующий	-	надо	разработка	,	закон
81.	лишь	-	весь	шаг	и	ряд
82.	также	-	можно	проект	не	течение
83.	всегда	-	мочь	товар	они	срок
84.	решение	-	только	анализ	же	урок
85.	начать	-	этот	техника	)	исследование
86.	весь	-	другой	возможность	она	соглашение
87.	они	-	его	очередь	мы	шаг
88.	надо	-	или	течение	то	разработка
89.	можно	-	по	срок	.	реализация
90.	мочь	-	тот	совет	но	проект
91.	который	-	не	число	:	техника
92.	его	-	он	организация	"	очередь
93.	только	-	она	часть	что	счёт
94.	тот	-	мы	счёт	быть	круг
95.	этот	-	и	время	в	возможность
96.	не	-	,	качество	а	совет
97.	она	-	)	право	как	число
98.	быть	-	быть	положение	"	время
99.	он	-	"	процесс	с	часть
100.	,	-	.	день	на	день
101.	и	-	в	дело	–	дело

In Table 2 and Figures 1, 2 and 3 one can see evaluation of the results of the first experiment for the spans [-1, 1] (see Fig. 1), [-2, 2] (see Fig. 2) and [-3, 3] (see Fig. 3).

Table 2. Evaluation of the results of verb-noun collocation extraction for the verb выполнять ('to carry out')

Span	Precision	Recall	F-measure
[-1, 1] (without POS-filtering)	0,287129	0,591837	0,488215
[-1, 1] (with POS-filtering)	0,625	0,816327	0,769231
[-2, 2] (without POS-filtering)	0,336634	0,693878	0,572391
[-2, 2] (with POS-filtering)	0,465347	0,959184	0,791246
[-3, 3] (without POS-filtering)	0,316832	0,653061	0,538721
[-3, 3] (with POS-filtering)	0,485149	1	0,824916

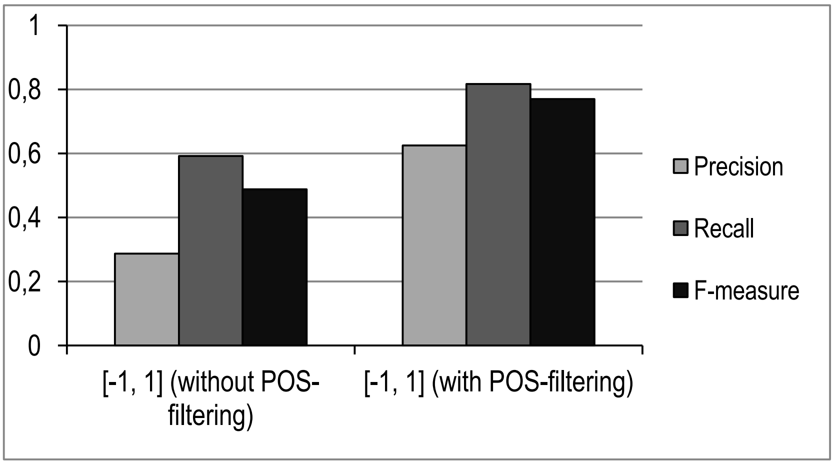


Fig. 1. Evaluation of the results of verb-noun collocation extraction for the verb выполнять ('to carry out') (span = [-1, 1])

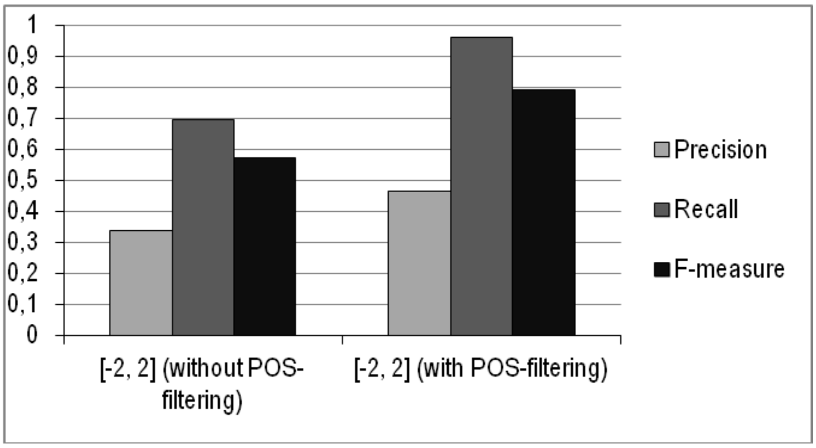


Fig. 2. Evaluation of the results of verb-noun collocation extraction for the verb выполнять ('to carry out') (span = [-2, 2])

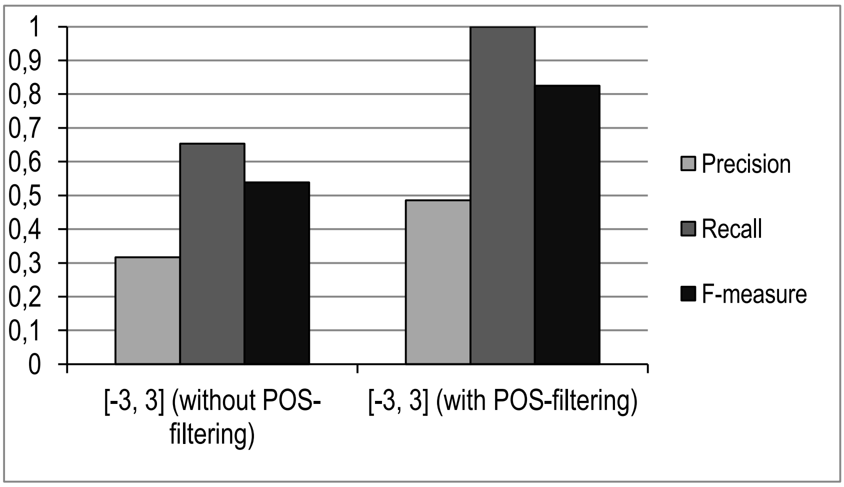


Fig. 3. Evaluation of the results of verb-noun collocation extraction for the verb выполнять ('to carry out') (span = [-3, 3])

Figures 1, 2 and 3 show the same tendency: POS-filtering raises precision, recall and F-measure. This observation leads us to conclude that POS-filtering increases the quality of verb-noun collocation extraction. At the same time, the influence of a span size on verb-noun collocation extraction is not essential.

The analysis of the phrases extracted by IntelliText in the first experiment has shown that the left context doesn't provide relevant collocates for the verbs within our research. It can be explained by the fact that in the Russian language the direct object (a noun in the accusative) comes after the verb. Therefore, in the second experiment we confined ourselves to the right context, limiting the span up to 5 words.

For the indication of a span size we use the signs [0, 1], [0, 2], [0, 3], [0, 4] and [0, 5], which mean a zero left context and a right context equal to one, two, three, four and five words respectively.

Results of the second experiment (the impact of a span size on verb-noun collocation extraction for the verb *выполнять* ('to carry out')) are presented in Table 3. Ranks of collocate candidates are indicated in the first column, other columns contain nouns in the accusative extracted by IntelliText for different right contexts of the verb. Nouns highlighted in white are collocates for the verb *выполнять* ('to carry out') according to our list of verb-noun collocations.

Table 3. Verb-noun collocation extraction for the verb *выполнять* ('to carry out') by IntelliText (span size changed)

rank	**[0, 1]**	**[0, 2]**	**[0, 3]**	**[0, 4]**	**[0, 5]**
1.	приказание	приказание	приказание	приказание	функция
2.	наказ	функция	функция	функция	приказание
3.	функция	поручение	поручение	поручение	поручение
4.	поручение	наказ	задание	задание	задание
5.	задание	задание	наказ	наказ	наказ
6.	приказ	предписание	обязанность	гидроизоляция	гидроизоляция
7.	упражнение	обязанность	предписание	обязанность	обязанность
8.	заказ	заказ	заказ	предписание	предписание
9.	указание	приказ	обещание	заказ	заказ
10.	предписание	обещание	приказ	обязательство	обязательство
11.	роль	указание	обязательство	обещание	обещание
12.	директива	обязательство	упражнение	приказ	приказ
13.	обязанность	упражнение	указание	упражнение	упражнение
14.	рейс	директива	миссия	указание	указание
15.	обещание	миссия	предназначение	миссия	миссия
16.	инструкция	трюк	долг	предназначение	предназначение

17.	завет	распоряжение	директива	директива	долг
18.	распоряжение	роль	трюк	долг	директива
19.	маневр	долг	роль	трюк	трюк
20.	обязательство	завет	распоряжение	каприз	распоряжение
21.	требование	предназначение	пожелание	распоряжение	роль
22.	задача	пожелание	каприз	роль	каприз
23.	пожелание	задача	завет	задача	задача
24.	полёт	рейс	задача	пожелание	пожелание
25.	заповедь	инструкция	рейс	завет	завет
26.	норма	каприз	инструкция	рейс	рейс
27.	просьба	требование	просьба	инструкция	инструкция
28.	воля	заповедь	требование	требование	присяга
29.	рекомендация	просьба	бросок	просьба	просьба
30.	план	маневр	рекомендация	бросок	требование
31.	блок	рекомендация	заповедь	рекомендация	бросок
32.	посадка	воля	маневр	заповедь	рекомендация
33.	работа	полёт	обряд	маневр	заповедь
34.	долг	обряд	ритуал	обряд	маневр
35.	удар	ритуал	воля	ритуал	обряд
36.	операция	норма	полёт	воля	ритуал
37.	команда	работа	норма	монтаж	воля
38.	условие	перевозка	работа	полёт	норма
39.	танец	план	монтаж	норма	монтаж
40.	поворот	стрельба	посадка	работа	полёт
41.	множество	прыжок	перевозка	перевозка	работа
42.	фигура	заявка	план	посадка	перевозка
43.	программа	посадка	операция	план	посадка
44.	ряд	операция	назначение	операция	план
45.	решение	блок	классификация	классификация	операция

46.	правило	команда	стрельба	прыжок	классификация
47.	расчёт	спектр	команда	назначение	прыжок
48.	желание	назначение	прыжок	стрельба	стрельба
49.	контроль	программа	заявка	команда	назначение
50.	закон	процедура	спектр	заявка	команда
51.	движение	ремонт	блок	спектр	заявка
52.	действие	удар	процедура	блок	процедура
53.	-	установка	желание	процедура	спектр
54.	-	условие	программа	действие	блок
55.	-	желание	удар	желание	программа
56.	-	фигура	действие	программа	действие
57.	-	контракт	ремонт	удар	желание
58.	-	действие	установка	ремонт	удар
59.	-	объём	условие	услуга	ремонт
60.	-	танец	правило	условие	установка
61.	-	правило	комплекс	установка	услуга
62.	-	расчёт	фигура	обслуживание	условие
63.	-	поворот	объём	приём	срок
64.	-	множество	расчёт	правило	правило
65.	-	контроль	контракт	комплекс	обслуживание
66.	-	решение	решение	фигура	приём
67.	-	комплекс	танец	объём	комплекс
68.	-	приём	приём	расчёт	фигура
69.	-	ряд	поворот	контракт	объём
70.	-	исследование	множество	решение	движение
71.	-	закон	контроль	контроль	расчёт
72.	-	движение	движение	движение	контракт
73.	-	проект	ряд	танец	решение
74.	-	разработка	урок	поворот	обработка

75.	-	анализ	закон	множество	контроль
76.	-	шаг	исследование	разработка	исследование
77.	-	совет	срок	срок	танец
78.	-	часть	проект	ряд	поворот
79.	-	число	услуга	исследование	мероприятие
80.	-	вид	разработка	мероприятие	множество
81.	-	дело	анализ	урок	разработка
82.	-	-	шаг	закон	ряд
83.	-	-	совет	соглашение	течение
84.	-	-	договор	течение	урок
85.	-	-	течение	проект	закон
86.	-	-	счёт	договор	соглашение
87.	-	-	часть	шаг	шаг
88.	-	-	процесс	счёт	проект
89.	-	-	число	анализ	десяток
90.	-	-	дело	предложение	предложение
91.	-	-	вид	совет	счёт
92.	-	-	-	очередь	договор
93.	-	-	-	деятельность	совет
94.	-	-	-	часть	анализ
95.	-	-	-	момент	очередь
96.	-	-	-	организация	деятельность
97.	-	-	-	цель	положение
98.	-	-	-	прогресс	момент
99.	-	-	-	дело	часть
100.	-	-	-	сторона	сторона
101.	-	-	-	время	время

Results of the second experiment show that the extension of a span size leads to the decrease of precision and F-measure whereas recall increases and reaches its maximum value for the spans [0, 4] and [0, 5] (see Fig. 4).

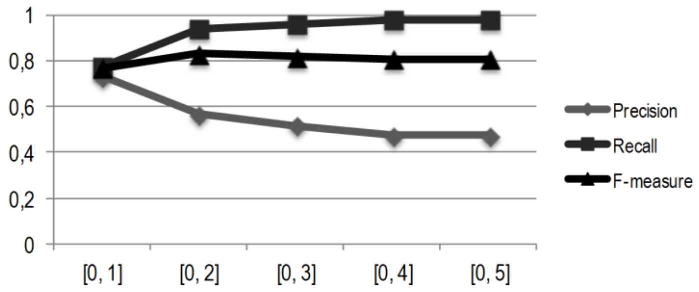


Fig. 4. Evaluation of the results of verb-noun collocation extraction for the verb выполнять ('to carry out')

The experimental data we have received allow us to make the following observations. On the one hand, the extension of a span size enables the extraction of distant collocations (collocations in which components are separated by other words). Below there are some examples of distant collocations for the verb *выполнять ('to carry out')* extracted from the corpus RNC2010-MOCKY for different spans.

Span = [0, 2].

…он продолжает работать, **выполняя** *основные* **функции**, которые не требуют большого потребления…

…выяснение, насколько Америка способна **выполнять** *эту* **роль** гипердержавы. И там нет уверенности…

Span = [0, 3].

…лётчики и подводники, **выполняющие** *особые правительственные* **задания**, офицеры военной разведки…

…Его ум позволяет радоваться, **выполняя** *приказы и* **указания**. Наверно, это круто…

Span = [0, 4].

…в приоритетном порядке **выполнять** *взятые по конвенции* **обязательства**. Президент США…

…постоянно **выполняло** *и перевыполняло государственный* **план**, занимая первое место…

Span = [0, 5].

…Ваша задача научиться **выполнять** *любые атакующие и защитные* **движения** так, будто они уже достигли…

...посетителей на стенде «Линии График», **выполнявших** *какие-то только им понятные* **действия**, привлекало…

On the other hand, the extension of a span size without taking into consideration syntactic relations between words leads to the extraction of erroneous collocates. As examples we have cited the main "false" collocations with the verb *выполнять ('to carry out')* extracted from the corpus RNC2010-MOCKY.

Span = [0, 3].

…Старайтесь с первых же шагов в обучении **выполнять** *предоставленные методы* **действия** без предварительной подготовки…

In this example the collocation **выполнять действия** *is wrong: the noun* **действия** *is related to the noun* **методы**, *not to the verb* **выполнять.**

Span = [0, 4].

…Свою работу эксперты наши **выполняют** *добросовестно, а дальнейшие* **действия** **-** кто в дальнейшем будет манипулировать…

For the verb **выполняют** *IntelliText extracts the noun* **действия** *as a collocate, which is a mistake because this noun refers to the second simple sentence in the compound sentence.*

…Можно заставить подчинённых **выполнять** *работу, отдав соответствующее* **распоряжение**, но такой процесс…

The collocation **выполнять распоряжение** *is erroneous because the noun* **распоряжение** *is a part of the dangling participle* **отдав соответствующее распоряжение** *and is syntactically related to the adverbial participle* **отдав.**

Experiments on verb-noun collocation extraction for the verbs *нарушать* ('to violate') and *принимать* ('to accept') have shown similar results.

In order to solve the problem of erroneous collocates we plan to make a program for collocation extraction which would combine statistical approach (association measures) and syntactic means (taking into account syntactic relations between collocation components).

Conclusion

Our experiments enable us to draw the following conclusions:

1. The quality of verb-noun collocation extraction with POS-filtering is higher than without POS-filtering.

2. In general, the extension of a span size with POS-filtering raises recall, but lowers precision and F-measure of verb-noun collocation extraction.

3. The extension of a span size with POS-filtering has an ambiguous effect. On the one hand, this method makes it possible to extract distant collocations, but on the other hand, it leads to errors in verb-noun collocation extraction, which pays therefore to consider not only the part of speech of a collocate, but syntactic relations between components of a collocation as well.

4. In both experiments one can see that MI is an effective association and ranking measure for verb-noun collocation extraction. Ranks of collocate candidates speak for their relation to collocations in real speech: if a collocate candidate has a high rank, it occurs more often as a component of widely used collocations; if a collocate candidate has a low rank, it is usually a part of free word phrases.

In conclusion, our analysis and comparison of methods of automatic verb-noun collocation extraction allow us to state that a solution to this problem requires a complex syntax-based approach, in other words, it demands a combination of statistical and syntactic methods of verb-noun collocation extraction.

Bibliography

[Pečina, 2009] Pečina P. Lexical Association Measures: Collocation Extraction. Praha, 2009.

[Church, 1991] Church, K., Gale, W., Hanks, P. and Hindle, D. 1991. Using Statistics in Lexical Analysis. Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon. New Jersey, Lawrence Erlbaum. P. 115-164.

[Evert, 2004] Evert, S. 2004. The Statistics of Word Cooccurrences: Word Pairs and Collocations. Ph.D. thesis, University of Stuttgart.

[Slovar' sochetaemosti slov russkogo jazyka, 1983] Slovar' sochetaemosti slov russkogo jazyka. P.N. Denisov, V.V. Morkovkin (ed.). M., 1983.

[Deribas, 1983] Deribas V.M. Ustojchivye glagol'no-imennye slovosochetanija russkogo jazyka. M., 1983.

[Slovar' russkogo jazyka, 1981 – 1984] Slovar' russkogo jazyka: V 4 t. (MAS). A.P. Evgen'eva (ed.). M.: Russkij jazyk, 1981–1984.

Authors' Information

Svetlana Koshcheeva – Student, Saint-Petersburg State University, Universitetskaya emb., 11, Saint-Petersburg 199034 Russia; e-mail: swetik-1995@mail.ru

Major Fields of Scientific Research: Corpus linguistics

Victor Zakharov – Associate Professor, Saint-Petersburg State University, Universitetskaya emb., 11, Saint-Petersburg 199034 Russia; e-mail: vz1311@yandex.ru

Major Fields of Scientific Research: Corpus linguistics, Computational Lexicography

PRACTICAL ASPECTS OF NATURAL LANGUAGE ADDRESSING

Krassimira Ivanova

***Abstract**: NL-addressing is approach for building of a kind of so called "post-relational databases". Some practical aspects of implementation and using of NL-addressing are discussed in this paper. The software realized in this research was practically tested as a part of an instrumental system for automated construction of ontologies "ICON" ("Instrumental Complex for Ontology designatioN") which is under development in the Institute of Cybernetics "V.M.Glushkov" of NAS of Ukraine. In this paper we briefly present ICON and its structure. Attention is paid to the storing of internal information resources of ICON realized on the base of NL-addressing and experimental programs WordArM and OntoArM.*

***Keywords**: Natural Language Addressing, Post-Relational Databases*

***ACM Classification Keywords**: H.2 Database Management; H.2.8 Database Applications*

Introduction

In this research we follow the proposition of Kr. Markov to use the computer encoding of name's (concept's) letters as logical address of connected to it information stored in a multi-dimensional numbered information spaces [Markov, 1984; Markov, 2004; Markov, 2004a]. This way no indexes are needed and high speed direct access to the text elements is available. It is similar to the natural order addressing in a dictionary where no explicit index is used but the concept by itself locates the definition. For this case we use the term: "Natural Language Addressing" (NL-addressing) [Ivanova et al, 2013a].

The idea of NL-addressing is to use encoding of the name both as relative address and as route in a multi-dimensional information space and this way to speed the access to stored information. For instance, let have the next definition: "London: The capital city of England and the United Kingdom, and the largest city, urban zone and metropolitan area in the United Kingdom, and the European Union by most measures".

In the computer memory, for example, it may be stored in a file at relative address "00084920" and the index couple is: ("London", "00084920"). At the memory address "00084920" the main text, "The capital … measures." will be stored. To read/write the main text, firstly we need to find name "London" in the index and after that to access memory address "00084920" to read/write the definition.

If we assume that name "London" in the computer memory is encoded by six numbers (letter codes), for instance by using ASCII encoding system London is encoded as (76, 111, 110, 100, 111, 110), than we may use these codes for direct address to memory, i.e. ("London", "76, 111, 110, 100, 111, 110").

Above we have written two times the same name as letters and codes. Because of this we may omit this couple and index, and read/write directly to the address "76, 111, 110, 100, 111, 110".

For human this address will be shown as "London", but for the computer it will be "76, 111, 110, 100, 111, 110".

Till now, NL-addressing has been presented in several publications [Ivanova et al, 2012a; 2012b; Ivanova et al, 2013a; 2013b; 2013c; 2013d; 2013e; Ivanova, 2013; Ivanova, 2014a].

Some practical aspects of implementation and using of NL-addressing are discussed in this paper. The software realized in this research was practically tested as a part of an instrumental system for automated construction of ontologies "ICON" ("Instrumental Complex for Ontology designatioN") which is under development in the Institute of Cybernetics "V.M.Glushkov" of NAS of Ukraine. In this paper we briefly present ICON and its structure. Attention is paid to the storing of internal information resources of ICON realized on the base of NL-addressing and experimental programs WordArM and OntoArM.

The transition to non-relational data models

Some of the world's leading companies and products which support extra large ontology bases are presented on page of W3C [LTS, 2012]. It should be noted, there exists a gradual transition from relational to non-relational models for organizing ontological data. The graph oriented approach for storing ontologies became one of the preferred. Perhaps the most telling example is the system AllegroGraph® 4.9 [AlegroGraph, 2012] of the FRANZ Inc. [Franz Inc., 2013]. AllegroGraph is a modern, high-performance, persistent graph database. AllegroGraph uses efficient memory utilization in combination with disk-based storage, enabling it to scale to billions of quads while maintaining superior performance. AllegroGraph supports SPARQL, RDFS++, and Prolog reasoning from numerous client applications [AlegroGraph, 2012]. The driving force has been AIDA platform of Amdocs Product Enabler Group (Amdocs). The "Amdocs Intelligent Decision Automation" (AIDA) is an engine that is powered by Franz AllegroGraph 4.0 real-time semantic technology [Guinn & Aasman, 2010]. AllegroGraph provides dynamic reasoning and DOES NOT require materialization. AllegroGraph's RDFS++ engine dynamically maintains the ontological entailments required for reasoning; it has no explicit materialization phase.

Materialization is the pre-computation and storage of inferred triples so that future queries run more efficiently. The central problem with materialization is its maintenance: changes to the triple-store's ontology or facts usually change the set of inferred triples. In *static* materialization, any change in the store *requires complete re-processing before new queries can run*. AllegroGraph's dynamic materialization simplifies store maintenance and reduces the time required between data changes and querying. AllegroGraph also has RDFS++ reasoning with built in Prolog.

Post-relational data bases give new possibilities but are not aimed to replace RDBMS. Both have one main goal – *to store data effectively*. Because of this, it is not correct to claim one against another.

In addition, many new approaches are built over the RDBMS platforms. In the same time, it is important to point main features of RDF triple stores which make them preferable.

Steve Harris, the CTO* of a company that extensively uses RDF triple stores commercially, has outlined the "five main features" of RDF triple stores which make them preferable [TSRD, 2012]:

- *Schema flexibility* - it's possible to do the equivalent of a schema change to an RDF store live, and without any downtime, or redesign - it's not a free lunch, you need to be careful with how your software works, but it's a pretty easy thing to do;
- *More modern* - RDF stores are typically queried over HTTP it's very easy to fit them into Service Architectures without performance penalties. Also they handle internationalized content better than typical SQL databases - e.g. you can have multiple values in different languages;
- *Standardization* - the level of standardization of implementations using RDF and SPARQL is much higher than SQL. It's possible to swap out one triple store for another, though you have to be careful you're not stepping outside the standards. Moving data between stores is easy, as they all speak the same language;

* CTO: Chief Technology Officer or Chief Technical Officer is an executive-level position in a company or other entity whose occupant is focused on scientific and technological issues within an organization.

- *Expressivity* - it's much easier to model complex data in RDF than in SQL, and the query language makes it easier to do things like LEFT JOINs (called OPTIONAL in SPARQL). Conversely though, if you data are very tabular, then SQL is much easier;
- *Provenance* - SPARQL lets you track where each piece of information came from, and you can store metadata about it, letting you easily do sophisticated queries, only taking into account data from certain sources, or with a certain trust level, on from some date range etc.

There are downsides though. SQL databases are generally much more mature, and have more features than typical RDF databases. Things like transactions are often much more crude, or nonexistent. Also, the cost per unit information stored in RDF vs. SQL is noticeably higher. It's hard to generalize, but it can be significant if you have a lot of data - though at least in our case it's an overall benefit financially given the flexibility and power [TSRD, 2012].

The flexibility of triple stores is very important for solving of two considerable practical problems: building and using of domain ontologies and, directly connected to it, building and using of ontologies of text documents.

Domain ontologies

Domain ontologies are formal descriptions of the classes of concepts and the relationships among those concepts that describe an application area. In other words, domain ontology models concepts and relationships that are relevant to the given domain (e.g., biology, architecture, software engineering) [Witte et al, 2010]. Building domain ontologies is not a simple task when domain experts have no background knowledge on engineering techniques and/or they have not much time to invest in domain conceptualization.

In order to develop domain ontology some methodology has to be followed. For instance, such methodology is the “METHONTOLOGY Framework” developed within the Ontological Engineering group at Universidad Politécnica de Madrid [Fernández et al, 1997]. This methodology enables the construction of ontologies at the knowledge level, and has its roots in the main activities identified by the IEEE software development process and in other knowledge engineering methodologies. METHONTOLOGY guides in how to carry out the whole ontology development through the specification, the conceptualization, the formalization, the implementation and the maintenance of the ontology [Corcho et al, 2005]. The METHONTOLOGY framework provides the idea of support activities: Knowledge Acquisition and Validation/Verification. It is divided into three phases: *Specification*, *Conceptualization* and *Implementation*. These phases constitute an iterative process [Brusa et al, 2006].

The “METHONTOLOGY Framework” reduced the existing gap between ontological art and ontological engineering [Fernández et al, 1997] mainly by:

- Identifying a set of activities to be done during the ontology development process. They are: plainly, specify, acquire knowledge, conceptualize, formalize, integrate, implement, evaluate, document, and maintain;
- Proposing the evolving prototype as the life cycle that better fits with the ontology life cycle. The life of ontology moves on through the following states: specification, conceptualization, formalization, integration, implementation, and maintenance. The evolving prototype life cycle allows the ontologies to go back from any state to other if some definition is missed or wrong. So, this life cycle permits the inclusion, removal or modification of definitions *anytime* of the ontology life cycle. Knowledge acquisition, documentation and evaluation are support activities that are carried out during the majority of these states;
- METHONTOLOGY highly recommends the reuse of existing ontologies.

Ontologies of text documents

Creating of ontologies of text documents is based on domain ontology and consists of Document annotation and Ontology population [Amardeilh, 2006]:

- *Document Annotation* consists in (semi-)automatically adding metadata to documents, i.e. providing descriptive information about the content of a document such as its title, its author but mainly the controlled vocabularies as the descriptors of a thesaurus or the instances of a knowledge base on which the document has to be indexed;
- *Ontology Population* aims at (semi-)automatically inserting new instances of concepts, properties and relations to the knowledge base as defined by the domain ontology.

Once Document Annotation and Ontology Population are performed, the final users of an application can exploit the resulting annotations and instances *to query, to share, to access, to publish documents, metadata and knowledge.*

Document Annotation and Ontology Population can be seen as similar tasks.

- Firstly, they both rely on the modeling of terminological and ontological resources (ontologies, thesaurus, taxonomies…) to normalize the semantic of the documentary annotations as well as the concepts of the domain;
- Secondly, as human language is a primary mode of knowledge transfer, they both make use of text-mining methods and tools such as Information Extraction to extract the descriptive structured information from documentary resources or Categorization to classify a document into predefined categories or computed clusters;
- Thirdly, they both more and more rely on the Semantic Web standards and languages such as RDF for annotating and OWL for populating [Amardeilh, 2006].

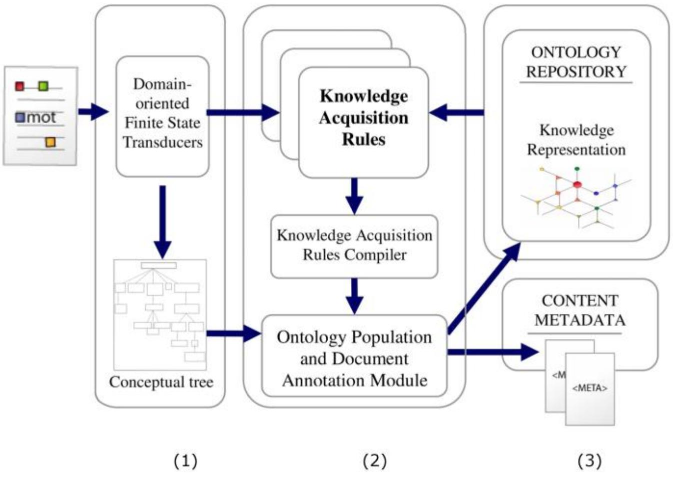

mot
Domain-oriented Finite State Transducers
Conceptual tree
Knowledge Acquisition Rules
Knowledge Acquisition Rules Compiler
Ontology Population and Document Annotation Module
ONTOLOGY REPOSITORY
Knowledge Representation
CONTENT METADATA
<M
<META>
(1)
(2)
(3)

Fig. 1. The OntoPop's platform [Amardeilh, 2006]

The document annotation and ontology population we will illustrate following the OntoPop platform [Amardeilh, 2006]. We have three phases (Figure 1):

(1) Extracting information from semi-structured texts - the text-mining solutions parse a textual resource, creating semantic tags to mark up the relevant content with regard to the domain of concern;

(2) Mapping between the results of the Information Extraction tool and the ontology model - the mediation layer maps the semantic tags produced by the text mining tools into formal representations, being the content annotations (RDF) or the ontology instances (OWL);

(3) Representing and managing the domain ontology, the thesaurus and the knowledge base - the semantic tags are used either to semantically annotate the content with metadata or to acquire knowledge, i.e. to semi-automatically construct and maintain domain terminologies or to semi-automatically enrich knowledge bases with the named entities and semantic relations extracted.

Operations with ontologies stored by NL-addressing

Operations for maintenance and integration of ontologies may be facilitated by using NL-addressing.

NL-addressing permits ontology operations to be realized by operations with corresponded layers of ontologies. It is possible to create a "virtual" ontology by combining only the paths to ontologies without any "real" creation a new one. In this case, the consistency has to be supported dynamically. For instance, after merging ontologies irrespective of the kind of operation result (virtual or real), new ontology will contain a union of the layers of source ontologies.

When same relation (layer) exists in both ontologies, the process of merging may be provided in depth for all existing concepts of layers. The problem to be solved is what to do if in different archives exist concepts (i.e. equal location) but different content. Here we have three variants:

(1) To select concept content of the first ontology;

(2) To select concept content of the second ontology;

(3) To keep both contents and dynamically to make decision what is appropriate.

Our preference is to create virtual ontologies because this will save resources (time and space) and will give new possibilities based on dynamical selection of the content.

Using natural language addressing for storing dictionaries, thesauruses and ontologies, facilitate its realization.

Not all of operations for maintenance and integration of ontologies can be made for all ontologies [Kalfoglou & Schorlemmer, 2003]. In general, these are very difficult tasks that are in general not solvable automatically [Obitko, 2007].

What is common and may be realized is developing of new tools for storing ontologies. At the first place, such tools are RDF-stores.

Building RDF-stores using NL-addressing

The Semantic Web and RDF triple stores are important research themes. Taking in account that NL-addressing is a possibility which may be used in addition to all already existing tools and approaches, below we will outline the main areas of its applicability. It is not correct to claim that NL-addressing will replace one or another tool. It has to be used where it is really effective.

In [Ivanova et al, 2012b] we presented main approaches for creating RDF-triple stores. Below, following that explanation, we will sketch some practicable solutions. Let remember that every RDF-triple consists of three elements – *Subject*, *Relation*, and *Object*.

NL-Addressing for ontology generic schemas

- ***Vertical representation:*** It is easy to realize vertical representation of a triple store via NL-addressing. The values of *Subject* will be the addresses and all couples (*Predicate*, *Object*) for given value may be stored at one and the same address. This way with one operation all edges of a node of the graph will be received. In the multi-layer variant, values of *Predicate* may be names of the layers (archives). In this case, additional operations for reading edges will be needed. The advantage is possibility to work only with selected layers and to reduce the time for access.
- ***Normalized triple store (vertical partitioning):*** The normalized triple store is ready for representing via NL-addressing. We may use *multi-layer variant* where values of *Predicate* may be names of the layers (archives). In this case, additional operations for reading edges will be needed. The advantage is possibility to work only with selected layers and to reduce the time for access. The *Subject* will be the NL-address and only *Object* will be saved. Possibility to concatenate all *Objects* for a given *Subject* reduces the size of memory and access time. In addition, the vertical partitioning approach may be realized directly by the Multi-domain Information Model [Markov, 2004] because it *directly supports the column-oriented DBMS (one column = one information space)*.

✓ **NL-Addressing for ontology specific schemas**

- ***Horizontal representation:*** The horizontal representation is an example of a set of layers. Storing every class in a separate layer (archive) gives possibility to add properties without restructuring existing tables.
- ***Decomposition storage model:*** The decomposition storage model is memory and time consuming due to duplicating the information and generation of too much search indexes. In the same time, it is similar to the NL-addressing style and may be directly implemented using NL-addressing but *this will be not efficient*. NL-addressing permits new possibilities due to omitting of explicit given information – names as well as balanced indexes. The feature tables may be replaced by NL-addressing access to corresponded points of the information space where all information about given *Subject* will exist. This way we will reduce the needed memory and time.
- ***Multiple indexing frameworks:*** The NL-addressing directly supports idea of multi-indexing because of the multi-layer structures and direct access to the *Object* values by NL-address computed on the base of the *Subject* and *Relation* values. Only the *Object*'s index has to be generated if it is really needed.

The above outlined ideas give basis for experiencing in a real software implementation of NL-addressing in ICON.

ICON - Instrumental complex for ontology designation

Design of ontologies, i.e. the formation sets of concepts, relations, axioms, and functions for interpretation, is a laborious process. Manual construction of these sets needs both time and many highly qualified specialists. This determines the development of tools (instrumental complexes) for automation of process of ontology design and distribution. The instrumental complexes for automated construction of ontologies are aimed to be used for the analysis and processing of large volumes of semi-structured data, such as linguistic corpuses in English, Dutch, Russian, Ukrainian, Bulgarian, and others languages.

Such instrumental complex is under development at the Kiev Institute of Cybernetics "V.M.Glushkov" of the National Academy of Sciences of Ukraine with the participation of Bulgarian experts. The complex is called

"**ICON**" ("**I**nstrumental **C**omplex for **O**ntology designatio**N**", from Russian "ИКОН": "**И**нструментальный **К**омплекс **О**нтологического **Н**азначения") [Palagin et al, 2011]. This research is a part of this project and continues work for intelligent systems memory structuring [Gladun, 2003] done during the years.

Information model of ICON is presented in Figure 2 below.

ICON consists of three subsystems: *"Information exchange", "Information processing", and "Internal information resources"*:

- "***Information Exchange***" subsystem is aimed to serve manual or automatic c*ollecting and distributing of information* as well as interface with other subsystems of ICON to support creating, storing, visualization and export of the ontological knowledge. It serves retrieval of relevant to solving problem text documents which are available in the Internet and/or in other electronic collections. It include graphical user interface for knowledge engineers and domain experts, who provide preliminary design of ontologies, control and verification of design results, deciding on degree of completion design and more. Via this subsystem the external information resources can be accessed. They include different sources from local or global information bases and networks, such as:
    - Knowledge resources from given domain - electronic collections of encyclopedic dictionaries, monolingual dictionaries, thesauruses, etc.;
    - Internet resources - sources of text documents and distributed knowledge bases to be used in the process of creating ontologies.
    - Collecting information from external sources is served by the ICON information-retrieval system. It is designed to detect and extract textual documents from various external sources and to create linguistic text corpora based on data from these documents;
- "***Information Processing***" subsystem is a set of *original software modules* that implement relevant algorithms for the ontology' design, and *finished tools*, freely available on the Internet, such as Protégé [protégé, 2012] used as one of the main components in module for visual design. Processing of information includes: automatic natural language processing; knowledge discovery, extraction, representation, construction and verification of semantic structures; integration of ontological knowledge, etc. There are two main groups of processing tools respectively for *Linguistic structures* and *Conceptual structures;*
- "***Internal information resources***" subsystem is aimed to support storing of large dictionaries, thesauruses, and ontologies in specialized electronic libraries based on NL-addressing tools realized in this research. It contains:
    - Linguistic libraries - a kind of electronic linguistic corpus which contains various dictionaries and thesauruses as well as document databases with source and/or processed information, for instance, a Linguistic corpuses of texts - a variety of text documents to be processed; and published documents with received results;
    - Conceptual libraries - they are built during the design or integration of ontologies. They are used to store both source information and finished ontological models.

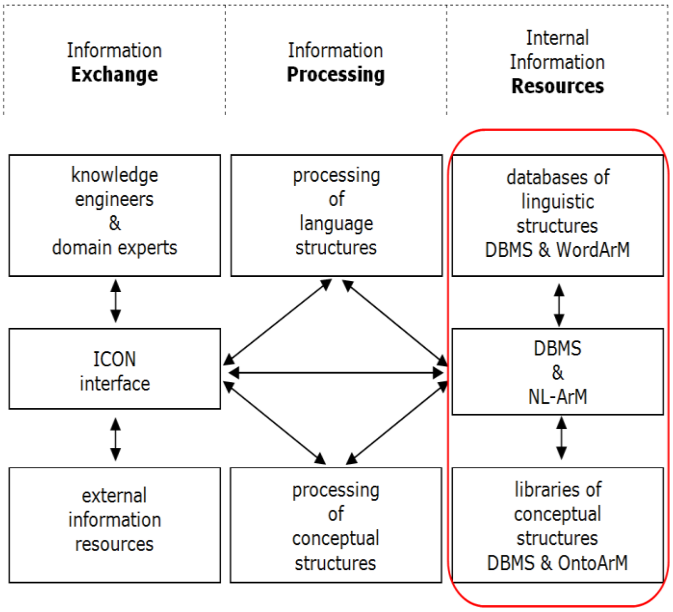

Information **Exchange**
Information **Processing**
Internal Information **Resources**
knowledge engineers & domain experts
processing of language structures
databases of linguistic structures DBMS & WordArM
ICON interface
DBMS & NL-ArM
external information resources
processing of conceptual structures
libraries of conceptual structures DBMS & OntoArM

Fig.2. Information model of ICON

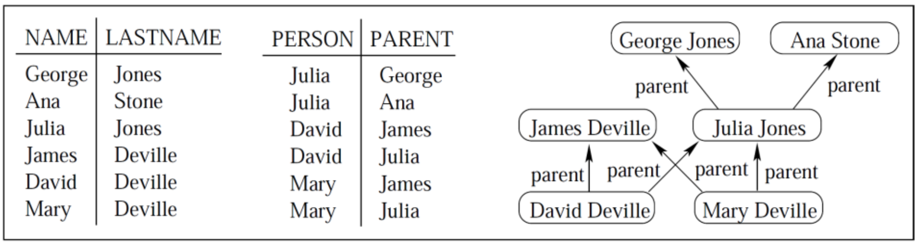
Storing of the internal information resources of ICON

Storing of the internal information resources of ICON is based on several relational DBMS as well as on program modules presented in current research [Ivanova, 2014a]: *WordArM* and *OntoArM*, outlined in [Ivanova, 2014b; Ivanova, 2014c]. The main idea is to extend possibilities of “conventional” tools for semi-structured datasets. Conventional DBMS are used to store some structured information, like sets of descriptions of text documents to be processed.

Some finished tools for processing ontological information have their own databases but they are not appropriate for storing semi-structured information. For instance, such tool is the system Protégé [protégé, 2012]. It is written in Java and allows users to create their own database plug-ins. This choice is also consistent with rest of the Protégé plug-in architecture. Protégé developers chose the simplest schema that one could think of and focused on "maximal change" usage where the class structure and hierarchy is undergoing constant change. For large ontological structures the Protégé approach is not effective and does not support functions for dictionaries and thesauruses. The OWL and RDF descriptions are heavy to be parsed by human.

The proper decision was to integrate Natural Language Addressing together with existing tools and this way to have available all needed functions.

The model which has been chosen is multi-layer storing of graph information. To outline it, let's look at an example - the family tree presented on Figure 4 [Angles & Gutierrez, 2008].



NAME	LASTNAME
George	Jones
Ana	Stone
Julia	Jones
James	Deville
David	Deville
Mary	Deville

PERSON	PARENT
Julia	George
Julia	Ana
David	James
David	Julia
Mary	James
Mary	Julia

Fig. 4. Family tree [Angles & Gutierrez, 2008]

The tree is represented by two tables: “NAME/LASTNAME” and “PERSON/PARENT”. For convenience, the children inherit the father's family.

The "multi-layer" representation of the family tree is given in Table 1.

Table 1. Multi-layer representation of the family tree

		addresses					
		George	Ana	Julia	James	David	Mary
layers	lastname	**Jones**	**Stone**	**Jones**	**Deville**	**Deville**	**Deville**
	parent_of			**George; Ana**		**James; Julia**	**James; Julia**

NL-addressing ***means direct access to content of each cell***. Because of this, for NL-addressing the problem of recompiling the database after updates does not exist. In addition, the multi-layer representation and natural language addressing reduce resources and avoid using of supporting indexes for information retrieval services (B-trees, hash tables, etc.).

Organization of ICON libraries

The ICON internal information resources are stored in libraries which may be of two main types:

- Common libraries, which contain general information used practically by all users and models;
- Local libraries, which contain specific information needed only for given user or model.

In addition, these information resources may be linguistic or conceptual. This way we have a simple taxonomy (Figure 5):

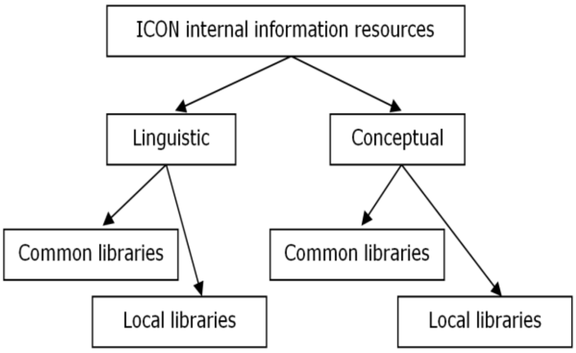


Fig. 5. Taxonomy of ICON internal information resources

Libraries may be installed on single computer or distributed on local network. Special description in a “context” table is used to establish correspondence between names, types, permissions, and allocations (paths) of library archives (files). Common archives are allocated in shared folders. It is possible to have more than one folder with common archives. Updating of common archives may be done after permission from the administrator. Local archives are stored in users’ folders, which may be shared or not, depending of user preferences.

Main difference between common and local archives is in the permissions for updating. Common archives have more strict discipline for making updates – it is obligation of and may be done only by administrators. Updating of local archives is under control of end-user.

ICON Libraries of linguistic structures

Libraries of linguistic structures are organized according different application areas (domains) covered by ICON. The tool for organization of these libraries is WordArM. As a rule there are no interconnections between linguistic archives (files) but there are many connections with conceptual structures where the linguistic information is used.

Common linguistic archives contain dictionaries and thesauruses of general purpose like Ukrainian-English dictionary or WordNet thesaurus of English.

Local linguistic archives contain thematic oriented dictionaries and thesauruses with specific information which concern given practical domain. For instance, it may be Medical thesaurus or Ukrainian-English dictionary of computer science.

One may note that the former ones have same general purposes as previous. This is quite right. What will be declared as common and what as local depends only on decision of administrators about the way of the updating. Common archives may be changed only by administrator, but not by end-user.

We have to point to a special "*Data base of text documents"* which consists of original text documents and linguistic corpuses which are sources for creating the ontologies. In addition, we have to mention the common and local archives with metadata about documents and other information resources. The metadata is closely connected to documents and corresponded resources which are source for conceptual structures. All these information sources are organized using the ArMSpeed tool which is not mentioned in this research and because of this it is not discussed here.

ICON libraries of conceptual structures

ICON conceptual libraries are built during the design or integration of ontologies. There are two kinds of such libraries:

- Library of domain ontologies;
- Library of ontologies of text documents.

These libraries are supported by OntoArM.

✓ ***ICON library of domain ontologies***

Creating and editing domain ontologies in ICON is supported by its original ontological editor [Velychko & Prihodnyuk, 2013]. It is able to read and store ontologies in OWL and XML formats. The ICON Ontological Editor uses functions of OntoArM for saving ontologies. Storing model chosen in ICON is multi-layer storing of ontology graph based on Natural Language Addressing. The preliminary evaluation of the number of layers needed for ICON is about 50 up to 100.

The domain ontology consists of an upper level ontology with a set of sub-ontologies subordinated to it. It is possible sub-ontologies to be stored in subfolders of those of the main ontology but this is not obligatory. Using links (local or global paths) ontology may subordinate several others. This way practically we have ontology network with unlimited size.

Domain ontology is stored in a separate folder. It contains all archives of all its layers. Link to ontology is the path to folder which contains it. Domain ontology may be connected to some linguistic resources –

dictionaries and/or thesauruses. Again the connections are links but this time they point the file of the resource, i.e. the path to it.

- ✓ **ICON library of ontologies of text documents**

A generalized view of OntoArM implementation is shown on Figure 6 (following [Witte et al, 2010]).

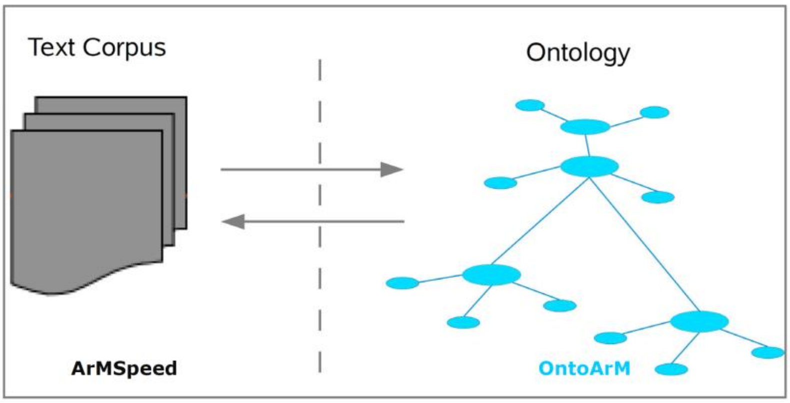


Fig. 6. Using OntoArM for storing ontologies of text documents (following [Witte et al, 2010])

Text corpus and its metadata are stored using ArMSpeed module. Beside NL-addressing, in this module is used search, based on balanced trees.

Ontologies are stored by OntoArM.

Creating and editing ontologies of text documents in ICON is supported by its Ontological Editor based on:

- ArMSpeed for storing documents;
- OntoArM for storing ontologies of text documents, using the same storing model as for domain ontologies. It is multi-layer storing of ontology graph based on Natural Language Addressing.

Ontology of a text document is stored in a separate folder. It contains all archives of all its layers. Link to ontology is the path to folder which contains it. Ontology of the text document may be connected to some linguistic resources – dictionaries and/or thesauruses. The connections are links (paths) to the files of linguistic resources.

ICON methodology for construction of ontologies

ICON follows similar methodology as the “METHONTOLOGY Framework” [Fernández et al, 1997].

It is important to point that ICON methodology permits inclusion, removal or modification of definitions *anytime* of the ontology life cycle. This is very important facility which causes serious problems to conventional databases which have to update permanently their indexing structures and this way to consume large (time and space) resources.

In addition, the processes of document annotation and ontology population ICON are similar to ones of OntoPop platform [Amardeilh, 2006] (Figure 1). NL-addressing is used for knowledge representation in the ontology repository.

NL-addressing facilitates the whole process of ontology development in ICON which includes specification, conceptualization, formalization, implementation and maintenance of ontologies.

Conclusion

Some practical aspects of implementation and using of NL-addressing were discussed in this paper. NL-addressing is approach for building a kind of so called “post-relational databases”. In accordance with this the transition to non-relational data models was outlined.

The implementation has to be done following corresponded methodology for building and using of ontologies. Such known methodology was discussed in the paper. It is called “METHONTOLOGY” and guides in how to carry out the whole ontology development through specification, conceptualization, formalization, implementation and maintenance of the ontology.

Special case is creating of ontologies of text documents which are based on domain ontologies. It consists of Document annotation and Ontology population which we illustrated following the known OntoPop platform [Amardeilh, 2006].

The software realized in this research was practically tested as a part of the instrumental system for automated construction of ontologies "ICON" (“Instrumental Complex for Ontology designatioN”) which is under development in the Institute of Cybernetics “V.M.Glushkov” of NAS of Ukraine.

In this paper we briefly presented ICON and its structure. Attention was paid to the storing of internal information resources of ICON realized on the base of NL-addressing and experimental programs WordArM and OntoArM.

Usefulness of the NL-addressing for creating ontological databases was successfully proved in the practical experiments. During solving concrete problems, new functions based on NL-addressing rise to be realized. For instance, such functions concern work with very large RDF structures. RDF is a graph based data format which is schema-less, thus unstructured, and self-describing, meaning that graph labels within the graph describe the data itself. The prevalence of RDF data is due to variety of underlying graph based models, i.e. almost any type of data can be expressed in this format including relational and XML data [Faye et al, 2012].

Our further research will be directed to several interesting areas of implementing the NL-addressing in business applications where flexibility of this approach will give some new possibilities. Implementing the NL-addressing in linguistic systems which work with large linguistic data sets is another direction for further work.

Bibliography

[AlegroGraph, 2012] AllegroGraph® 4.8, http://www.franz.com/agraph/allegrograph/ (accessed: 25.08.2012).

[Amardeilh, 2006] Florence Amardeilh, “OntoPop or how to annotate documents and populate ontologies from texts”, In Proceedings of the Workshop on Mastering the Gap: From Information Extraction to Semantic Representation (ESWC-06), Budva, Montenegro, 2006 http://hal.archives-ouvertes.fr/docs/00/11/52/55/PDF/amardeilh_ESWC06.pdf (accessed: 31.07.2013)

[Angles & Gutierrez, 2008] Angles R., C. Gutierrez “Survey of Graph Database Models”, ACM Computing Surveys, Vol. 40, No. 1, Article 1, Publication date: February 2008, DOI 10.1145/1322432.1322433, http://doi.acm.org/10.1145/1322432.1322433, pp. 1 – 39

[Brusa et al, 2006] Graciela Brusa, Ma. Laura Caliusco, Omar Chiotti, “A Process for Building a Domain Ontology: an Experience in Developing a Government Budgetary Ontology”, In: M. A. Orgun and T. Meyer, Eds. Proceedings of the second Australasian Workshop on Advances in ontologies (AOW 2006), Hobart, Australia; Conferences in Research and Practice in Information Technology, Vol. 72, pages 7-15; Australian Computer Society, Inc. Darlinghurst, Australia, 2006. ISBN: 1-920-68253-8 http://dl.acm.org/citation.cfm?id=1273661 (accessed: 31.07.2013)

[Corcho et al, 2005] Oscar Corcho, Mariano Fernández-López, Asunción Gómez-Pérez, Angel López-Cima, "Building Legal Ontologies with METHONTOLOGY and WebODE", In: Law and the Semantic Web, Lecture Notes in Computer Science Volume 3369, 2005, pp 142-157, http://link.springer.com/chapter/10.1007%2F978-3-540-32253-5_9 (accessed: 31.07.2013)

[Faye et al, 2012] David C. Faye, Olivier Cure, Guillaume Blin. A survey of RDF storage approaches. Received, December 12, 2011, Accepted, February 7, 2012, ARIMA Journal, vol. 15 (2012), pp. 11-35.

[Fernández et al, 1997] Mariano Fernández, Asunción Gómez-Pérez, Natalia Juristo. "METHONTOLOGY: From Ontological Art towards Ontological Engineering", Spring Symposium on Ontological Engineering of AAAI; Stanford University, California, AAAI TR SS-97-06, 1997, pp 33–40. http://oa.upm.es/5484/1/METHONTOLOGY_.pdf (accessed: 31.07.2013)

[Franz Inc., 2013] Semantic Web Technologies http://www.franz.com/ (accessed: 16.05.2013).

[Gladun, 2003] Gladun, V. P "Intelligent systems memory structuring", International Journal Information Theories and Applications, 10(1), 2003, pp. 10–14.

[Guinn & Aasman, 2010] Guinn B., J. Aasman "Semantic Real Time Intelligent Decision Automation", STIDS 2010 Proceedings, pp. 125-128. http://ceur-ws.org/Vol-713/STIDS_P1_GuinnAasman.pdf (accessed: 15.08.2012)

[Ivanova et al, 2012a] Krassimira Ivanova, Vitalii Velychko, Krassimir Markov. "About NL-addressing" (К вопросу о естествено-языконой адрессации) In: V. Velychko et al (ed.), Problems of Computer in Intellectualization. ITHEA® 2012, Kiev, Ukraine - Sofia, Bulgaria, ISBN: 978-954-16-0061 0 (printed), ISBN: 978-954-16-0062-7 (online), pp. 77-83 (in Russian).

[Ivanova et al, 2012b] Krassimira Ivanova, Vitalii Velychko, Krassimir Markov. "Storing RDF Graphs using NL-addressing", In: G. Setlak, M. Alexandrov, K. Markov (ed.), Artificial Intelligence Methods and Techniques for Business and Engineering Applications. ITHEA® 2012, Rzeszow, Poland; Sofia, Bulgaria, ISBN: 978-954-16-0057-3 (printed), ISBN: 978-954-16-0058-0 (online), pp. 84 – 98.

[Ivanova et al, 2013a] Krassimira B. Ivanova, Koen Vanhoof, Krassimir Markov, Vitalii Velychko, "Introduction to the Natural Language Addressing", International Journal "Information Technologies & Knowledge" Vol.7, Number 2, 2013, ISSN 1313-0455 (printed), 1313-048X (online), pp. 139–146.

[Ivanova et al, 2013b] Krassimira B. Ivanova, Koen Vanhoof, Krassimir Markov, Vitalii Velychko, "Introduction to Storing Graphs by NL-Addressing", International Journal "Information Theories and Applications", Vol. 20, Number 3, 2013, ISSN 1310-0513 (printed), 1313-0463 (online), pp. 263 – 284.

[Ivanova et al, 2013c] Krassimira B. Ivanova, Koen Vanhoof, Krassimir Markov, Vitalii Velychko, "Storing Dictionaries and Thesauruses Using NL-Addressing", International Journal "Information Models and Analyses" Vol.2, Number 3, 2013, ISSN 1314-6416 (printed), 1314-6432(online), pp. 239 - 251.

[Ivanova et al, 2013d] Krassimira B. Ivanova, Koen Vanhoof, Krassimir Markov, Vitalii Velychko, "The Natural Language Addressing Approach", International Scientific Conference "Modern Informatics: Problems, Achievements, and Prospects of Development", devoted to the 90th anniversary of academician V. M. Glushkov. Kiev, Ukraine, 2013, ISBN 978-966-02-6928-6, pp. 214 - 215.

[Ivanova et al, 2013e] Krassimira B. Ivanova, Koen Vanhoof, Krassimir Markov, Vitalii Velychko, "Storing Ontologies by NL-Addressing", IVth All–Russian Conference "Knowledge-Ontology-Theory" (KONT-13), Novosibirsk, Russia, 2013, ISSN 0568 661X, pp. 175 - 184.

[Ivanova, 2013] Krassimira Ivanova, “Informational and Information models”, In Proceedings of 3rd International conference “Knowledge Management and Competitive Intelligence” in the frame of 17th International Forum of Young Scientists “Radio Electronics and Youth in the XXI Century”, Kharkov National University of Radio Electronics (KNURE), Kharkov, Ukraine, Vol.9, 2013, pp 6-7.

[Ivanova, 2014a] Krasimira Ivanova, “Storing Data using Natural Language Addressing”, PhD Thesis, Hasselt University, Belgium, 2014

[Ivanova, 2014b] Krassimira Ivanova, “WordArM - A System for Storing Dictionaries and Thesauruses by Natural Language Addressing”, International Journal “Information Theories and Applications”, Vol. 21, Number 4, 2014, ISSN 1310-0513 (printed), 1313-0463 (online), (in print).

[Ivanova, 2014c] Krassimira Ivanova, “OntoArM - A System for Storing Ontologies by Natural Language Addressing”, International Journal "Information Technologies & Knowledge" Vol. 8, Number 4, 2014, ISSN 1313-0455 (printed), 1313-048X (online), (in print)

[Kalfoglou & Schorlemmer, 2003] Yannis Kalfoglou, Marco Schorlemmer, “Ontology mapping: the state of the art”, The Knowledge Engineering Review, Vol. 18:1, pp. 1–31, Cambridge University Press, United Kingdom, USA, 2003. ISSN = 0269-8889, DOI: 10.1017/S0269888903000651 http://dl.acm.org/citation.cfm?id=975028 (accessed: 31.07.2013)

[LTS, 2012] LargeTripleStores http://www.w3.org/wiki/LargeTripleStores (accessed: 29.08.2012)

[Markov, 1984] K.Markov. A Multi-domain Access Method. Proceedings of the International Conference on Computer Based Scientific Research. PLOVDIV, 1984, pp.558-563.

[Markov, 2004] Markov, K. Multi-domain information model, Int. J. Information Theories and Applications, 11/4, 2004, pp. 303-308.

[Markov, 2004a] Markov, K. Co-ordinate based physical organization for computer representation of information spaces. (Координатно базирана физическа организация за компютърно представяне на информационни пространства.) Proceedings of the Second International Conference “Information Research, Applications and Education” i.TECH 2004, Varna, Bulgaria, Sofia, FOI-COMMERCE – 2004, стр.163-172 (in Bulgarian).

[Obitko, 2007] Obitko M. Ontologies and Semantic Web, 2007 http://www.obitko.com/tutorials/ontologies-semantic-web/operations-on-ontologies.html (accessed: 09.08.2012)

[Palagin et al, 2011] Palagin A.V., Krivii S.L., Petrenko N.G. “Ontological methods and instruments for processing domain knowledge”, (А. В. Палагин, С. Л. Крывый, Н. Г. Петренко. Онтологические методы и средства обработки предметных знаний: монография/Луганск: изд-во ВНУ им. В. Даля, 2011. – 300 с.), (in Russian)

[protégé, 2012] http://protege.stanford.edu (accessed: 25.05.2012)

[TSRD, 2012] Triple Stores vs Relational Databases http://stackoverflow.com/questions/9159168/triple-stores-vs-relational-databases (accessed: 11.01.2013).

[Velychko & Prihodnyuk, 2013] Velychko V.U., Prihodnyuk V.V. Technological tool for graphical design of computer ontologies. (Величко В. Ю., Приходнюк В. В. Технологическое средство графического проектирования компьютерных онтологий.) In: Troitzsch K. G., Debicki R., Chernyshenko S. V., Romaniuk V.V., Kyrychenko K. I. (eds.) Conference Proceedings “Actual problems of training specialists in ICT”, Part 2; Sumy State University, Sumy 2013, pp. 38-43 (in Russian).

[Witte et al, 2010] René Witte, Ninus Khamis, Juergen Rilling, "Flexible Ontology Population from Text: The OwlExporter", International Conference on Language Resources and Evaluation (LREC), Valletta, Malta: ELRA, pp. 3845--3850, 2010 http://www.lrec-conf.org/proceedings/lrec2010/pdf/932_Paper.pdf (accessed: 31.07.2013)

Authors' Information

Ivanova Krassimira – University of National and World Economy, Sofia, Bulgaria; e-mail: krasy78@mail.bg

Major Fields of Scientific Research: Software Engineering, Business Informatics, Data Mining, Multidimensional multi-layer data structures in self-structured systems

Agent-Oriented Software Engineering Models

REDUCING SEMANTIC GAP IN DEVELOPMENT PROCESS OF MANAGEMENT INFORMATION SYSTEMS FOR VIRTUAL ORGANIZATIONS

Jacek Jakieła, Paweł Litwin, Marcin Olech

Abstract: *The paper describes experience gained by developing an agent-oriented methodology suitable for design and implementation process of Management Information Systems for modern business structures called virtual organizations. It starts with description of semantic gap problem, shows main concepts related to agency and virtual organizations and describes similarities between virtual organizations and multi-agent systems. Next we shortly present the state-of-the-art of currently used methodologies aimed at inter-organizational modeling and show motivations which have become the rationale for our approach. What is more we present the main methodology assumptions made, international standards the methodology follows and the framework we have developed so far. The paper ends with conclusions and further research plans.*

Keywords*: business modeling, agent-oriented development methodology, agent-oriented-software engineering, virtual organization, semantic gap, agent-oriented management information systems*

ACM Classification Keywords*: I. Computing Methodologies; I.2 Artificial Intelligence; I.2.11 Distributed Artificial Intelligence; Multi-Agent Systems*

Introduction

Nowadays business organizations are becoming increasingly complex systems. This complexity is well visible during the development process of information system supporting modern organizations' activities. Organizations are no more monolithic structures with clear boundaries and areas of operations, well defined according to functional hierarchies. Modern business structures can take different forms. Enterprise can consist of distributed independent organizations with physical presence, that share resources to achieve common goals [Franke 2002], or can be fully virtual and operate primarily via electronic means e.g. various forms of business webs [Tapscott et al. 2000]. Therefore the structural and behavioral characteristics of business firms have profoundly changed. All these changes have to be taken into consideration during the process of designing the management information systems.

Apart from the complexity of business domain there is also the problem of complexity of software, which is its essential property, not an accidental one. This inherent complexity derives from such elements as: the difficulty of managing the development process, the flexibility possible through software, and the problems of characterizing the behavior of discrete systems [Brooks 1995, Booch et al. 2007].

The third problem is that contemporary methodologies for software development are not equipped with modeling methods for preparing system specification including all new characteristics of business problem domains. As a result semantic gap arises between business models and software models used for implementation of management information system. This gap is the source of many problems during the process of transforming business specification into the software architecture. The main problem is that many aspects of organization operations that should be supported by system under development are not taken into account during development process and therefore software does not support properly business goals of the enterprise. As there will be shown in the paper, the solution of this problem is to use appropriately selected modeling concepts for all of the stages of software development. The concepts defined for every stage have to have high semantic proximity, and thanks to this, the process of transforming business model into software model is an intuitive and unambiguous mapping of specifications' artifacts.

The paper presents the skeleton of management information system development methodology that has two main goals. Firstly it improves the process of business and software complexity management. Secondly it helps to reduce the semantic gap between business and system specifications prepared during software development process. The proposed methodology is based on the concept of software agent and its characteristics which are used as modeling constructs. As will be shown, such approach enables to better manage the complexity of modeling process. What is more the semantic proximity of the agent and modern organizations' characteristics will lead to significant reduction of semantic gap between modeling artifacts used at different stages of development process.

The Semantic Gap Problem

The semantic gap characterizes the difference between two descriptions of an object by different linguistic representations. In computer science, the concept is relevant whenever human activities, observations, and tasks are transferred into a computational representation. More precisely the gap means the difference between ambiguous formulation of knowledge related to the application domain in a business specification and its computational representation in a formal language – at first system specification and then programming language.

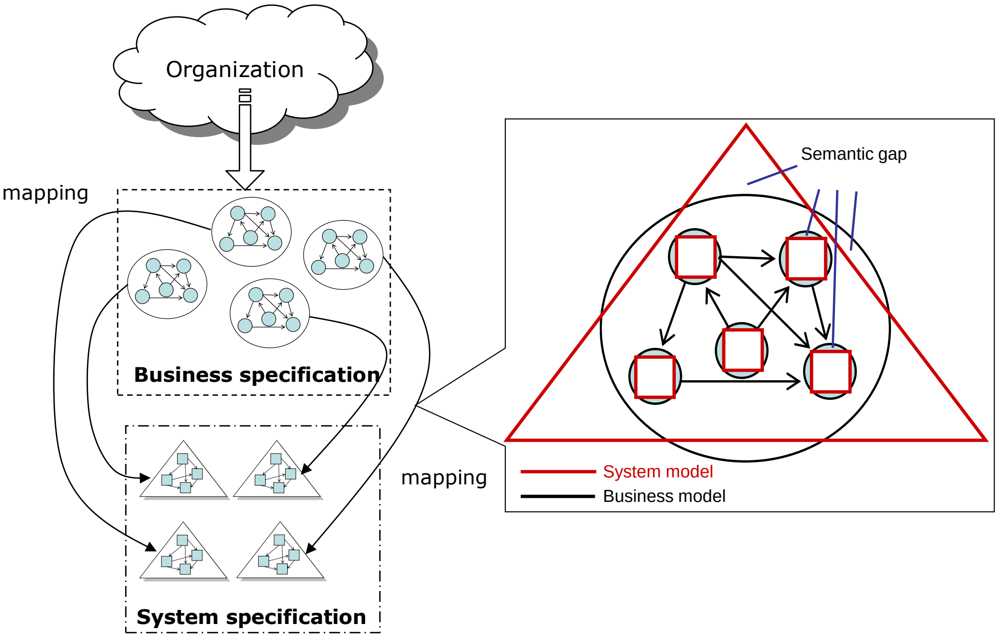


Fig. 1. Semantic Gap Problem

It is a fundamental task of software engineering to close the gap between application specific knowledge and technically doable formalization. For this purpose domain specific (high-level) knowledge must be transferred into algorithms and their parameters (low-level) [Dorai, Venkatesh 2003]. When developing management information system there are two main analysis and design perspectives: business system perspective and software system perspective. In order to describe application domain, analyst with stakeholders prepare business model that formalizes all important aspects of organization's operations and structure. It typically includes organization chart, strategy related documents (vision, mission and goals statements), business processes models and data models.

This constitutes the basis for application domain analysis and requirements elicitation as well as specification. In the next stage the system specification is derived from business specification. It includes the software architecture and functionalities description as well as all the models related to static and dynamic aspects of the system under development. The main question here is *how to express business artifacts in terms of implementation constructs?* At this point the semantic gap problem usually arises because every perspective uses its own set of concepts – different for business and software modeling, and therefore there is no intuitive and unambiguous mapping between business and software mindsets. The problem is visualized on the figure 1. In the following sections there will be shown how to tackle this problem. The cornerstone of proposed solution is to base development methodology on an agent concept.

The Essence of Agency

Before advantages of agent oriented organization modeling and system development will be presented, it seems advisable to explain the essence of agency and define two main concepts our methodology is based on – an *agent* and a *multi-agent system*.

The Concept of Agent

Over the last two decades the concept of an intelligent agent has become really popular. A number of researchers dealing with artificial intelligence domain focused on agency. Consequently numerous definitions of an agent have been coined. Two of them are provided below.

Michael Wooldridge and Nicholas R. Jennings [Wooldridge, Jennings 1995] describe an agent as: *a hardware or (more usually) software-based computer system that enjoys the following properties*:

- *autonomy*: agents operate without the direct intervention of humans or others, and have some kind of control over their actions and internal state;
- *social ability*: agents interact with other agents (and possibly humans) via some kind of agent-communication language, which enables to exchange their knowledge;
- *reactivity*: agents perceive their environment, (which may be the physical world, a user via a graphical user interface, a collection of other agents, the INTERNET, or perhaps all of these combined), and respond in a timely fashion to changes that occur in it;
- *pro-activeness*: agents do not simply act in response to their environment, they are able to exhibit goal-directed behavior by taking the initiative.

Another definition has been proposed by S. Franklin and A. Graesser in their paper attempting to distinguish software agents from regular computer programs [Franklin, Graesser 1996]: "*An autonomous agent is a system situated within and a part of an environment that senses that environment and acts on it over time, in pursuit of its own agenda and so as to effect what it senses in the future.*"

Based on definitions presented few vital attributes of an agent can be abstracted:

- an agent exists in a certain *environment* and thus it ceases to be an agent when extracted from such environment,
- an agent *senses its environment*, *acts on this environment* and its *actions can affect what an agent will sense in the future*,

- an agent *operates over time* and acts whenever it “feels” it’s necessary; unlike regular program which terminates once its mission is accomplished,
- an agent *operates autonomously* pursuing its own goals and *is able to undertake pro-active behavior*.

All these basic characteristics constitute conceptual framework that will be used later when trying to show how agent oriented methodology may help to improve complexity management and reduce semantic gap during management information system development process.

The Concept of Multi-Agent System

A Multi-Agent System may be defined as a set (society) of decentralized software components (where every component exhibits the properties of an agent, mentioned in the previous section), that are carrying out tasks collaboratively (often in parallel manner) in order to achieve a goal of the whole society. Later in the paper this definition has been disaggregated and all the properties are used to show why the multi-agent system can be considered as a very intuitive virtual organization modeling metaphor.

As presented definition reveals, software agents have the ability to collaborate with each other what enables the creation of multi-agent systems. Collaboration is defined as a process in which society coordinate its actions in order to achieve common goals. Software agents are able to collaborate with one another as well as human agents.

The corner–stones of inter–agent collaboration are: *communication and knowledge sharing*. *Communication* is basically an exchange of information among agents (agents can send messages to each other, observe each other’s state and behavior, however, communication takes place on the knowledge level). To enable knowledge sharing agents must have common goals and decompose the process of achieving these goals into sequence of actions providing that every agent is capable of performing task assigned to it.

Inter–agent collaboration requires also a *communication language*. Currently the most popular agent communication languages are: Knowledge Query and Manipulation Language (KQML) developed in early 90’s and FIPA-ACL developed by Foundation for Intelligent Physical Agents. Both rely on speech acts theory and define a set of performatives, their meaning and protocol for perfomatives exchange.

Although there are many frameworks and agent architectures developed so far by AI community, we have decided to base our methodology on *Belief-Desire-Intension* approach [Georgeff et al. 1999], which is most widely used framework for multi-agent systems development and offers implementation constructs that are semantically closest to virtual organization characteristics.

The Essence of Virtual Organization

Virtual organizations are (often temporary) value-added partnerships of independent, autonomous actors, such as individuals, companies or research institutes that have established a pre–partnership relationship in order to work together for achieving common goals. As we can see there is very close semantic proximity between basic characteristics of virtual organization and multi-agent system. More detailed insights will be presented in the next section.

Virtual organization can be viewed as a hub of partner firms that are selected according to an actual need in order to carry out a given task on a temporary basis. They can be partnerships of independent firms with physical presence. In such case every partner delegates specific unit of organization which constitutes the part of the virtual organization structure. It can also be the whole organization that takes part in such alliance with all departments and resources it possesses. Possible configuration of virtual organization is presented on figure 2.

Organizations may also take fully virtual forms. With the development of the Internet and modern ICT, new ways of conducting business have evolved. Many firms without physical presence are conducting business on electronic marketplaces. In such case virtual organization is a business web platform which provides the environmental condition, such as trust and coordination mechanisms and tools, necessary for the dynamic configuration of market and customer-driven value chain constellations [Franke 2002]. Therefore the task of designing organization boils down to developing the software (in the form of business web) that will support all operations of e-business partners related to extended supply chain (see figure 3).

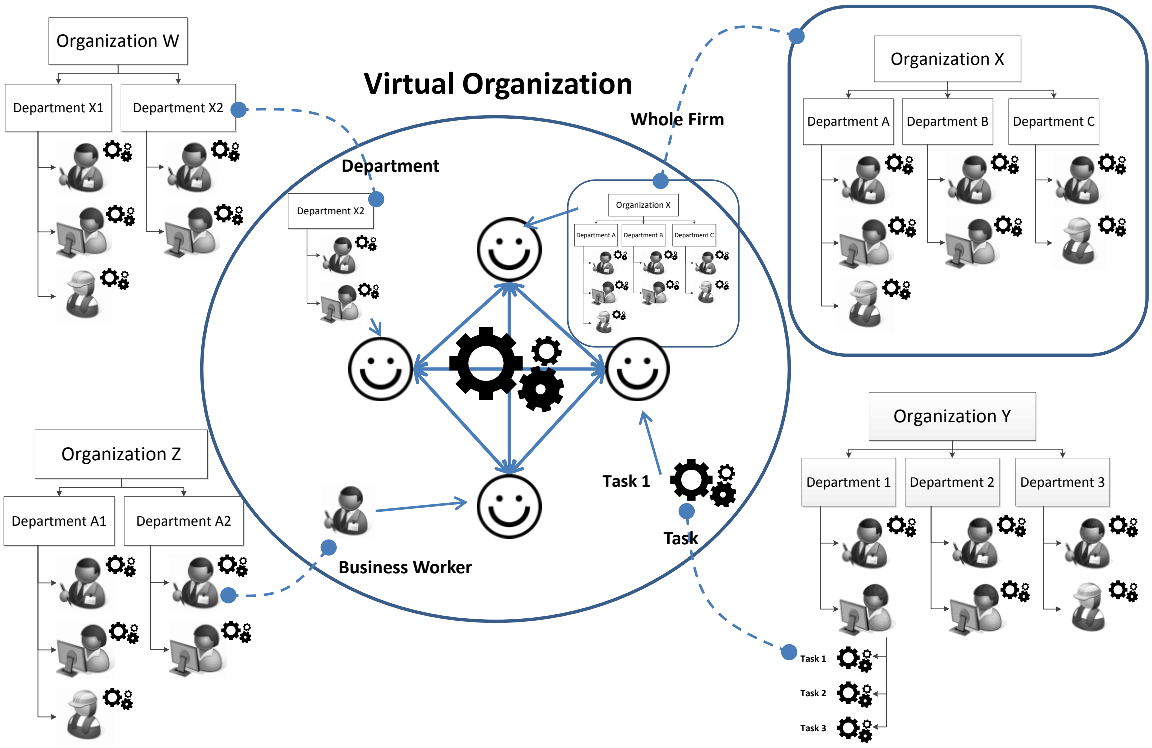


Fig. 2. The Structure of Virtual Organization

In both cases it is possible to abstract the common characteristics of virtual organizations. Most important are the following:

- process oriented organization structure rather than functional hierarchy,
- dynamic nature of organization structure that can change during its lifecycle,
- high autonomy of partners that constitute virtual organization,
- every business partner is self-contained what means that it has all needed competences and resources for conducting specific category of tasks,
- goal orientation, what means that all the tasks are organized toward achievement of common goals,
- physical distribution of partners.

As there will be argued later in the paper, in order to solve the semantic gap problem, all of the presented characteristics have to be taken into account when developing the software supporting operations of virtual organizations.

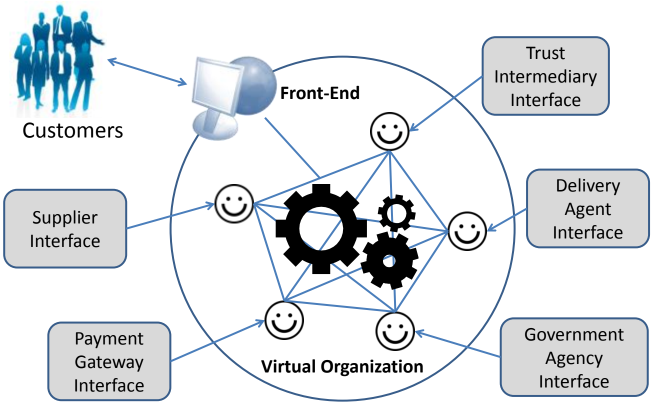

Customers
Front-End
Trust Intermediary Interface
Supplier Interface
Delivery Agent Interface
Payment Gateway Interface
Government Agency Interface
Virtual Organization

Figure 3. The Structure of Business Web Virtual Organization

Agent Paradigm as a Modeling Framework and Complexity Management Tool

When considering agent paradigm as a modeling framework it is essential to answer two fundamental questions. The first one is *why agent and multi-agent system characteristics make agents so natural and intuitive organizational modeling constructs*? The second is *why agent orientation is optimal choice for complexity management*?

One of the fundamental assumptions for our methodology is that business modeling process bases on multi-agent system metaphor. It leads to perceiving and understanding of virtual organization in the way typical for multi-agent system software engineering, but also takes under consideration business aspects including most important virtual organization characteristics.

The similarities between multi-agent systems and business organizations are particularly visible in case of virtual organizations. The table 1 presents observed similarities. In the first column there are basic structural and behavioral characteristics of multi-agent systems, and in the second characteristics of virtual organizations [Jakieła, Pomianek 2009].

Table 1. Resemblances between structural and behavioral characteristics of virtual organizations and multi-agent systems

Multi-agent system	Virtual organization
Multi-agent system is a set of **decentralized** software components.	**Virtual organization** is value-added partnership of **decentralized business actors**, such as individuals, companies or research institutes.
Multi-agent system is a set of **autonomous** software components.	Decentralization of partners that form virtual organization requires **autonomy delegation**. It has drastically changed the role of business actors, because “controlled positions” have been replaced by positions which give full competence. In case of **virtual organization** it is impossible to avoid situation when **partners**, who perform tasks related to common goals, **are fully autonomous entities**.
Multi-agent system is a set of **goal-oriented** software components.	Virtual organization is partnership of independent actors that **work together in order to achieve common goals**.

Multi-agent system is a set of **self-contained** software components that **may carry out tasks in parallel manner**.	Instead of artificial operations order, in virtual organization natural operation order is used. Business processes conducted by virtual organization are de-linearized. It allows for performance acceleration, because **tasks are performed in parallel by self-contained business partners**. Every partner carries out the tasks in the most effective and efficient manner because every virtual organization component possesses competences and resources needed for conducting specific category of activities.
The organization of multi-agent system **may change dynamically** depending on the current goals of society.	The structure of virtual organization **can change during its lifecycle**, depending on the current goals that have to be achieved.

Now let's move to the issue of complexity management. As Grady Booch says "*The task of the software development team is to engineer the illusion of simplicity*" [Booch et al. 2007], however as we have already mentioned, complexity management is a serious problem in case of modern organizations modeling and software development. In order to better explain how an agent paradigm, used as a mindset of our methodology, can improve complexity management it is useful to define the concept of complex system. Booch [Booch et al. 2007] relying on Simon's work [Simon, 1996] has defined the basic characteristics of complex systems:

- complexity frequently takes a form of hierarchy, where the system is composed of sub-systems connected with each other, which have their sub-systems, which in turn have their sub-systems and so on until the elementary level is reached. This hierarchy does not mean the superior-subordinate relation. Thanks to the fact, that complex systems are nearly decomposable we can fully understand them, describe or even perceive. Simon claims that it is highly probable that in reality only the systems that have a hierarchical structure can be understood [Simon 1996]. Looking at virtual organizations from this perspective, it is possible to distinguish such levels of hierarchy as organization actors level, business process level, singular organization level and specific configuration of few firms in a form of virtual organization (see figure 4),

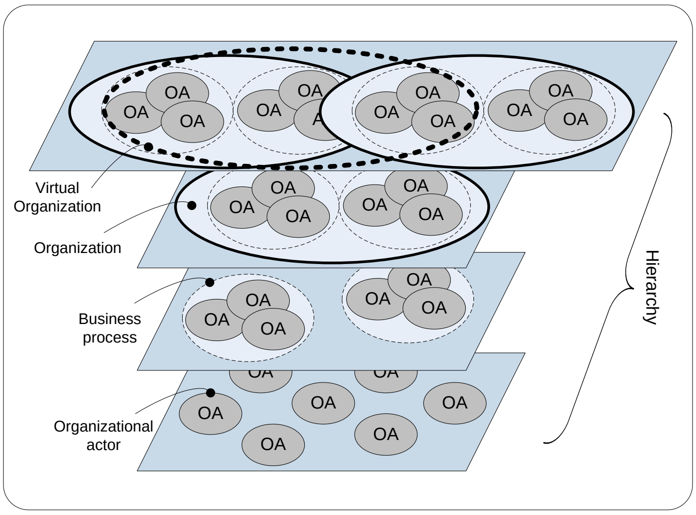


Fig. 4. Virtual Organization as a Complex System

- the choice which components of a system should be treated as elementary is arbitrary and depends on the system observer's decision,
- it is possible to identify interactions taking place between sub-systems as well as inside sub-systems, between their components, however interactions of the second type have one row higher frequency and are more predictable. The interaction frequency will differ depending on the level of hierarchy. For example, within a business organization more interactions will take place between employees working on the same process than between teams of employees working on different processes. The differences in interaction frequency within and between sub-systems allow decomposition and lead to the clear division between domains of analysis. In case of social systems, and undoubtedly every organization can be seen as such system, nearly decomposable character is clearly visible, therefore it is possible to exploit advantages of the decomposition method,
- complex systems are mostly sets of similar elements composed in various combinations. In other words there are certain common templates created on the basis of reuse of similar elementary components or more complex structures in the form of sub-systems,
- systems organized hierarchically tend to evolve over time, and hierarchical systems evolve faster than non-hierarchical ones. Simon claims that complex systems will evolve out of simple systems, if certain intermediary forms exist [Simon 1996].

Taking into consideration characteristics of complex systems as well as agent paradigm it is possible to show advantages of agent approach in the context of complexity management in the organization modeling as well as management information system development process [Jakieła 2006].

As the first argument it can be noticed, that agent oriented decomposition of a problem domain is an effective way to divide the problem space, while modeling organizations and information systems. It can be concluded from a number of factors. Firstly, hierarchical structure of complex systems causes, that modularization of organization components in terms of goals, that are to be achieved is a really intuitive solution. As Jennings and Wooldridge claim hierarchical organization of complex systems causes that at each level of the hierarchy, the purpose of the cooperation between sub-systems is achieving a functionally higher level. Whereas within sub-systems, components which these sub-systems are composed of, cooperate in order to achieve total functionality of a sub-system. As a consequence, decomposition oriented on goals that are to be reached is very natural division [Wooldridge, Jennings 2000]. Applying this schema to an organization the situation emerges where organization actors cooperate in order to achieve goals of the process, in turn processes are realized in order to achieve the goal of the specific firm, and firms combine their inherent competences in order to achieve the goals of virtual organization. It is worth to remember that goal orientation is one of the main characteristics of an agent and thus agent concept can be used without any additional effort.

Another vital issue is presentation of such characteristic of a modern organization as decentralization in the area of information processing and control. In this case agent oriented decomposition seems to be an optimal solution due to such characteristics of an agent as thread of control encapsulation in the form of autonomy property. The distributed organizational components may be thus modeled with autonomous agents as a basic modeling constructs.

Agent oriented approach allows also to solve problems connected with the design of interactions taking place between system components. It is a serious issue due to the dynamics of interactions between organization components. It is really frequent, that organization components enter an interaction in time which is difficult to predict and for unknown at the design stage reasons. As a consequence it's really challenging to predetermine parameters of such interactions. The solution to this problem is existence of system components that can make decisions concerning the type and range of interaction during runtime.

Another argument for an agent oriented approach is that it allows to eliminate semantic gap between agent abstraction used during the information system design phase and structures used during organization modeling. It is directly connected with similarities which appear between structural and behavioral characteristics of a multi-agent system and organization (table 1). Continuing this thread it is advisable to point out the following conveniences:

- Mutual interdependencies among organization actors and organization sub-systems can be naturally mapped into the system architecture in terms of high-level social interactions which take place among agents.
- In virtual organizations dependencies of this type are present in the form of really complex network of dynamically changing relationships. Agent based approach includes mechanisms which allow to describe such relations. For example, interaction protocols such as Contact Net Protocol can be used in order to dynamically create virtual organization structure, which can be, in case of such need, activated and after reaching particular goals deactivated. What is more, there are off-the-shelf structures, which can be used during the community modeling, what is really useful when modeling organization actors and sub-systems [Wooldridge, Jennings 2000].
- The process of organization modeling and system design frequently requires to perceive modeled object from the perspective of various abstraction levels, treating set of elements as atomic modeling structure. The idea of an agent is flexible enough to be used in an elementary level or on any detail level depending on the analyst's needs. For example, an agent could be organization actor, department or whole organization and components treated as elementary interact only in an integrated form omitting details concerning intra-interactions.
- Organization modeling and system design with agent oriented approach leads to the structure, which has numerous stable intermediary forms, what is really important concerning complexity management. Among others it means that system components in the form of agents can be created rather independently and, in case of such a need, added to the system providing a smooth functionality growth.

After introduction to basic advantages of taken approach, the next section describes the methodology for management information system development, based on all the insights that have been presented so far.

The Skeleton of Development Methodology

Related Works and Methodology Motivations and Assumptions

Motivations for our work have been derived from detailed analysis of research concerning development methodologies for inter-organizational and virtual organizations management information systems [Huemer et al. 2008, Yu 1995, Mylopoulos et al. 2002, Zaborowski 2006, Mili et al. 2010, Telang et al. 2012]. The analysis discovered opportunities for improvements.

The first problem identified is the lack of guidelines for business modeling stage in the system lifecycle and in some methodologies this stage has not been taken into consideration at all. As best practices, prepared by Object Management Group show, business specification including all organization stakeholders' needs and business goals, is a key determinant of quality of management information system under development.

Next important issue is the ease of methodology adoption to practical applications. We assumed that the key factor of fast methodology adoption is to make use of unified languages that are considered as international standards for business and software systems modeling. Unfortunately part of methodologies in use (e.g. COMMA, TROPOS) is based on non-standard, original notations what considerably decreases the speed of adoption for business and industrial applications and narrows the circle of prospect users.

Analysis of research works enabled to formulate the following motivations for our methodology:

- The designing of information systems for virtual organizations requires the detailed business model that describes structural and behavioral characteristics of application domain for which system is being developed.
- The business model has to be precisely mapped into architecture and functionalities of the software system that will support virtual organization operations.
- The system development should be supported by the process, analysis and design methods for business modeling and system implementation as well as unified modeling language adopted by software industry.

The motivations presented have been used for preparing the following methodology assumptions:

- Development methodology should be equipped with detailed business modeling stage that includes such aspects as business motivation model, business processes model, business rules model and organizational structure model. This enables to include in business specification the most important characteristics of application domain and considerably improves the process of system requirements elicitation and specification.
- All modeling methods developed as a part of methodology should be based on unified languages, what will increase the speed of its adoption to business and industrial applications.
- Methodology should be agent oriented what will enable to reduce the semantic gap between business and system requirements and facilitate the management of business and software modeling complexity. Agent orientation of the methodology means that all modeling concepts used are derived from agent paradigm and related to virtual organizations characteristics. What is more, during implementation stage management information system should be developed as a multi-agent software solution based on *Belief-Desire-Intension* (BDI) *architecture*.

Modeling Standards, Frameworks and Implementation Architectures Used

In order to increase the speed of methodology adoption it has been based on international modeling standards. All the standards were used as the meta-models that have been extended and adjusted for our methodology modeling methods. The following standards have been used:

- *Business Motivation Model* (BMM) – a standard developed by Object Management group which allows a business plan to be developed, communicated and managed in an organized manner. Business strategy is modeled in terms of Vision, Goals, Objectives, Mission, Strategies and Tactics, and internal as well as external Influences. These influences are then assessed to identify the potential impact they may have on the business. What is important all elements of the BMM are developed from a business perspective. The main idea is to create a business model for the elements of the business plan, before system design or technical development is begun. Thanks to this, the business strategy can become the foundation for system requirements specification and connects system solutions to their business intent [OMG 2013a].
- *Business Process Modeling Notation* (BPMN) – it is graphical notation that depicts the steps in business processes. BPMN depicts the end to end flow of a business process. The notation has been specifically designed to coordinate the sequence of processes and the messages that flow between different process participants in a related set of activities. BPMN is targeted at a high level for business users and at a lower level for process implementers. The business users should be able to easily read and understand a BPMN business process diagram. The process implementer should be able to adorn a business process diagram with further detail in order to represent the process in a physical implementation [OMG 2013b].
- *Unified Modeling Language* (UML) – this is an industry standard modeling language with a rich graphical notation, and comprehensive set of diagrams and elements. We have based all modeling

methods on UML notation. As we have already mentioned the rationale for this was to increase the speed of methodology adoption to business and industrial applications [OMG 2011].

- *Eriksson and Penker Business Patterns* – this is a set of UML extensions that enables to conduct business modeling with the use of UML notation. Part of these patterns is used as a foundation for organization layer in business modeling stage of our methodology [Eriksson, Penker 2000].
- *Belief-Desire-Intension Architecture* (BDI) – is one of the major approaches to building agents and multi-agent systems, including commercial agent software. It is inspired by logics and psychology. The main idea is to build agents using symbolic representations of agents' beliefs, desires, and intentions. It provides a mechanism for separating the activity of selecting a plan (from a plan library) from the execution of currently active plans. Consequently, BDI agents are able to balance the time spent on deliberating about plans (choosing what to do) and executing those plans (doing it) [Borodini et al. 2007, Georgeff et al. 1999]. We used this architecture as a foundation for implementation discipline. According to methodology assumptions presented before, the management information system is developed as a multi-agent software solution which conforms to BDI architecture.

Structure of Methodology

Our methodology has multi-layer organization. Its structure may be presented in two dimensions – *static* and *dynamic*. Static structure describes how process elements are logically grouped into core process disciplines. Basic process elements are: modeling methods, disciplines, artifacts, and roles. Dynamic structure shows how the process, expressed in terms of cycles, phases, iterations, and milestones, unfolds over the lifecycle of a project (figure 5). We borrowed dynamic structure from *Rational Unified Process* (RUP), which defines four main phases: inception, elaboration, construction and transition. In *inception phase* a good understanding of what system to build is gotten. It is done by getting a high-level understanding of all the requirements and defining the system's scope. In this stage the focus is also on mitigating business risks, and producing the business case for building the system. Finally it is important to get acceptance of all stakeholders and decide whether to proceed with the project. During *elaboration phase* most technically difficult tasks such as: design, implementation, testing, and baselining an executable architecture (including subsystems, their interfaces, key components, and architectural mechanisms) are undertaken. What is more, major technical risks are addressed by code implementation and validation [Barnes 2007].

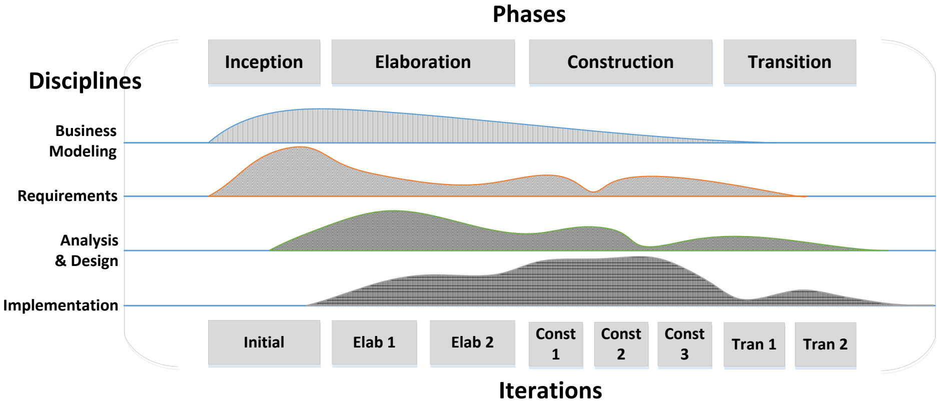


Fig. 5. The Agent Oriented Development Methodology Dynamic Structure

Most of the implementation is done during *construction phase*. Programmers are developing first operational version of the system on the basis of executable architecture. Then they deploy alpha releases to verify if system under development meets stakeholders' needs. At the end of this stage fully functional beta version is deployed, however system still requires improvements and tuning related to overall functional and non-functional requirements as well as quality.

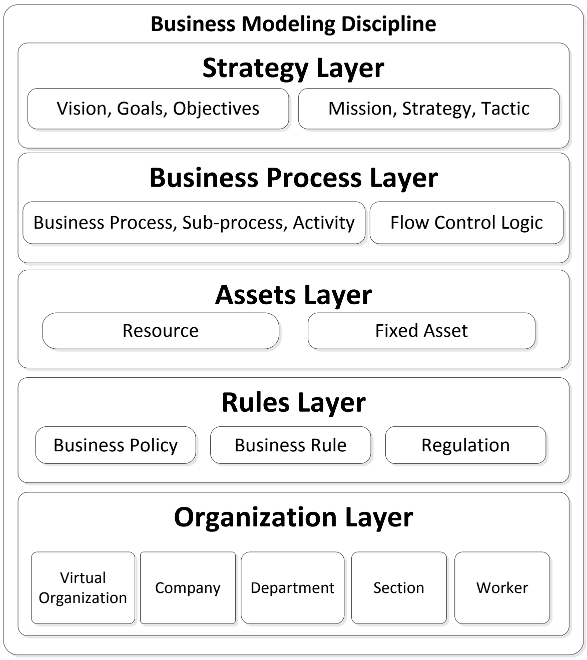


Fig. 6. Structure of Business Modeling Discipline

Main aim of the *transition stage* is to collect final feedback and ensure the release of the system under development addresses needs of all stakeholders. During this stage testing and minor adjustments are made. Basic activities include fine-tuning of the product, configuration and usability analysis. Focus is also on users training and integration issues.

Our original contribution is related to static structure of the methodology. We have extended business modeling, requirements as well as analysis and design disciplines. What is more, according to methodology assumptions, we have made implementation discipline agent oriented and prepare the transformations of design artifacts to implementation constructs. The figure 6 presents our approach in the area of the business modeling discipline.

Every layer is responsible for modeling specific aspect of the virtual organization. They are as follows:

- *Strategy layer* is split into two sections: Ends and Means. Ends section is used for describing state the virtual organization wants to achieve. This section includes such elements as: Visions, Goals and Objectives. Means section describes what courses of action need to be taken and how should be executed. In this section there are store elements like Missions and Courses of Action expressed in terms of Strategies and Tactics. Modeling concepts in the strategy layer have been adapted from Business Motivation Model. Elements of the strategy layer are visually modeled with the use of UML extensions developed for every modeling concept.
- *Business Process layer* is meant to model business processes, sub-processes, activities and finally the logic of control flows. Elements of the business process layer are visually modeled with Business Process Modeling Notation (BPMN) constructs.
- *Assets layer* is responsible for modeling all the resources and fixed assets which are used by virtual organization business processes. Concepts in the assets layer are based on BMM and are visually modeled with UML extensions.

- *Business Rules layer* is used for modeling directives which govern or guide business processes. There are three categories of directives: *Business Policy, Business Rules and Regulations*. Concepts in the business rules layer are based on BMM and its visual notation is developed in the form of UML extensions.
- *Organization layer* is meant to model the structure of the virtual organization. It includes such elements as: workers, sections, departments and companies that form virtual organization. Concepts in this layer are based on Eriksson & Penker patterns visually modeled with extended UML language.

All layers are interconnected and therefore it is possible to trace any artefact between layers. For every layer modeling methods have been developed according to framework published in [Mayer et al. 1995]. Every modeling method is described in the structure consisted of name, method definition (concepts and motivation), discipline (dictionary, grammar, detailed procedure) and use (how the method is used in the system development process).

Integral part of every layer is a visual modeling language. The notation of the language has been based on UML and concepts derived from standards presented before. The visual language was developed for every layer for both business modeling and analysis and design disciplines with the use of meta-modeling approach provided by OMG [OMG 2013c]. Sample meta-model of modeling language created for *strategy layer* and *means section* is presented on figure 7.

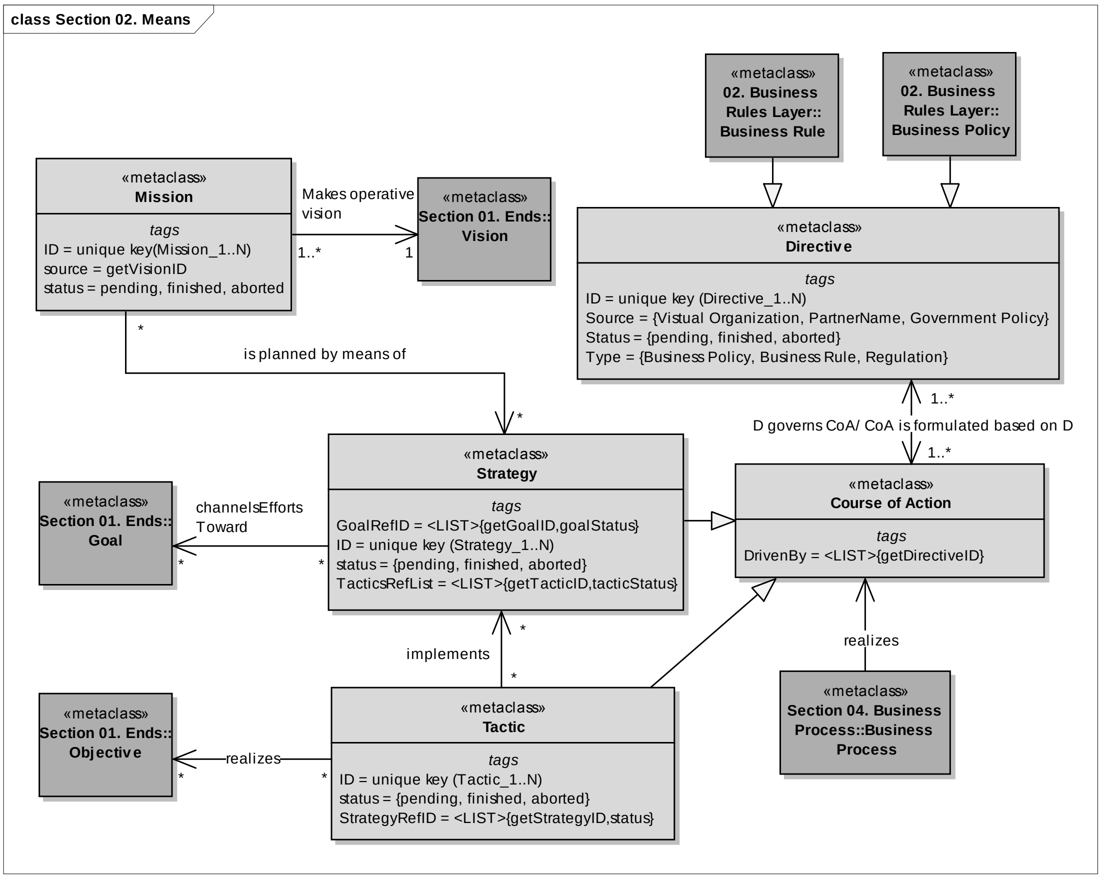

class Section 02. Means
«metaclass»
02. Business Rules Layer:: Business Rule
«metaclass»
02. Business Rules Layer:: Business Policy
«metaclass»
Mission
tags
ID = unique key(Mission_1..N)
source = getVisionID
status = pending, finished, aborted
Makes operative vision
1..*
1
«metaclass»
Section 01. Ends:: Vision
«metaclass»
Directive
tags
ID = unique key (Directive_1..N)
Source = {Vistual Organization, PartnerName, Government Policy}
Status = {pending, finished, aborted}
Type = {Business Policy, Business Rule, Regulation}
*
is planned by means of
*
1..*
D governs CoA/ CoA is formulated based on D
1..*
«metaclass»
Strategy
tags
GoalRefID = <LIST>{getGoalID,goalStatus}
ID = unique key (Strategy_1..N)
status = {pending, finished, aborted}
TacticsRefList = <LIST>{getTacticID,tacticStatus}
«metaclass»
Course of Action
tags
DrivenBy = <LIST>{getDirectiveID}
«metaclass»
Section 01. Ends:: Goal
channelsEfforts Toward
*
*
implements
*
*
realizes
«metaclass»
Section 04. Business Process::Business Process
«metaclass»
Section 01. Ends:: Objective
realizes
*
*
«metaclass»
Tactic
tags
ID = unique key (Tactic_1..N)
status = {pending, finished, aborted}
StrategyRefID = <LIST>{getStrategyID,status}

Fig. 7. Part of Visual Language Meta-model for Strategic Layer of Business Modeling Discipline

Elements filled with darker color comes from different sections/layers. Elements with brighter background are native to the current layer and section. As can be seen, basic element in the Means section is a *Mission*, which indicates the ongoing operational activity of the enterprise. Mission makes *Vision* operative. Every

Mission is planned by means of *Strategy*, which represent the essential *Course of Action* to achieve, *Ends — Goals* in particular. Every Strategy is implemented by *Tactic*, which is a Course of Action that represents part of the Strategy details. Tactics generally channel efforts towards *Objectives*. Every Strategy and Tactic is Course of Action which is formulated based on *Directive*. Directives indicate how the Courses of Action should, or should not be carried out — in other words, they govern Courses of Action. Specifically, a Directive defines constraints or liberates some aspect of an enterprise.

After all activities related to business modeling discipline planned for specific iteration are done, Analysis and Design discipline activities are carried out. The structure of Analysis and Design discipline is presented on figure 8.

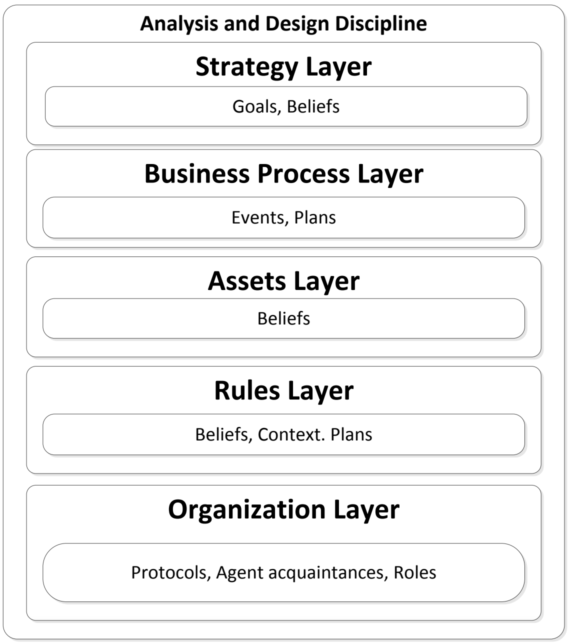


Fig. 8. The Structure of Analysis and Design Discipline

As has already been mentioned, analysis and design as well as implementation disciplines have been developed according to an agent paradigm. Therefore all the modeling concepts that are used for preparing agent oriented system specification conform to BDI architecture. What is more all the concepts are related to business modeling discipline concepts in such a way that there is a direct and unambiguous mapping between business and systems specifications. Finally implementation discipline uses the programming constructs from AgentSpeak language interpreted by Jason [Borodini et al. 2007].

Agent oriented modeling concepts are the following:

- *Beliefs*, which are used to represent information agent stores about environment, other agents and itself. Interesting fact about beliefs in Jason is that they are annotated and therefore may be maintained on the meta-level. There are three main annotations such as: percept, self and agent name. Percept is used to denote information from the agent sensors (received from environment). Self means that the belief is created by agent as mental note. Agent name suggests that source of the belief is other agent.
- *Goals* represent the state of affairs the agent strives for. The representation of goal is the same of a belief except that it is prefixed by exclamation mark.

- *Plans* constitute courses of action an agent will execute in order to achieve its goals or to react to changes in its environment. Every agent has the library of plans that determine its behavior. The plan is structured as presented below.

```
triggering_event : context <- body.
```

The *triggering event* represents event that will be handled by plan. *Context* describes circumstances under which the plan is suitable to handle the event. The *body* is a sequence of actions that will be executed or new goals for the agent to achieve. The agent behavior may change over time if new plans are acquired during the communication with other agents. Events result from changes in beliefs and goals. Beliefs may be updated and new goals set or received from other agents as a part of the delegation process. Events trigger execution of plans, provided that event matches the triggering event and is applicable in the time it is selected.

- *Protocols* specify rules of how all messages will be exchanged between agents. Agent acquaintances describe how agents are related to one another.
- *Roles* describe what agent is responsible for.

According to methodology assumptions all elements modeled in business modeling discipline have to be mapped into agent paradigm concepts in analysis and design discipline and finally implemented as management information system solution. What is also very important this conversion should be done in the way that reduces semantic gap between business and system aspects.

Figure 9 presents the mapping between modeling concepts. Elements from Organization Layer are represented using agents' roles, agents' acquaintances and composed protocols. Different resources modeled in Assets Layer are mainly described using agents' belief base. Concepts from Strategy Layer and Ends section are mapped into the agents' goals and beliefs. Concepts from Means section are mapped into agents' goals and plans. Business Policies, Business Rules and Regulations are used in plans mainly to check their context. Business processes are converted to events or messages and plans.

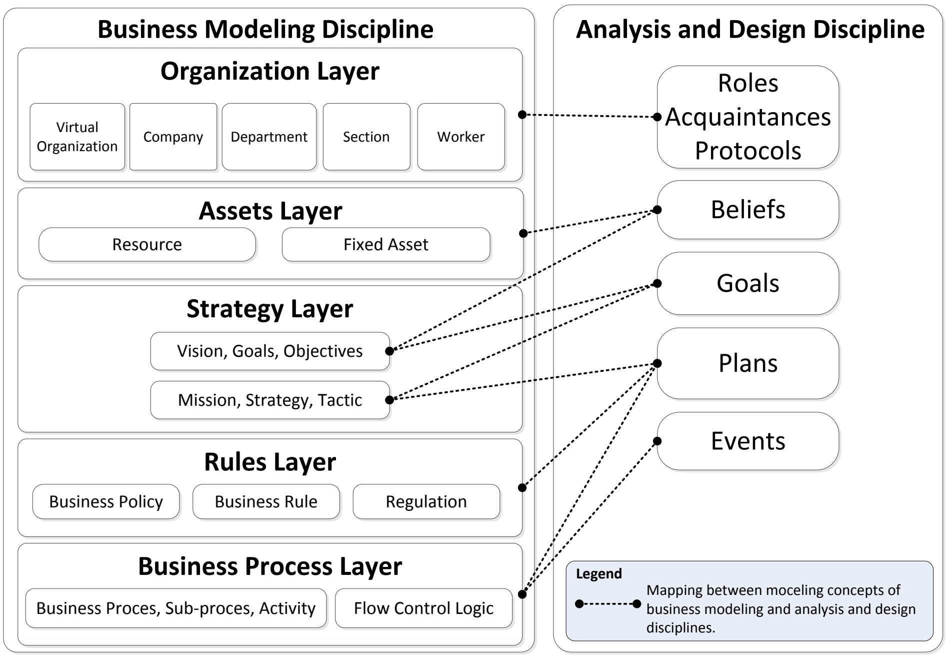


Fig. 9. Mappings between Modeling Concepts

Final development activities are related to implementation of the design model. This is done with *AgentSpeak* language and Jason interpreter. Because in Analysis and Design discipline all of the constructs come from BDI approach, the implementation boils down to the conversion of design model to agent oriented implementation constructs. However this is behind the scope of the paper as its aim was to present of how an agent paradigm may be used as tool for reducing semantic gap and improving development complexity management.

Conclusions and Further Research

High complexity and newly emerged characteristics of contemporary business structures require new approach to management information systems development. This approach should address to main problems: enable to reduce sematic gap between business model of application domain and design model of the system under development as well as improve the process of complexity management. Paper presents the solution in the form of agent oriented development methodology for systems supporting virtual organizations' operations. As was shown in the paper, agent orientation is an effective way to capture and include in business model the structure and behavior of modern enterprises. It is possible because of very high semantic proximity of agent paradigm modeling constructs and contemporary organizations characteristics. Our methodology is generic what means that it can be used in system development for organizations of any specificity and industry sector. Thanks to unified modeling language and process used in our methodology, it can be fast adopted to business and engineering applications. What is more, because of very detailed business modeling discipline it is possible to model all the important aspects of every organization and finally elicit the system features that will fully support virtual enterprise's business goals and objectives.

The methodology is still under development. The future research will concern the detailed meta-model for implementation discipline as well as automatic Model-Driven-Architecture transformations between specifications artifacts. However the methodology in its current state of development may be used in management information system prototype projects. This should be done in order to verify our approach and improve its meta-models of visual language and the process (disciplines).

Bibliography

[Barnes 2007] Barnes J.: Implementing the IBM Rational Unified Process and Solutions: A Guide to Improving Your Software Development Capability and Maturity. IBM Press, 2007.

[Booch et al. 2007] Booch G., Maksimchuk R., A., Engel M., W., Young B., J.: Object-Oriented Analysis and Design with Applications. Addison-Wesley, 2007.

[Borodini et al. 2007] Borodini R. H., Hubner J. F., Wooldridge M.: Programming Multi-Agent Systems in AgentSpeak Using Jason. Wiley, Chichester, 2007.

[Brooks 1995] Brooks F., P.: The Mythical Man-Month: Essays on Software Engineering. Addison-Wesley, 1995.

[Dorai, Venkatesh 2003] Dorai, C., Venkatesh, S.: Bridging the Semantic Gap with Computational Media Aesthetics. IEEE MultiMedia, 2003.

[Eriksson, Penker 2000] Eriksson H. E., Penker M.: Business modeling with UML: Business patterns at work. John Wiley & Sons, 2000.

[Franke 2002] Franke U., J.: Managing Virtual Web Organizations in the 21st Century: Issues and Challenges. Idea Group Publishing, 2002.

[Franklin, Greasser 1996] Franklin, S., Greasser, A.: Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents. University of Memphis, 1996.

[Georgeff et al. 1999] Georgeff, M., Pell B., Pollack M., Tambe M., Wooldridge M.: The Belief-Desire-Intention Model of Agency. Intelligent Agents V: Agents Theories, Architectures, and Languages. Spronger-Verlag, 1999.

[Huemer et al. 2008] Huemer Ch. et al.: The Development Process of the UN/CEFACT Modeling Methodology. 10th International Conference on Electronic Commerce (ICEC) '08 Innsbruck, Austria.

[Jakieła 2006] Jakieła, J.: AROMA – Agentowo zoRientowana metOdologia Modelowania orgAnizacji. WAEiI, Politechnika Śląska, Gliwice, 2006

[Jakieła, Pomianek 2009] Jakieła 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.

[Jennings, Wooldridge 2000] Jennings N., R., Wooldridge M.: Agent-oriented software engineering. Proceedings of the 9th European Workshop on Modeling Autonomous Agents in a Multi-Agent World : Multi-Agent System Engineering, 2000.

[Mayer et al. 1995] Mayer, R., J., Crump, J., W., Fernandes, R., Painter, M., K., Keen A.: Information Integration for Concurrent Engineering. Compendium of Methods Report. Interim Technical Paper. Wright-Patterson Air Force Base. Ohio, 1995.

[Mili et al. 2010] Mili H. et al.: Business Process Modeling Languages: Sorting Through the Alphabet Soup, ACM Computing Surveys, Vol. 43, No. 1, Article 4, 2010.

[Mylopoulos et al. 2002] Mylopoulos J., Castro J., Kolp M.: Towards requirements-driven information systems engineering: the Tropos project, Information Systems 27 (2002) 365–389. Elsevier, 2002.

[OMG 2011] Object Management Group: Unified Modeling Language, ver. 2.4.1, August 2011.

[OMG 2013a] Object Management Group: Business Motivation Model, ver. 1.2b2, August 2013.

[OMG 2013b] Object Management Group: Business Process Model and Notation, ver. 2.0.2, December 2013.

[OMG 2013c] Object Management Group: Meta-Object Facility, ver. 2.4.1, June 2013.

[Simon 1996] Simon, H.: The Sciences of Artificial. MIT Press, 1996.

[Tapscott et al. 2000] Tapscott, D., Lowy, A., Ticoll, D.: Digital Capital: Harnessing the Power of Business Webs. Harvard Business Review Press, 2000.

[Telang et al. 2012] Telang R. P., Singh P. M.: Comma: A Commitment-Based Business Modeling Methodology and its Empirical Evaluation, Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012), Conitzer, Winikoff, Padgham, and van der Hoek (eds.), 4-8 June 2012, Valencia, Spain.

[Wooldridge, Jennings 1995] Wooldridge, M., Jennings, N., R.: Agent Theories, Architectures, and Languages: A Survey. Springer-Verlag, 1995.

[Yu 1995] Yu E.: Modelling Strategic Relationships for Process Reengineering, Ph.D. Thesis, University of Toronto, Department of Computer Science, Toronto, 1995

[Zaborowski 2006] Zaborowski M. „Model informacyjno-decyzyjny struktury danych o obiektach zarządzania", W: Kozielski St. i inni (red.) „Bazy danych. Modele, technologie, narzędzia. Analiza danych i wybrane zastosowania", WKŁ, 2006.

Authors' Information

Jacek Jakieła, Ph.D., Eng. – Department of Computer Science FMEA RUT; W. Pola 2, 35-959 Rzeszow, Poland; e-mail: 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

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

Major Fields of Scientific Research: Multiagent Based Simulations, Application of Artificial Intelligence in Industry.

Computational Simulation Models

TOOLS FOR ANALYSIS OF PROCESSES MEASURED ON SPARSE AND IRREGULAR SPATIAL-TEMPORAL GRID WITH APPLICATION TO DATA OF NATIONAL CENSUSES[3]

Alexander Temruk, Mikhail Alexandrov

Abstract: *In the paper we shortly describe mathematical and program tools for analysis spatial- temporal processes. Speaking mathematical tools we mean filling data on thick and regular grid in space and time and also use of principal component method for process decomposition. Speaking program tools we mean software for end-user, which implements all mentioned operations and provides dynamic visualization for results of modeling. The proposed technology proves to be useful for processing data of National censuses and we demonstrate it on the real example reflecting dynamic of population density in one of the Russian regions. Such a technology was described almost 15 years ago by P. Makagonov, who applied it for analysis of migration flows in the State of Oaxaca in Mexico. In our work we use this experience, but our system is developed on the new platform and it contains new algorithms.*

ACM Classification Keywords: *I.2 Artificial Intelligence*

Keywords: *principal component analysis, density of population, National census, software*

Introduction

1.1 Problem settings

Subject of this article are processes unfolding in time and space. We use the following limitations:

- Process is described by a single parameter. The case of several interrelated parameters is not considered.
- Position of measurement points in time does not change. So we have a grid, but not a chaotic set of points in space and time
- Space is an area of the plane. Thus, we deal with a function defined on the grid 2Dx1D

These limits are illustrated in Fig. 1. It shows the points on the plane, in which the values of function were measured. The numbers indicate the year. These timestamps refer to the years when the censuses were conducted in the USSR and in Russia.

When the grid of observation is dense and regular, one can implement analysis without filling. When this grid is sparse and irregular, we need a stage of filling data. In this paper we consider this case as the general one. To fill the data we offer our own algorithms taking into account the property of smoothness for the function under consideration.

[3] Work done under partial support of the British Petroleum grant (BP-RPANEPA 2013).

We deal with a sparse and irregular grid of observations. Such cases occur when the measurements are either very expensive or they are impossible for several reasons. It may be, for example, analysis of dynamics of parameters reflecting ecology, demography, and economics of large areas. In this paper we consider the dynamics of the population density in a region of Russia for demonstration of the proposed technology.

The term “analysis” involves the two following operations:

- recovery and visualization of spatial dynamics of a parameter under consideration
- decomposition of spatial-temporal processes and their visualization

One should note that both of these operations are possible only when data are given on a regular grid. Naturally that the presence of dense grid in space and time increases the quality of the analysis.

Visualization is to provide a film which can be displayed to an expert. Here each frame reflects distribution of the parameter at plane at successive moments of time. The main problem here is the presentation, which should emphasize changes of the parameter in time. Convenient form of presentation can stimulate intuition of an expert to explain the reasons of the observed dynamics. In this paper, we use our own algorithms to preprocess data before using standard procedures of visualization. Direct utilization of these procedures is ineffective.

To identify processes of different orders, we use principal component analysis (PCA). Despite of its simplicity principal components are rarely used in the analysis of spatial-temporal data. For this reason PCA is not a part of well-known software packages for processing such a data. So, in this paper we briefly describe the operations that allow to calculate these principal components. Dynamics of the first and second principal components (PCs) is usually sufficient to explain the spatial dynamics of the parameter under consideration. The developed system allows to implement its visualization using the method of preprocessing that we use in the film.

The final result of our applied research is the software system in the form of application oriented to ordinary end-user. The software system is developed in Matlab and includes algorithms, which have been described above. It is: filling spatiotemporal data, calculating PCs, dynamics visualization. To demonstrate the developed software system we use two examples. The artificial example shows the result of decomposition on PCs. The real example shows all steps of data processing: filling in time, filling in space, the decomposition on PCs. In this example, we process the data of censuses related to one of the Russian regions

1.2 Related works

Analysis of spatiotemporal data is used in climatology [Torne, 2007], meteorology [Kunitsyn, 2008], physics [Galanin, 2007] and many other applications. Typically, such an analysis only display data without revealing the factors of hidden dynamics.

There are several known software packages for the analysis of spatial data. The most advanced of them are is represented in a the group of software products ArcGIS. [ArcGIS, http: https://www.arcgis.com]. All of them have developed tools for maps presentation in different scales, interactive design of maps, matching maps. However, these packages can’t satisfy our needs for the following reasons: algorithms of filling spatial data are rough enough (linear interpolation and triangulation), there is no filling of data in time; there is no analysis of latent factors in dynamics. We do not mention here the time-series analysis. This analysis is available in many packages, but it is not accompanied by spatial analysis.

Our research use results of the work [Makagonov, 2003]. In this work the authors describe the system they developped to study migration flows in Oaxaca, one of the Mexican states. The data they used were data of the National censuses. The difference between Makagonov´s system and our system is: we use our own algorithms for filling data and visualization of results. Besides we use standard MatLab platform for system development and our codes are the open ones.

The paper is structured by the following way. In the section 2 we present algorithms. In the section 3 we describe developed software and results of experiments. Section 4 includes conclusions.

Algorithms

Filling in time, local splines

The function is assumed to be smooth and given on irregular grid in time If we consider the problem of interpolation under these conditions then we can use the cubic spline interpolation [Bahvalov, 1999]. Spline interpolation procedure is available in the package MatLab and it is named *spline()*. To find parameters of cubic interpolation the following conditions are used: (a) equality of polynomials function value $S_k(t_{k-1})= y_{k-1}$; $S_k(t_k)=y_k$; (b) continuity of the first derivative of splines $S'_{k-1}(t_k)=S'_k(t_k)$; (c) continuity of the second derivative of splines $S''_{k-1}(t_k)=S''_k(t_k)$. These conditions lead to a linear system of three diagonal type, which is solved by sweep method. We call this spline as a global spline.

In our system we use the so-called local cubic splines. It does not require equality of the second derivatives at the points of measurement. To calculate these splines the following conditions are used: (a) for all segments $S_k(t_{k-1})=y_{k-1}$, $S_k(t_k)=y_k$; (b) for the first segment $S''_1(t_0)=0$, $S''_1(t_1)=0$; (c) for the second and subsequent segments $S'_k(t_{k-1})= S'_{k-1}(t_{k-1})$, $S''_k(t_k)=0$. In the last condition $S'_{k-1}(t_{k-1})$ is known from the previous calculation of the spline on the previous segment.

The described algorithm is a modification of the algorithm described in [Kostyuk, 1977].

As a result of local spline interpolation we get regular and dense grid in time. It is illustrated on Fig. 2.

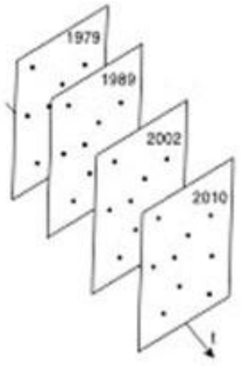


Fig. 1. Original data for analysis

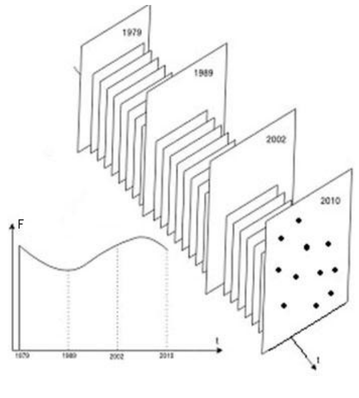


Fig. 2. Filling data in time

Filling in space, method of shades

At each temporary layer we have a grid of irregular points. To fill it to a regular and dense grid the well-known triangulation method is usually offered [Kalitkin, 1978]. This method is simple, but it does not take into account the smoothness of the function. In this paper we use the method of shadows. This method was proposed for filling two-dimensional distribution of geological parameters on a surface. The algorithm of this method is described in [Hakimov, 1986].

1. Function value in a point is determined by linear combination of function values at the points with known values. This contribution decreases with the distance between points and it depends on a characteristic radius *R*.
2. Only *n*-nearest neighbors affect function value in a given point.
3. Each point near a given point creates a shadow, which reduces the influence of points located behind. The shadow is defined by angle *θ*

Value function in some *i*-th point on the plane is given by the formula:

$$\varphi_i = F_0 + \Sigma_j w_j (\Delta F_j) /(1+\alpha r_j), \quad \Sigma_j w_j = 1, \quad j=1,\dots n$$

Here: φ_i is function value in the *i*-th point, F_0 is the mean value of functions in the known points, ΔF_j is deviation from the mean value in the *j*-th point (where the value of the function is known), w_j is the weight of this point, r_j is the distance from the *j*-th point to a given *i*-th point, *n* is the number of points with the known values of function near *i*-th point, $\alpha=1/R$ is the proportionality factor. Weight w_j is calculated by means of simple formulas that take into account the distances between points and shadow effect. Fig. 3 illustrates this effect: the influence of the points behind the shadow is less. One should note that instead of the function $1/(1+\alpha r_j)$ other functions can be used. For example, $1/(1+(\alpha r_j)^2)$ or $\exp(-\alpha r_j)$. When filling is done we have the regular and dense grid instead of the irregular and sparse one, see Figure 4.

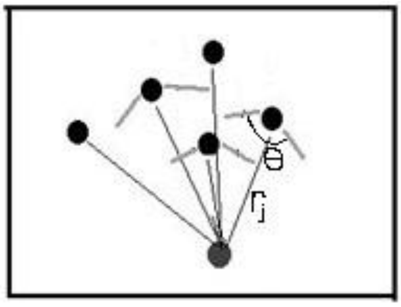


Fig. 3. Simple illustration of the method of shadows

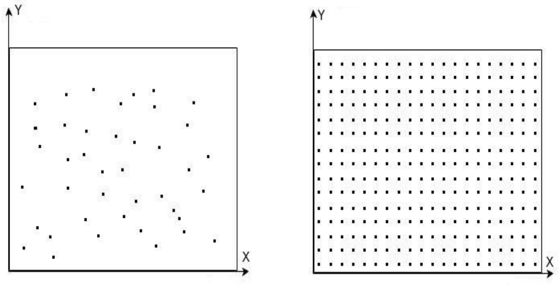


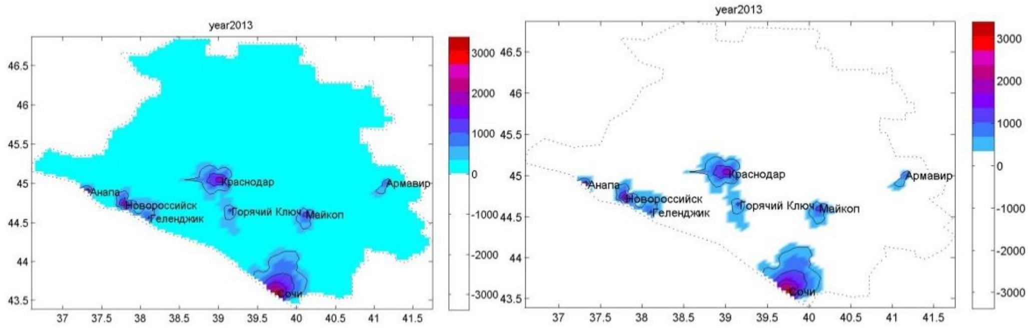
Fig. 4. The original and the resulting grids

Here: *R, n, θ* are parameters of the method.These parameters should be tuned to specific data to be processed. We did such experiments with the data of Primorskiy region of Russia and found the best values of parameters. Namely, *R* is equal to a quarter of the maximum distance between known points, *n* is taken equal to 10% of all points, and *θ* is equal to π/4:

Algorithm of visualization, procedure of 'whitening'

The traditional way for representing 2D functions are isolines. Matlab includes procedures *contour()*, *meshgrid()* to present functions in the form of isolines. However, the direct application of these procedures proves to be non-effective: user doesn't see changes in function dynamics.

We propose preprocessing procedure before visualization, which we call the 'whitening'. Here whitening is the use of white color for those parts of the plane where the function value is less than a certain threshold. To determine the optimal threshold we implemented a series of experiments. The best threshold is the level, which corresponds to the band with the largest square in this area. Figure 5 shows a map of Krasnodarskiy region in Russia with the most major cities. Color reflects the population density in different parts of the region. The numbers on the color scale on the right is the density of the population in relation to the mean density. On Figure: 5a the parts of the map where the density is less than its mean value in the region are whitened. On Figure 5b the parts of the map associated with the band of the largest square are whitened.



(a) (b)

Fig. 5. Map of Krasnodarskiy region, effect of whitening

Calculation of spatial-temporal principal components

The PCA is one of the traditional tools of multivariate data analysis. For the case of objects that are not related to space and time, principal components (PCs) are calculated according to the known algorithm:

- Calculation of correlation matrix of parameters
- Determination of eigenvalues and eigenvectors of the matrix, the eigenvectors define the directions of PCs
- Objects are projected on PC directions

The distribution of objects along the axis of the first PC is considered as an impact of the first factor, the distribution of objects along the axis of the second PC is considered as an impact of the second factor, etc.

Calculating PC for the case of spatial-temporal data is not well known. So, we explain the process of calculating on the simple example. Suppose we have m=20 points on the plane P_1,P_2,... P_{20}. Assume that at each point we have time series, which contain n=50 values: T_1, T_2, ...T_{50}. All of these values can be represented by the matrix:*U.*

Table 1. Sourse matrix

$U(m,n)$ =

Points	T_1	T_2	...	T_{50}
P_1	1.2	1.35	...	1.10
P_2	0.95	1.14	...	1.56
...	...	...	...	...
P_{20}	3.16	2.68	...	3.05

Then matrix $R(n,n)= U^T(m,n) \times U(m,n)$ is formed (values are not real).

The last matrix is matrix of correlation of time series. Its eigenvalues and eigenvectors define the spatial-temporal PCs.. Each eigenvector contains 50 elements. of n=50. We denote these eigenvectors as: γ_k = ($\gamma_{k,1}$, $\gamma_{k,2}$,……. $\gamma_{k,50}$). Here *k* is the number of *k*-th orthonotmal vector. Then we calculate the values of PCs for all points P_1,P_2,... P_{20} and all moments of time T_1, T_2, ...T_{50} using the formulas: C_1= $U\,\gamma_1$, C_2= $U\,\gamma_2$, etc. Each C_k is the matrix $C_k\,(m,n)$.

Table 2. Correlation matrix

$R\ (n,n)$ =

Points	T_1	T_2	…	T_{50}
T_1	1	0.95	…	0.28
T_2	0.95	1	…	0.36
…	…	…	…	…
T_{50}	0.28	0.36	…	1

Software and experiments

Architecture and functions of the system

Software system was developed on the software platform MatLab. The system is presented in the form of exe-application. It is oriented on end user, so it contains a user-friendly interface. One of the Interface windows is shown on Figure.6.

The system has a modular structure. It contains:

- 3 computing modules: interpolation in time, interpolation in space, calculation of PCs
- 3 input-output modules: reading and writing files, visualization, user interface

Data transfer between modules is implemented through external files. Such an autonomy allows easily to implement modifications of the system.

Typical steps of data processing are the follows:

1. Import tables with the original data;
2. Spline interpolation on a given temporal grid;
3. Spatial interpolation in each temporal layer;
4. Calculation of the first and the second PCs;
5. Visualization of the spatial dynamics for the function under consideration;
6. Visualization of the spatial dynamics for the first and the second PCs.

In those cases when the original data are given on regular and dense grid the steps (2) and (3) are absent.

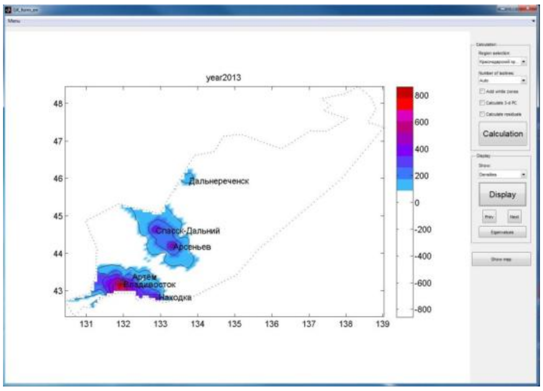


Fig. 6. Interface of the system

Testing the system on artificial example

The problem

In the artificial example the values are generated by the known function on a given spatial-temporal grid. These values are considered as the experimental data. The grid is dense and regular. So, the problem of filling function is absent here. The function itself contains components that can be explicitly linked to the factors of the first and second order. The purpose of the experiment is to identify these factors on the basis of artificial experimental data and their visualization.

Original data

We consider the mathematical function containing 3 "caps":

$$S_0(x,y,t) = \frac{1+2*t}{1+\alpha R(x,y)} \qquad S_1(x,y,t) = S_2(x,y,t) = \frac{2^t}{1+\alpha R(x,y)}.$$

Here we denote:

$$R(x,y) = \sqrt{(x-x_0)^2 + \left(y-y_0\right)^2}$$

In these formulae: x_o, y_o are the centers of the caps; x, y are coordinates of the grid; α is a parameter specifying variability of the cap. These functions are defined on the plane in the region [0;1]x[0;1] and on time interval [0;5]. Coordinates of centers for the following caps are: $x_o= y_o=0{,}5$ for S_0, $x_o=y_o=0{,}1$ for S_1, $x_o=y_o=0{,}9$ for S_2. Parameter $\alpha =5$ is for the central caps, and $\alpha =10$ is for the caps near corners. Calculation of the function is performed with the steps $\Delta x=\Delta y=0{,}02$ and Δt = 1 in space and in time respectively. Thus we have a spatial grid 50 x 50 for 6- temporal layers.

It is easy to see that there are two developing processes in the example. One is a slow process. It is associated with the cap S_0. Amplitude of the cap grows arithmetically. Another process is the fast one. It is associated with the caps S_1 and S_2. Amplitude of these caps grows exponentially.

Results

Calculations were carried out for the first and second PCs. Values of the first PC in the first and last time points are shown on Fig. 7. The first component shows a growth of the function in the same proportion on the entire region. Values of the second PC in the first and the last time points are shown on Fig. 8. The second PC shows an additional growth of the function caused by the influence of the caps located at corners.

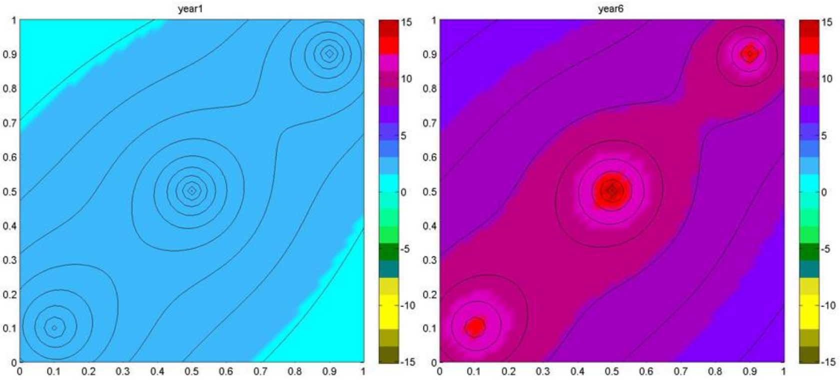


Fig. 7. The first PC in the first and the last moments of time

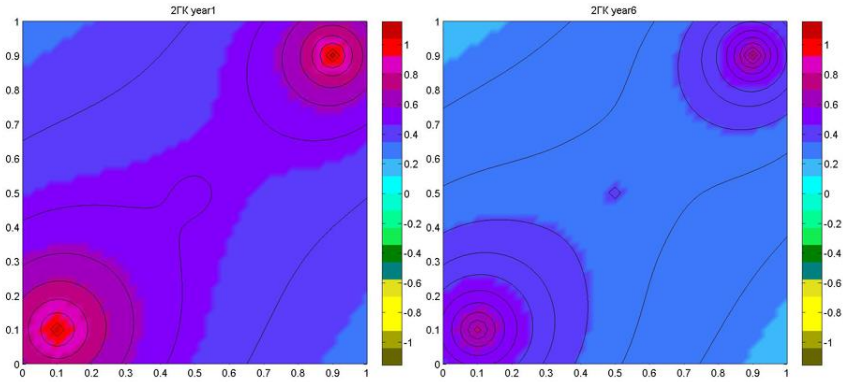


Fig. 8. The second PC in the first and the last moments of time

Therefore, one can see that the software surely distinguishes both processes.

Analysis of census data in Krasnodarskiy region

Geography of the region

Krasnodarskiy region is located in the southern Russia near Ukraine, Abkhazia, Georgia. Sochi, the capital of Olympic Games 2014, is a city in Krasnodarskiy region. The big cities in this region are Krasnodar (capital of the region), Sochi (resort town) and Novorossiysk (port). Smaller cities are Anapa, Gelendzhik and Armavir. The region has very good natural environment, and therefore agriculture is well developed here. As a consequence the percentage of rural population is high in the region. .

The last censuses were held in the USSR in 1979 and 1989 and in Russia in 2002 and 2010. Using this data the prediction for 2013 year was calculated. Thus, we have 5 time layers irregularly distributed along the time axis: 1979, 1989, 2002, 2010 and 2013.

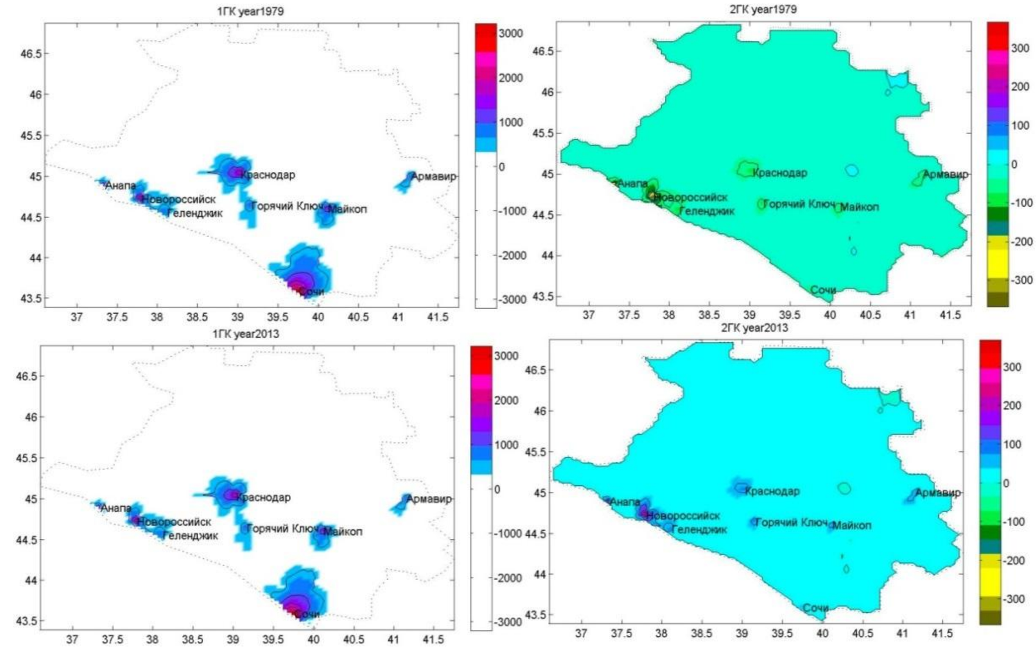
Krasnodarskiy region consists of 38 districts. The data concerning square and population in each district are known. It makes possible to calculate the density of population in these districts.

On the initial stage of processing the data of population density is calculated for each district and for each year 1979,…2013. Thus we have the data on sparse and irregular grid. This step should be considered as preprocessing. It was performed in semi-automatic mode using Excel.

Data processing

Further calculations are performed according to the scheme described in the section 3.1. Namely, the data are imported into the system and then the spatial-temporal interpolation and the calculation of the first two PCs are implemented. The results of calculation are saved in two files. The first file contains a “film”. It is a sequence of maps with population density in the region in the deferent moments of time. The second file contains values of the first and the second PCs. Fig. 9 shows the contents of this file.

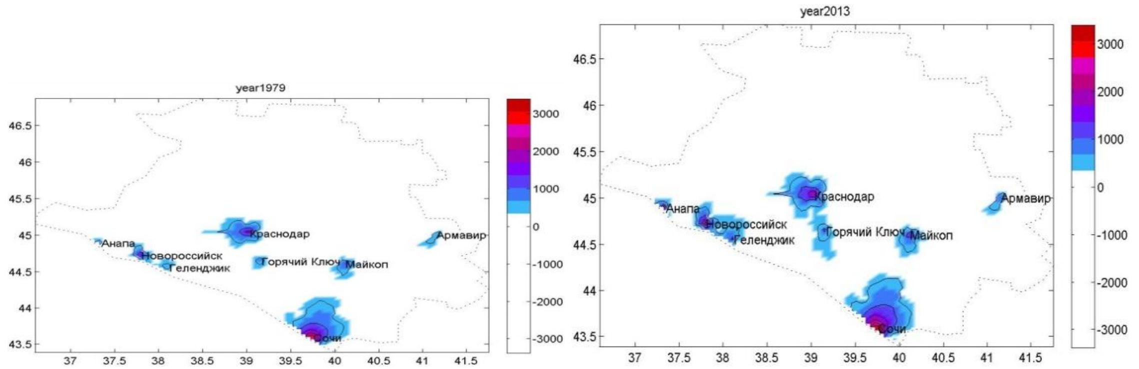
Consider the map on the top left. In the middle of the map there is Krasnodar, the capital of the region. The most southern point is Sochi. Three cities are situated on the west: Anapa, Novorossiysk and Gelendzhik. Krasnodar, Sochi and Novorossiysk are our main points. It is easy to find these cities on all other maps. On the right part of maps there is a color scale. Each color is associated with a certain population density.



(a) (b)

Fig. 9. Principal components, Krasnodarskiy region:
(a) first PC in 1979 and in 2013 years; (b) second PC in 1979 and in 2013 years;

The first PC demonstrates the process of a small growth of population density in the whole region. This growth is 18%. First of all the growth is observed in the cities of Krasnodar, Sochi, Novorossiysk mentioned above. The second PC reflects the second-order process. This is the process of redistribution of population during 44 years and an additional growth of population in Novorossiysk city. The latter can be explained by the accelerated construction of the port with its military infrastructure.



(a) *(b)*

Fig. 10. Density of population in Krasnodarskiy region: (a) in 1979 and (b) in 2013 years

Post-analysis

Effect of PCA can be seen on Fig.10. There is shown the dynamics of population density in Krasnodarskiy region in 1979 and 2013 years. White areas correspond to the threshold of whitening described in the section 2.3. In our case this value is slightly higher than the mean density in the region. The maps say almost nothing about the districts having impact on population redistribution in the region. PCA allows to reveal these hidden processes.

Conclusion

In this paper we propose the technology and software for analysis of spatial dynamics of the data given on sparse and irregular grid. In this work:

- method of shadows for filling data in space is described and realized
- algorithm of local spline interpolation for filling data in time is described and realized
- module for calculation of spatial-temporal PCs is developed; it contains open codes and can be used in other programs
- functionality of developed system is demonstrated on artificial and real data

In future we plan to implement the following modifications in the system:

- to include possibility to manage parameters of methods related to filling data using interface (now these parameters are fixed by default)
- to include module especially oriented on processing data of National censuses

Acknowledgements

The authors are grateful to Dr. Makagonov for his interest to our work and valuable advice.

Bibliography

[ArcGIS, URL] ArcGIS resourse: http://www.arcgis.com/features/

[Bahvalov , 1975] Bahvalov N. *Numerical methods* - Moscow: Nauka, 1975 (rus.)

[Hakimov , 1986] Hakimov B., Garris V. *New approach in interpolation of geological fields, Mathematical methods of investigation in gerology*. N_11, 1986 - pp.6-13 (rus.)

[Galanin, 2007] Galanin M., Guzev M., Nizkaya T. *Numerical solution of thermal plasticity problem with additional parameters* - Preprint, Inst. Appl. Math., the Russian Academy of Science, Moscow 2007 (rus.)

[Kalitkin, 1978] Kalitkin N. *Numerical methods*. Nauka, Moscow, 1978 (rus.)

[Kostyuk, 1977] Kostyuk V. *Overview of graphical output*. / / Automation experiment and computer graphics. - Tomsk: TSU, 1977. - P. 90-102 (rus.)

[Kunitsyn] Kunitsyn V. *Satellite radio probing and tomography of the atmosphere*, (rus.) URL:http://atm563.phys.msu.su/rus/text_direct.htm

[Lomtadze, 1984] Lomtadze V. *Interpolation taking into account field anisotropy and evaluation of result's accuracy, Mathematical methods of investigation in geology*. N_5, 1984. - pp.11-18 (rus.)

[Makagonov , 2003] Makagonov P. Sboychakov, K: *Interaction y diferencia en la utilization de los metodos de analisis de sistemas y metodos estadisticos en las diferentes etapas de la mineria de datos en problemas sociales* - Pachuca, Hidalgo Mexico, 2003, pp.12-15

[Torne, 2007] Torne P., *Tropical tropospheric trends*.

URL:http://www.realclimate.org/index.php/archives/2007/12/tropical-troposphere-trends/

Authors' Information

Temruk Aleksandr – M.Sc. student, Russian Presidential Academy of national economy and public administration; Prosp. Vernadskogo 82, bld. 1, Moscow, 119571, Russia; Moscow Institute of Physics and Technology (State University); Institutskii per 9., Dolgoprudny, Moscow Region, 141700, Russia) e-mail: atem@mail.ru

Major Fields of Scientific Research: mathematical modeling, data mining

Mikhail Alexandrov – Professor, Russian Presidential Academy of national economy and public administration; Prosp. Vernadskogo 82, bld. 1, Moscow, 119571, Russia; fLexSem Research Group, Autonomous University of Barcelona, 08193 Bellaterra (Barcelona), Spain; e-mail: MAlexandrov@mail.ru

Major Fields of Scientific Research: data mining, text mining, mathematical modelling

SIMPLE FREE-SHARE PACKAGE FOR VISUAL ANALYSIS OF MULTIDIMENSIONAL DATA SETS4

Alina Nasibullina, Mikhail Alexandrov, Alexander Kovaldji

Abstract: *We describe a simple free share package to study a structure of multidimensional object sets and to classify objects on these structures. The package includes 4 methods: visualization of inter-object distance distribution, object presentations in 2D subspaces of parameters, 2D subspaces of principal components (factors), 2 different alternatives. The first method is a new one: it allows to evaluate the possible number of clusters (compact groups) in data. The second and the third methods are well-known: they give possibility to see structures in data. The forth method is the new one as well: it is a convenient way to see in 2D subspace the relation to alternative multidimensional objects. A user can mark some objects as the 'good' or 'bad' ones. Their position on structures allows to implement a visual binary classification using the principal of neighborhood. All methods are managed in the interactive mode. The functionality of the system is shown on analysis of a real data set reflecting business activity of Russian companies of mobile communication.*

ACM Classification Keywords*: I.2 Artificial Intelligence*

Keywords*: visual cluster analysis, visual binary classification, software*

Introduction

Analysis of large data sets of multidimensional objects needs application of automatic data processing. But automatic procedures are practically useful if an expert can interpret results and justify the revealed regularities. Visualization is one of the effective ways for such interpretation.

In this paper we describe the developed software system that includes 4 methods of visual analysis. The first method uses inter-object distance distribution to predict the possible number of clusters (compact groups) in data. It is a new method. The second and the third methods provide object presentations in 2D subspaces of object parameters and principal components related with these parameters. It is the traditional methods. The forth method is a convenient way to see in a plane the relation to alternative multidimensional objects. It is a new method. A user can mark some objects as the 'good' or 'bad' ones. Their position on structures allows to implement visual binary classification using the principal of neighborhood. All methods are managed in the interactive mode.

By the moment we could not find simple programs for multidimensional data visualization, which would include convenient manipulations with subspaces and marked objects. We mean here such power tools for Data Mining as R, Rapid Miner, Weka [R, http], [Rapid Miner, http], [Weka, http]. From the other hand we would like to make accessible the new methods of data visualization mentioned above. These circumstances defined the actuality of the completed work.

The paper is structured by the following way. In the section 2 we describe traditional and new algorithms. In the section 3 we present the software package and show the result of experiments on real data set. Section 4 includes conclusions.

[4] Work done under partial support of the British Petroleum grant (RPANEPA - 2013).

Algorithms and software

Preprocessing

The first operation an expert should use is normalization and clearing. Speaking clearing we mean determination of outliers. We use the interval, statistical and logarithmic normalization for positive numbers:

$$x_{i\,new} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \qquad x_{i\,new} = \frac{x_i - M(x)}{\sqrt{D(x)}} \qquad x_{i\,new} = log_a(x_i + 1)$$

Here: x_i is the parameter value in the i-th observation, $M(x)$ is the expectation of parameter, $D(x)$ is the variance of parameter.

We find oulliers using Chebychev's inequality [Cramer, 1999]. Parameter *k* is selected by the user:

$$P\{|x - M(x)| < k\sigma\} \geq 1 - {}^{1}/_{k^2}$$

Traditional methods of visualization

If the source data contains more than three parameters, data representation in the space of four or more coordinates in the usual sense is impossible. There are special ways to visual data in multidimensional space, for example: parallel coordinates, Chernoff faces and flap position.

This software package uses the traditional way of visualizing multidimensional objects in 2D space of any two selected parameters. The only difficulty is how to choose these two parameters for expert could see groups of objects, because cluster structure will not be observed in each sub-space of the parameters.

Another way to visualize multidimensional data is a reduction of dimension and data presentation in 2D space of principal components. Using the principal components allows to reduce the dimension of the original data and to distort the geometric characteristics of clouds of points in the parameter space as little as it is possible.

To calculate new coordinates in the space of principal components it is necessary to find the eigenvectors and the eigenvalues of the covariance matrix of object parameters. Such a matrix is calculated as

$$M = A^T A$$

Here: *M* is a covariance matrix (*p,p*), *A* is matrix objects/attributes (*n,p*), *n* is number of objects, *p* is number of parameters. Parameter values should have the dimensionless form, i.e. it should be normalized in accordance with the problem to be solved. Traditionally the correlation matrix is used, where all the parameters have zero mean and unit variance. Matrix A is supposed to have normalized parameters. With this the covariance matrix is the correlation matrix.

Simultaneously with visualization of multidimensional data in 2D spaces the program uses the method of so-called "labeled atoms". Some specially selected points are marked on the plane.

Here the certain amount of representatives of "good" and "bad" classes is supposed to be known. Belonging of some objects within of selected structures to these classes can be determined on the basis of their neighborhood. For example, objects that are closer to "good labeled atoms" can be considered as the "good" ones.

New methods of visualization

In addition to traditional methods the program realizes 2 new methods. The first one is building diagram of inter-object distance distribution. It allows to determine the lower limit of the number of clusters

[Nasibullina, 2014]. The method was proposed 10 years ago by A.Kovaldji but by the moment it was not published.

Consider the set of N objects in a multidimensional space. Description of the algorithm includes the following steps:

1) Calculation of distances between all objects and determination of scattering (minimum and maximum distance). For *N* objects we have *N* (*N*-1) / 2 distances.

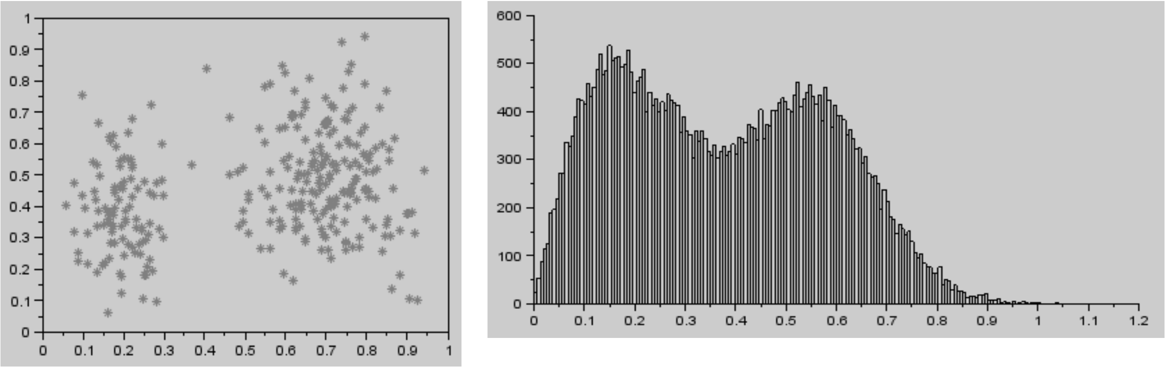
2) Building the histogram with the step equal to the several minimum inter-object distances. There is a distance on the X-axis and frequency of interval occurrence on the Y-axis. Thus, we obtain a function of the distance distribution.

3) If it is necessary, a diagram is smoothed.

Remote objects or groups of objects give far removed peaks from the origin. On the contrary, close objects or groups of objects give peaks that are near to the origin. The number of local maxima is equal to *K*=*n* (*n*+1) / 2, where *n* is the number of clusters (Figure 1). Such a formula is valid if scatter of objects in clusters is different and intercluster distances are different too. Otherwise some peaks will coincide. If scatter in clusters is more or less equal and all intercluster distances are different then number of peaks is equal *K*= *n* (*n*-1) / 2 + 1. Just figure 1 illustrates last formula.

Smoothing is necessary in order to make it easier to reveal essential peaks and to eliminate minor peaks. There are various methods of smoothing, but here we use the simple moving average.

Diagram of inter-object distances has 2 restrictions: a) intergroup distances are supposed to be greater than the inter-object ones; b) the method is suitable for clusters in the form of clouds, but no chains! Thus, this method is well suited to clearly separable clusters in the form of approximately round clouds.



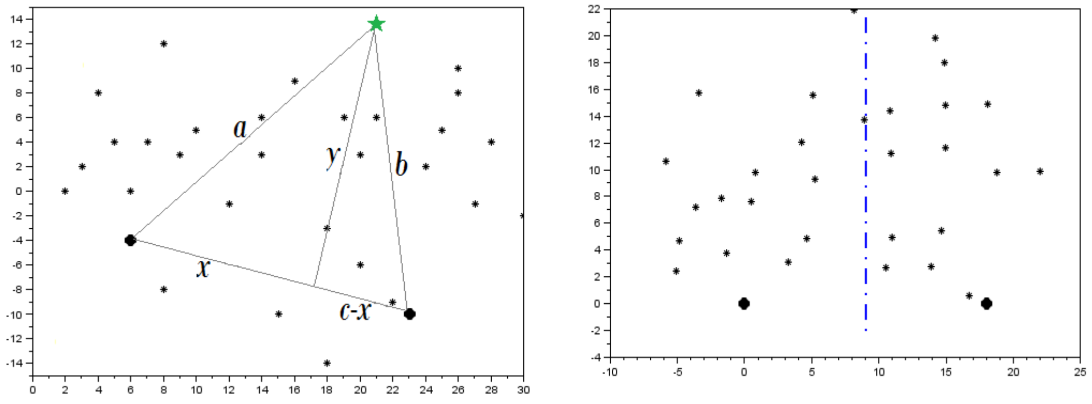
(a) (b)

Fig.1. Example on synthetic data: (a) Sets of points on the plane, (b) Diagram of inter-object distance distribution

The second new method is visualization of objects on the diagram of two alternatives. It allows to determine the proximity of objects to two points, which are known to have alternative properties. Speaking alternatives we mean an expert himself/herself assigns such objects. The principal idea is to represent multidimensional objects in 2D space. New coordinates can be calculated by the following formula:

$$x = \frac{a^2 - b^2 + c^2}{2c} \qquad y = \sqrt{a^2 - \frac{(a^2 - b^2 + c^2)^2}{4c^2}}$$

Here: x, y are new coordinates in 2D space; a, b are distance from the object to the first and to the second alternatives in multidimensional space; c is a distance between two alternatives in multidimensional space (Figure 2). The perpendicular at the center on Figure 2(b) gives possibility to say what alternative object is closer to a given object.



(a) (b)

Fig.2. (a) Points in the original space (2D as an example), (b) Points on the diagram of two alternatives

Software and example of application

Short package description

The program was developed on the platform SciLab [SciLab, http], [Alekseev, 2008]. It contains 4 modules: the computing one, man-machine interface (including visualization), and interface with data base (including preprocessing). These modules are schematically presented in Figure 3. HM-interface and BD-interface mean here human-machine interface and interface with data base respectively.

The main procedures of the computing module are:

- Calculation of distances between all objects and identification of scattering (minimum and maximum distance)
- Building a histogram with a step, which is equal to the several minimum inter-object distances (in the next version of package the step will be given as a part of maximum inter-object distance)
- Smoothing the histogram
- Calculation of geometric coordinates to represent objects in parameter sub-space for their subsequent visualization in the form of points in the selected window
- Calculation of eigenvectors and eigenvalues of the correlation matrix related to parameters of the original data
- Calculation of the principal components (PC)
- Calculation of geometric coordinates to represent objects in principal component sub-space for their subsequent visualization in the form of points in the selected window
- Calculation of the new coordinates of the objects relatively to two alternatives and their subsequent visualization in the form of points in the diagram "the worst – the best".

Interface allows to point source of information, to complete preprocessing, assign subspaces and labeled objects, select method of visualization. We demonstrate interface in the process of analysis of data related to business activity of Russian companies of mobile communication.

Source data and preprocessing

As real data to test the program we have taken data by 89 mobile companies of SPARK database. SPARC system is the largest database of companies in Russia, Ukraine and Kazakhstan. In this paper we used the following indicators of mobile companies: 1) current liquidity ratio, 2) quick liquidity ratio, 3) cash liquidity ratio, 4) equity ratio, 5) gearing ratio, 6) ratio of maneuverability, 7) average number of employees, 8) normalized net assets. The part of data are presented in the Table 1. We have conducted the interval data normalization and search for outliers using the parameter K=20 (it was recommended us). Here we used the interface shown in Figure 4.

Table 1. Data of Russian mobile companies

	Current liquidity ratio	Quick liquidity ratio	Cash liquidity ratio	Equity ratio	Gearing ratio	Ratio of maneuvera bility	Average number of employe es	Normaliz ed net assets
	2010, RUR	2010, RUR	2010, RUR	2010, RUR	2010, RUR	2010, RUR	2010	2010, RUR
1	0,621	0,617	0,490	0,351	4,391	0,879	149,4	21,520
2	0,563	0,464	0,062	0,055	23,194	-7,518	162,6	13,035
3	0,747	0,686	0,379	0,439	2,056	0,231	39,5	31,974
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
89	0,217	0,211	0,110	0,341	12,165	0,701	154,7	-1,880

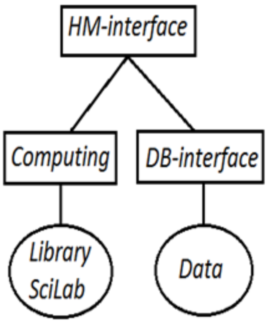
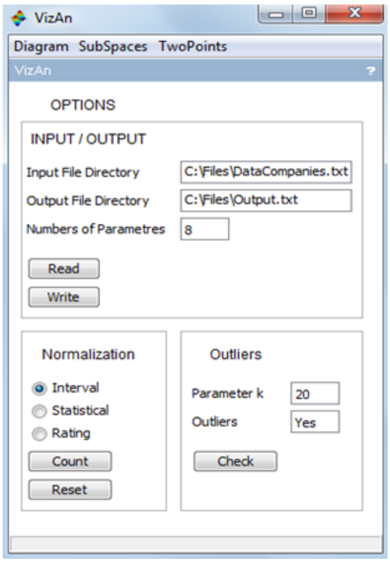


Fig.3. Structure of the system



Input File Directory: C:\Files\DataCompanies.txt

Output File Directory: C:\Files\Output.txt

Numbers of Parametres: 8

● Interval
○ Statistical
○ Rating

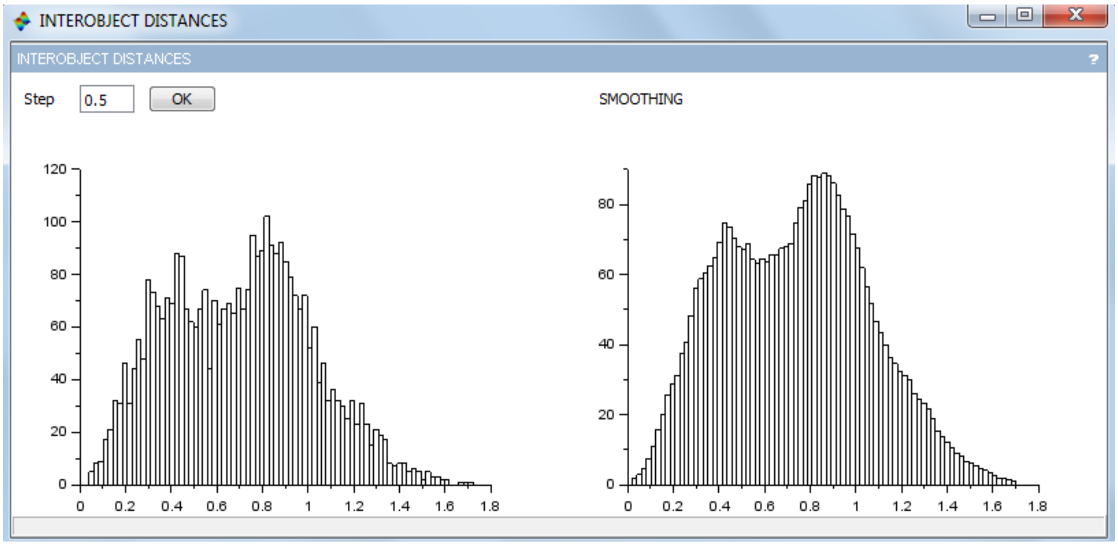
Parameter k: 20

Outliers: Yes

Fig.4. Window for preprocessing

Application of diagram of inter-object distance distribution

Generally a user tries to evaluate the approximate number of possible compact groups in data. For this the user builds the diagram of inter-object distance distribution The user can vary the step of the chart and complete smoothing with simple moving average method. The results are presented on Figure 5. We can distinguish two local maxima, so the lower bound of the number of clusters is two and we expect to see 2 compact groups of objects in one of the sub-spaces.



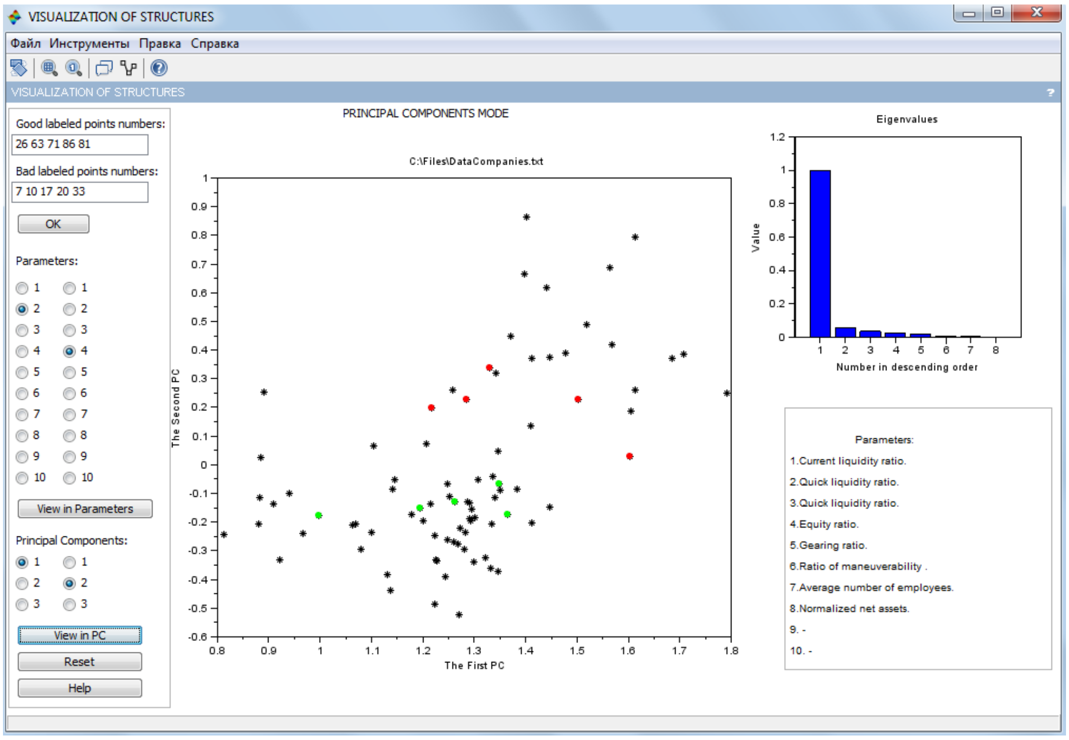
Step 0.5

Fig.5. Window with diagram of inter-object distance distribution

Visualization in the 2D parameter space and in the space of principal components

The package allows to handle objects with no more than 10 parameters. The number of PC does not exceed 3. It is enough for almost all problems being met in practice. As the procedures of choice and the presentation of objects in the space of parameters and in the space of PC are is similar then we consider only the second case.

A user specifies a pair of principal components being interested to him/her using interface on Figure. 6. The diagram of eigen-values at the corner shows that the first component can explain variations of object parameters almost completely. With the second and the third ones such a presentation will be more than enough. On Figure 6 one can see object distribution in the space of the first and the second PCs. The user can assign “good objects” and “bad objects” using fields at the upper left corner. They are presented as green and red points in the same window. The number of good and bad objects should not exceed 5 respectively. It is clearly seen that the labeled points spaced apart from each other. It allows to see the set of good and bad points on the basis of neighborhood.



VISUALIZATION OF STRUCTURES

Файл Инструменты Правка Справка

VISUALIZATION OF STRUCTURES

Good labeled points numbers: 26 63 71 86 81

Bad labeled points numbers: 7 10 17 20 33

OK

Parameters:

○ 1	○ 1
● 2	○ 2
○ 3	○ 3
○ 4	● 4
○ 5	○ 5
○ 6	○ 6
○ 7	○ 7
○ 8	○ 8
○ 9	○ 9
○ 10	○ 10

View in Parameters

Principal Components:

● 1	○ 1
○ 2	● 2
○ 3	○ 3

View in PC

Reset

Help

PRINCIPAL COMPONENTS MODE

C:\Files\DataCompanies.txt

Parameters:

1.Current liquidity ratio.

2.Quick liquidity ratio.

3.Quick liquidity ratio.

4.Equity ratio.

5.Gearing ratio.

6.Ratio of maneuverability .

7.Average number of employees.

8.Normalized net assets.

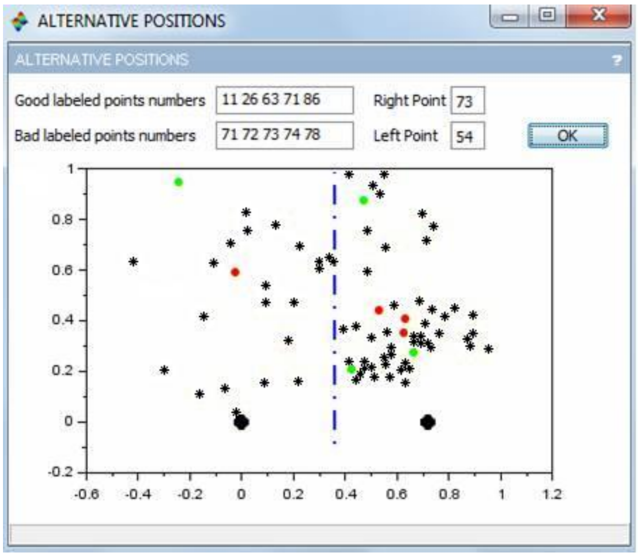
9. -

10. -

Fig.6. Visualization window in the parameter space and in the space of principal components

Visualization of objects on the diagram of two alternatives

To use such a presentation one should select two objects. One of them is considered as a good one, and the other object as the bad one. Here the user can assign “good objects” and “bad objects” as he/she did it early. The number of good and bad objects should not exceed 5 respectively. The results are presented on Figure 7. Objects located right of the central line are closer to the good object. And objects located left of the central line are closer to the bad object.



ALTERNATIVE POSITIONS

ALTERNATIVE POSITIONS

Good labeled points numbers 11 26 63 71 86

Bad labeled points numbers 71 72 73 74 78

Right Point 73

Left Point 54

OK

Fig.7. Companies in the diagram of two alternatives

One can see that the majority of labeled objects (the red and the green ones) are located right of the line. It means that both left and right alternative objects are not too contradictive.

Conclusion

This article presents possibilities of software package that allows to reveal structures in data set and select objects on these structures. We shortly described algorithms included to the package and demonstrated their work on a real data of Russian mobile phone companies.

Identifying structure in data sets is one of the problems to be resolved in Data Mining [Manning, 2008] and just inside this area Visual Data Mining occupies an essential niche [Ankerst, 2000]. So, the developed package can be considered as a tool of Data Mining.

Autumn-winter this year we suppose to release a second version of the package, which will include:

1) The algorithms providing more effective visual analysis. For this we plan to propose measures / metrics that allow to use the diagram of inter-object distance distribution for chain-like structures. Also we hope to develop algorithms to select the most interesting projections.
2) Typical algorithms of cluster analysis providing an automatic search for groups of different structure [Alexandrov, 2007]. We will use here various modifications of K-means, nearest neighbors and MajorClust methods. In these methods, we intend to use traditional measures: Euclidean, cosine linear, binary. We will use also untraditional measures, taking into account probabalistic and diffuse nature of the data parameters.

Bibliography

[Alekseev, 2008] E. Alekseev, O. Chesnokov, E. Rudchenko. Scilab. Solving engineering and mathematical problems. BINOM. 2008 (rus).

[Alexandrov, 2007] M. Alexandrov, P. Makagonov. Introduction to Technique of Clustering. In: Proc. of 3-rd Intern. Summer School on Computational Biology, Masaryk Univ. of Brno, Czech Rep., 2007, pp. 55-80.

[Ankerst , 2000] V.M. Ankerst. Visual Data Mining, Master thesis, 2000.

[Cramer, 1999] H.Cramer. Mathematical Methods of Statistics, Princeton University Press, 1999.

[Manning, 2009] C. Manning, P. Raghavan, H. Schutze. An introduction to information retrieval. Online edition. Cambridge UK, 2009.

[Nasibullina, 2014] A. Nasibullina. Evaluation of cluster number on the basis of diagram of inter-object distance distribution (this proceedings).

[RapidMiner, http] electronic resource, http://rapid-i.com .

[R-studio, http] electronic resource, http://www.rstudio.com, http://www.r-project.org.

[SciLab, http] electronic resource, http://www.scilab.org.

[Weka, http] electronic resource, http://www.cs.waikato.ac.nz/ml/weka/.

Authors' Information

Alina Nasibullina – M.Sc.Student, Russian Presidential Academy of national economy and public administration; Prosp. Vernadskogo 82, bld. 1, Moscow, 119571, Russia; Moscow Institute of Physics and Technology (State University); Institutskii per 9., Dolgoprudny, Moscow Region, 141700, Russia e-mail: *nasibullina.alinka@yandex.ru*

Major Fields of Scientific Research: Visual Data Mining

Mikhail Alexandrov – Professor, Russian Presidential Academy of national economy and public administration; Prosp. Vernadskogo 82, bld. 1, Moscow, 119571, Russia; fLexSem Research Group, Autonomous University of Barcelona, 08193 Bellaterra (Barcelona), Spain; e-mail: *MAlexandrov@mail.ru*

Major Fields of Scientific Research: data mining, text mining, mathematical modelling

Alexander Kovaldji - Deputy director for science, Director of multidisciplinary evening school, Moscow mathematical lyceum “Vtoraya Shkola”; St. Fotiyeva 18, Moscow, 119333, Russia; e-mail: *koval-dji@yandex.ru*

Major Fields of Scientific Research: mathematics, mathematical modeling

EFFICIENT DECOMPOSITION ALGORITHMS FOR SOLVING LARGE-SCALE TSP

Roman Bazylevych, Marek Pałasiński, Roman Kutelmakh, Bohdan Kuz

***Abstract**: The decomposition algorithms for solving specific Traveling Salesman Problems (TSPs) are presented. The test instances are based on the national geographic data and range in size from 9,847 cities in Japan to 115,475 cities in the USA. The proposed algorithms have a few stages: partitioning of the input set of points into small subsets, finding the partial high quality solutions, merging them into the whole initial solution, and optimization the final solution. Experimental results prove the efficiency of the proposed algorithms. Developed methods provide high quality solutions for large-scale TSP within close to the linear-logarithmic computational complexity.*

***Keywords**: TSP, large-scale, decomposition, algorithm, optimization, NP-hard.*

***ACM Classification Keywords**: G.2.1 Combinatorics - Combinatorial algorithms; I.2.8 Problem Solving, Control Methods, and Search - Heuristic methods.*

Introduction

The Travelling Salesman Problem (TSP) is extensively applied in transportation systems, automated design, testing and manufacturing of integrated circuits and printed circuit boards, X-ray crystallography and many other fields. The TSP is referred to the class of *NP*-hard combinatorial problems due to its factorial computational complexity, which unables obtaining exact solutions for large-scale problems within a reasonable runtime.

The TSP research began in the 50s of the previous century. In 1954 Dantzig, Fulkerson and Johnson defined the TSP as a discrete optimization problem and proposed a branch-and-bound method, which provides finding the optimal solution [Dantzig, 1954]. They solved an instance with 49 points and proved that no other route could be shorter. Flood [Flood, 1956] was one of the first scientists who introduced heuristic method for the problem. Lin and Kernighan devised one of the most efficient heuristic methods [Lin, 1973]. In 1972 Karp substantiated the *NP-completeness* of the problem [Karp, 1972]. The problem was also studied by many other researchers [Papadimitriou, 1977, Christofides, 1979, Reinelt, 1994, Johnson, 2002].

The studies by Applegate and others focused on finding the optimal solutions [Applegate, 1995, 1999, 2003, 2006, 2009]. They developed the "Concorde" software for providing exact solution to the problem. Recently Helsgaun has improved the classic version of the Lin-Kernighan method (LKH), which is considered as the best heuristic method so far [Helsgaun, 1998, 2006].

The Travelling Salesman Problem is formulated as follows: given is a set of points *P*, described by the their coordinates $P=\{p_1, p_2, \ldots, p_N\}$, $p_i=(x_i, y_i)$ for $i \in \{1, 2, \ldots, N\}$;

and metric *dist*: $P \times P \rightarrow R$ on the set *P*, for instance:

R_E: $dist_E(p_i, p_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$ (euclidean metrics),

or R_O: $dist_O(p_i, p_j) = | x_i - x_j | + | y_i - y_j |$ (orthogonal metrics), $i, j \in \{1, 2, \ldots, N\}$.

The problem consists in finding the closed route *M* (Hamiltonian cycle), which visits all points of the set *P* and has the minimum length: *len M* → min, where

$$len\ M = \sum_{i=1}^{N-1} dist(m_i, m_{i+1}) + dist(m_N, m_1), \qquad (1)$$

is the function of route length $M = <m_1, m_2, \ldots, m_N>$, $\forall i,j\ [m_i, m_j \in P, m_i \neq m_j]$, $|M| = N$.

Function $f_{quality}$ is used to measure the quality of the solution:

$$f_{quality} = \frac{len\ M - len\ M^*}{len\ M^*}, \qquad (2)$$

where *M** is the optimal solution.

When *M** is minimum length route: $len\ M \geq$ len M*, and $f_{quality}(M, M^*) \geq 0$.

Given $f_{quality}(M, M^*) = 0$ the route *M* (problem solution) is considered as an optimal, otherwise it is a suboptimal. The smaller function fquality (M, M*) value is, the closer to the optimal the problem solution is.

When the optimal problem solution (route *M**) is not given, the quality function $f'_{quality}$ for problem solution (route *M*) is applied, which for the given set of point *P* instead of optimal route length uses values of Held-Karp lower bound [Held, 1970, 1971, 1974]:

$$f'_{quality}(M, HKbound) = \frac{len\ M - HK_{bound}}{HK_{bound}}, \qquad (3)$$

$$f'_{quality}(M, HKbound) > 0, \qquad (4)$$

where *HKbound* is value of Held-Karp lower bound for the set of points *P*. The smaller the value of function $f'_{quality}$ is, the better the solution *M* is obtained.

When the optimal problem solution (route *M**) and the value of Held-Karp lower bound are not given, a compare function with the best known solution M_{known} found by the existing methods is used in order to estimate quality. If $f_{compare}(M, M_{known}) < 0$, the solution *M* is considered as the better comparatively with existing known one.

There are *N*! different alternative routes (solutions) via the given set of points *P*, and finding the optimal route can be either through the search of all possible options, that for the large-scale problems is impossible, either by branch and bound method, which also requires considerable computational cost.

Despite the fact that the general number of possible routes is finite, even advanced or future supercomputers are not able to conduct such search for many thousands or larger number of points. Therefore, many contemporary research works are focused on finding the suboptimal solutions for reasonable time, which are close to the optimal.

Decomposition and finding initial solution

The solving process involves the following main stages:

1) partitioning of the input set *P* into set $U = \{U_1, U_2, \ldots, U_K\}$ with *K* subsets: $P = U_1 \cup U_2 \cup \ldots \cup U_K$, $U_i \cap U_j = 0$, $D_{min} \leq |U_i| \leq D_{max}$, where D_{min} and D_{max} – respectively the minimum and maximum number of points in the subsets;
2) selection of the initial subset U_1 and finding its TSP solution (route M_1);
3) sequential extention of existing in *i*-step solution M_i by merging it with the partial solution ΔM_{i+1} for the adjacent subset U_{i+1} of points. New solution M_{i+1} is created;

4) continuation of the previous procedure until the inclusion of all points of set *P* into solution M_0 that is considered as an initial solution.

In order to extend for the (i+1)-step solutions M_i we consider the two subsets of points: *Ui* (all previous points) and additional U_{i+1}, that have overlapping $\Delta U_{i,i+1} = U_i \cap U_{i+1}$. The numbers of points in the set U_{i+1} and points in the set $\Delta U_{i,i+1}$ are the method's parameters which affects the quality of solution and running time. *Boundary entry* and *exit points* are defined for the set U_{i+1}. The rest part of the route M_i, which is not included in the set U_{i+1} is replaced by the fixed edges of the zero length.

With the help of the chosen method the TSP solution ΔM_{i+1} for the points within the set U_{i+1} is found. A new route M_{i+1} is formed by the route ΔM_{i+1} and segments of the route M_i, as a result of merging of the solutions in the subsets U_i and U_{i+1} (Fig. 1).

The procedure of the solutions' merging in the subsets continues till all subsets of the set *U* are united. The resulting route, covering all points of the set *P*, is viewed as the *initial solution* M_0 of the problem.

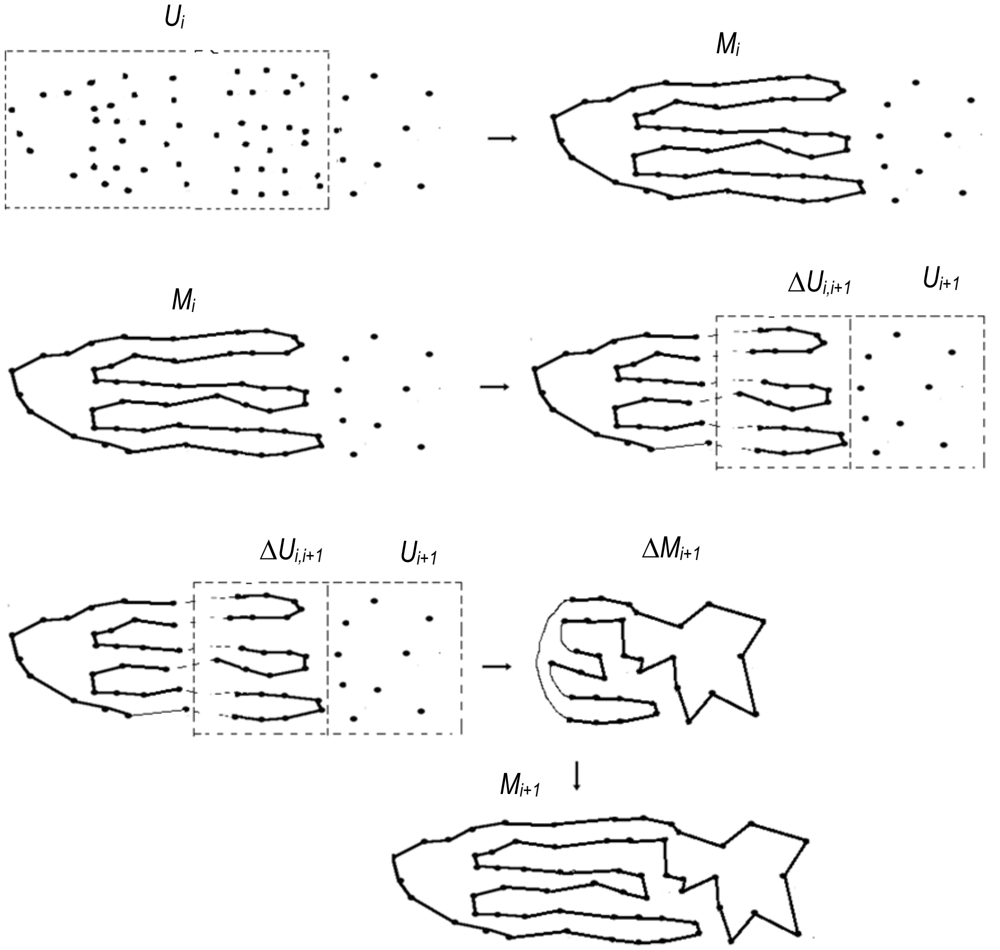


Fig. 1. Solution extension process

There are a number of algorithms of selecting the initial subset and subsequent subsets for solution extending. For example, from the left to the right merging of subsets, or alternatively, zigzag, spirally from some corner or center etc [Bazylevych, 2012].

Solution optimization

The results of experiments show that applying the extension method allows finding the initial route M_0 which on average 0,2-2% exceeds the length of the optimal one. It is required to use optimization methods to improve its quality. The reduction of the length of the route M_0 is provided through its iterative reduction in the certain Local Optimization Areas (LOA).

The method of optimization [Bazylevych, 2008, 2009] have the following features:

- size of the *optimization area* – the number of its points;
- size of the *overlapping area* – the number of points of the intersection area of two or more adjacent *optimization areas;*
- *strategy* (sequence, or direction) of optimization;
- *basic method* - the known method used for the TSP in a given LOA.

For the solution optimization the certain LOA is selected. In case of route length reduction, this area is replaced with a new one. The result is the route M_1, where $len\ M_1 < len\ M_0$. The process is repeated for all LOAs until all points of the route M_0 are reviewed.

The sequence of routes M_0, M_1, M_2, …, M_k, is obtained, where $len\ M_{i+1} < len\ M_i$ for $i \in \{0, 1, …, k\}$, and k is the number of area replacements on the route M_0 for shorter ones. Complete optimization process can be repeated several times until the length stops changing or the changes are insignificant.

The results of experiments prove that with the LOA size increase the quality of the solution improves, but computation time also increases. The quality also depends on the selected *basic method*. We recommend applying efficient *Lin-Kernighan* or *Lin-Kernighan-Helsgaun* methods.

Delaunay triangulation based optimization

This method is aimed to decrease the length of the route M_0 by sequential “scanning” the different LOAs along the initial route including also not only the points belonging to this route segment, but also points of other segments, which may be far away from this route segment, but close geometrically.

The initial route M_0 is divided into set of segments (LOAs) $S = \{S_1, S_2, …, S_r\}$, each of which has given number D of points (its size), and every two adjacent LOAs have the *overlapping* area which given number C of points (its size). The third parameter of the method is the *Depth* of the LOA (Fig. 2).

The set of points P is triangulated by the Delaunay algorithm [Guibas, 1985], obtaining the set of triangles $T=\{t_1, t_2, …, t_w\}$, $|T|=w$, $w \approx 3N$, every of which is described by their points $t_i=(p_{i1}, p_{i2}, p_{i3})$; $p_{i1}, p_{i2}, p_{i3} \in P$ for $i \in \{1, 2, …, w\}$. At the first step we choose an arbitrary point on the existed M_i road and spread around it the waves in triangles until the resulting region (LOA) includes the desired D number of points (dot line in the Fig. 3a). At the second step (Fig. 3b) we eliminate all pieces of existing road M_i and replace it external pieces (dashed line outside of LOA in the Fig. 3a) by fictitious pieces of zero length (continuous line beyond the LOA). At the third step (Fig. 3c) we solve the TSP in selected LOA. Finally, at the last step (Fig. 3d), the external fictitious pieces are replaced by the real ones (dashed line).

The replacement of the segments S_i ($i= 1,…, r$) continue until the optimization of all areas of the route M_0. As a result, the route M_1 is obtained, which is considered as *optimized*. The computational complexity of the optimization method is $O(N \log N + KD^{2,2})$, where K is the number of LOAs of optimization. Since the value D

is constant, K is the linear function of N and $K \ll N$, the computational complexity is $O(N \log N + K) \approx O(N \log N)$.

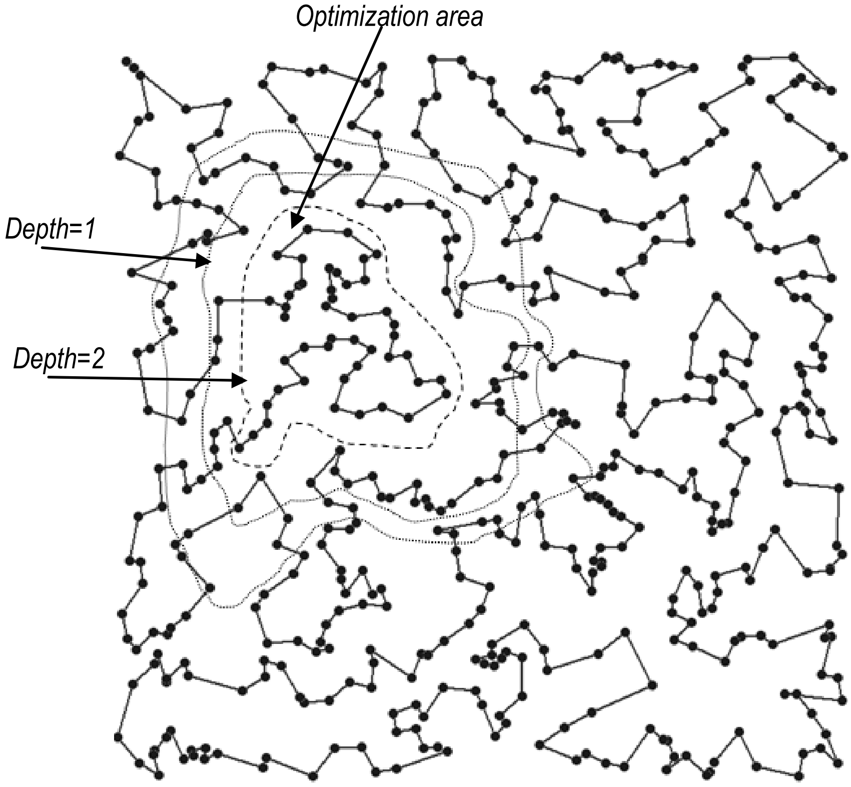


Fig. 2. Delaunay triangulation based optimization method and its features: size and depth

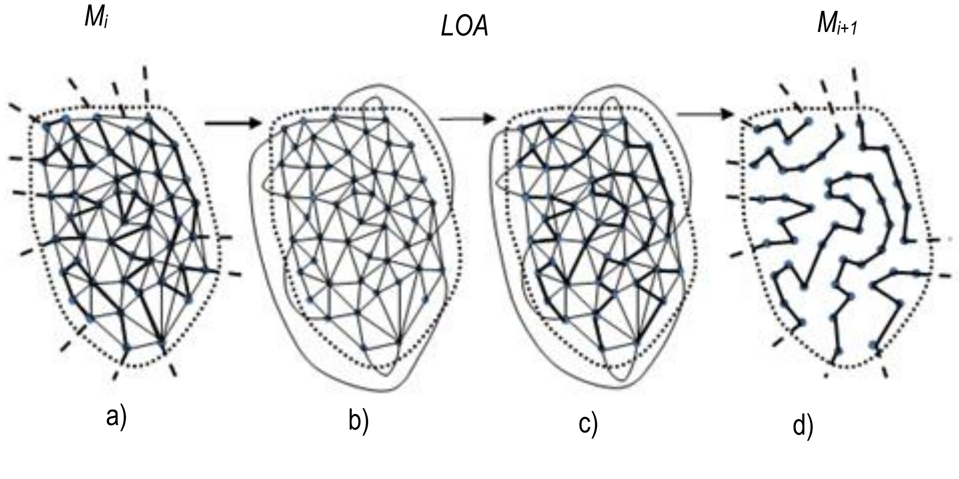


Fig. 3. Steps of replacement of route areas for shorter ones using optimization by route scanning method with Delaunay triangulation

Experimental results

Investigated how the parameters of optimization affect on quality of the solution optimization. The problem ch71009 was chosen for testing [National TSPs]. Initial route, 0,53% longer than the current best solution, was obtained by solution extending method [Bazylevych, 2012]. The following parameters were investigated: *size* of the LOA and *overlapping*. The size of the LOA varied from 100 to 2000 points, overlapping value varied from 10% to 80%. Test were performed on 3,5 GHz CPU. Table 1 provides the quality and running time of solution optimization.

The set of test instances, based on the national geographic data were chosen [National TSPs]. They vary in size from 9,847 cities in Japan to 115,475 cities in the USA (Fig. 4). The solving algorithm had a few stages: partitioning of the input set of points into small subsets, finding the initial solution, and optimization phase. Table 2 provides the results of finding solutions for test instances using the proposed decomposition and optimization algorithms.

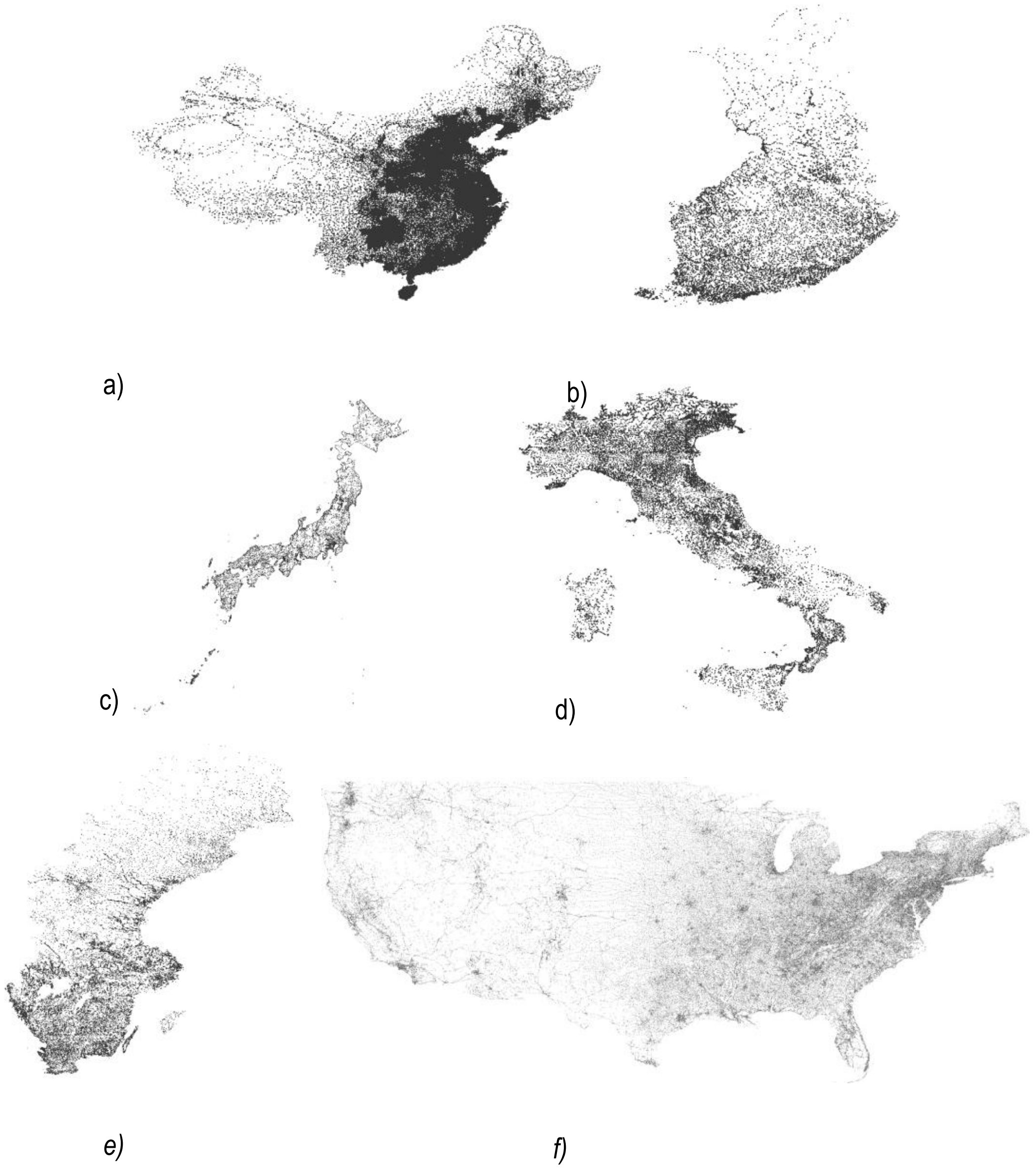


Fig. 4. Test instances: a) China (71009 points), b) Finland (10639), c) Japan (9847),
d) Italy (16862), e) Sweden (24978), f) the USA (115475)

The following pictures show the comparison in some areas of the route between the initial route and the optimized one. In some cases, the optimized route has no “inefficient” areas with long edges (Fig. 5). Also there are some changes in “global” route (Fig. 6). Figure 7 also shows some changes of the route segments.

Table 1. Quality (in %) and runtime (in seconds) of the optimized solutions (ch71009)

Size of LOA / Overlapping	100	500	1000	1500
10%	0,36% 3,3s	0,27% 4,7s	0,24% 6,7s	0,17% 8,4s
30%	0,29% 4,5s	0,30% 6,2s	0,19% 7,6s	0,13% 9,9s
40%	0,27% 4,9s	0,26% 6,9s	0,14% 10,3s	0,13% 11,8s
80%	0,26% 13,8s	0,19% 19,8s	0,08% 26,4s	0,10% 31,5s

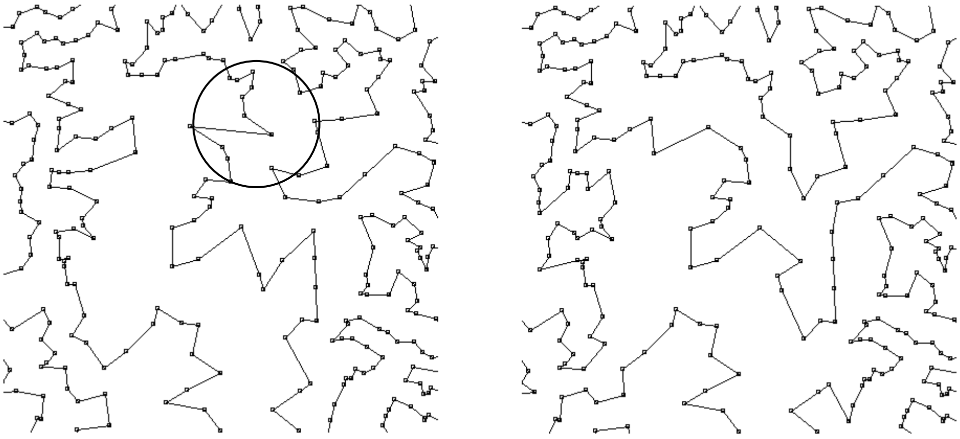


Fig. 5. Comparison of some segments of initial route and the optimized route (long edge)

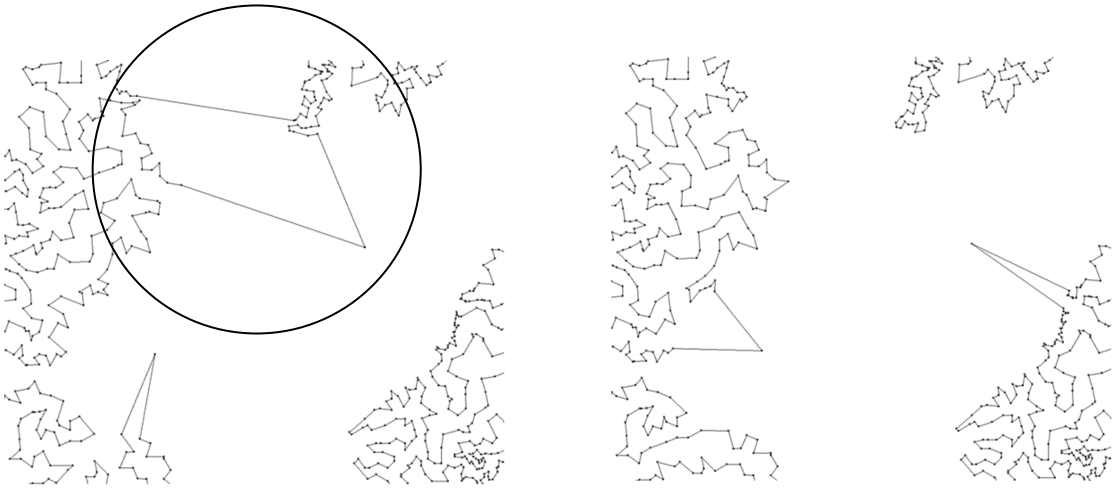


Fig. 6. Comparison of some segments of initial route and the optimized route (optimized more "globally")

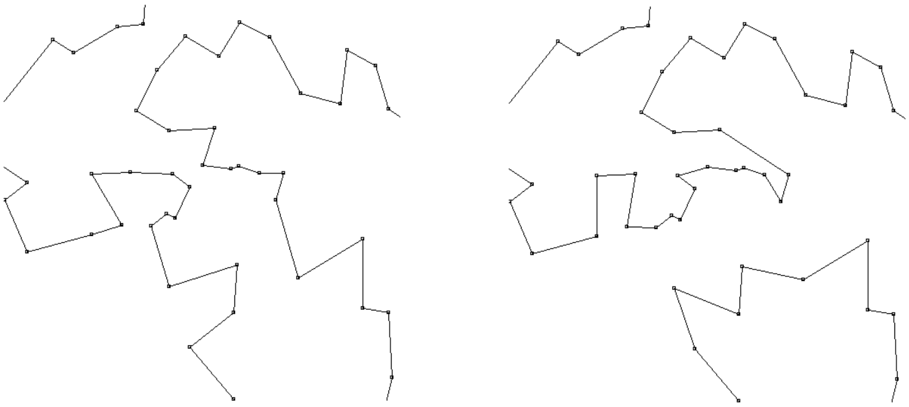


Fig. 7. Comparison of some segments of initial route and the optimized route

Table 2. Experimental results

Test problem	Number of points	Time, minutes	Quality
ch71009, China	71009	50,7	0,07%
fi10639, Finland	10639	10,7	0,05%
it16862, Italy	16862	15,9	0,02%
ja9847, Japan	9847	9,5	0,02%
sw24978, Sweden	24978	22,5	0,06%
usa115475, USA	115475	85,8	0,08%

Conclusion

New efficient decomposition and optimization methods, based on Delaunay triangulation, have been investigated for solving the large-scale travelling salesman problem. The computational complexity is close to linear-logarithmic. The problem is solved in several stages: partitioning the input set of points into subsets of limited sizes (≈ 800-2000 points); receiving the initial solution by merging partial solutions and its improvement by the developed optimization method. Methods provide at most 0.08% deviation from the best known solutions of the investigated problems of national TSPs and require much less time in comparison with the best existing heuristic or exact methods.

The work was performed as part of the project that has been partially funded by the State Agency for Science, Innovation and Informatics Implementation of Ukraine.

Bibliography

[Applegate, 1995] D. Applegate, R. Bixby, V. Chvátal V, W. Cook. Finding Cuts in the TSP // DIMACS Technical Report 95-05. – Rutgers University. – 1995.

[Applegate, 1999] D. Applegate, R. Bixby, V. Chvátal V, W. Cook. Finding Tours in the TSP // Technical report 99885, Research Institute for Discrete Mathematics. – Universität Bonn. – 1999.

[Applegate, 2003] D. Applegate, R. Bixby, V. Chvátal V, W. Cook. Implementing the Dantzig-Fulkerson-Johnson Algorithm for Large Traveling Salesman Problems // Mathematical Programming (2003). – (Series B) 97. – pp. 91-153.

[Applegate, 2006] D. Applegate, R. Bixby, V. Chvátal V, W. Cook. The Traveling Salesman Problem – A Computational Study // Princeton Series in Applied Mathematics. – Princeton University Press. – 2006.

[Applegate, 2009] D. Applegate, R. Bixby, V. Chvátal V, W. Cook, D. Espinoza, M. Goycoolea, K. Helsgaun. Certification of an Optimal TSP Tour Through 85,900 Cities // Operations Research Letters. – 37 (2009). – pp. 11—15.

[Bazylevych, 2008] R. Bazylevych, R. Kutelmakh, B. Prasad, L. Bazylevych. Decomposition and Scanning Optimization Algorithms for TSP // Proceedings of the International Conference on Theoretical and Mathematical Foundations of Computer Science. – Orlando, USA. – 2008. – pp. 110-116.

[Bazylevych, 2009] R. A. Bazylevych, B. Prasad, R. Kutelmakh, R. Dupas, L. Bazylevych. A Decomposition Algorithm for Uniform Traveling Salesman Problem / Bazylevych R., // Proceedings of the 4th Indian International Conference on Artificial Intelligence. – Tumkur, India. – 2009. – pp. 47-59.

[Bazylevych, 2009] R. Bazylevych, R. Kutelmakh. Optimization of TSP solutions by sequential scanning method // Visnyk of Lviv Polytechnic National University. – 2009. – № 638: "Computer sciences and information technologies". – pp. 254-260 (In Ukrainian).

[Bazylevych, 2012] R.P. Bazylevych, M. Palasinski et al. "Decomposition methods for large-scale TSP". In book: G. Setlak, M. Alexandrow, K. Markow. "Artificial intelligence methods and techniques for business and engineering application". ITHEA, Rzeszow–Sofia, 2012, pp. 148 – 157.

[Christofides, 1979] N. Christofides. The Traveling Salesman Problem // Combinatorial Optimization. N. Christophides, A. Mingozzi, P. Toth and C. Sandi. Eds. – John Wiley and Sons, New York. – 1979.

[Dantzig, 1954] G. Dantzig, R. Fulkerson, S. Johnson. Solution of a Large-Scale Traveling-Salesman Problem // Operations Research. – 1954. – Vol. 2. – pp. 393-410.

[Flood, 1956] M. Flood. The traveling-salesman problem // Oper. Res. – 1956. – N. 4. – pp. 61-75.

[Guibas, 1985] L. Guibas, J. Stolfi. Primitives for the manipulation of general subdivisions and the computation of Voronoi // ACM Transactions on Graphics (TOG) . – 1985. – Volume 4, Issue 2. – pp. 74-123.

[Held, 1970] M. Held, R.M. Karp. The Traveling Salesman Problem and Minimum Spanning Trees // Operations Research. – 1970. – Vol. 18. – P. 1138–1162.

[Held, 1971] M. Held, R.M. Karp. The Traveling Salesman Problem and Minimum Spanning Trees: part II // Mathematical Programming. – 1971. – Vol. 1. – P. 6-25.

[Held, 1974] M. Held, P. Wolfe, H.P. Crowder. Validation of subgradient optimization // Mathematical Programming. – 1974. – Vol. 6. – P. 62-88.

[Helsgaun, 1998] K. Helsgaun. An Effective Implementation of the Lin–Kernighan Traveling Salesman Heuristic // Datalogiske Skrifter (Writings on Computer Science). – 1998. – No. 81. – Roskilde University.

[Helsgaun, 2006] K. Helsgaun. An Effective Implementation of k-Opt Moves for the Lin–Kernighan TSP Heuristic // Datalogiske Skrifter (Writings on Computer Science). – 2006. – No. 109. – Roskilde University.

[Karp, 1972] R. Karp. Reducibility among combinatorial problems // In Raymond E.Miller and James W.Thatcher, editors, Complexity of Computer Computations. – 1972. – Plenum Press, New York. – pp. 85-103.

[Lin, 1973] S. Lin, B.W. Kernighan. An Effective Heuristic Algorithm for the Traveling-Salesman Problem // Operations Research. – 1973. – Vol. 21, No. 2. – pp. 498-516.

[National TSPs] http://www.math.uwaterloo.ca/tsp/world/countries.html

[Papadimitriou, 1977] C.H. Papadimitriou. The Euclidean traveling salesman problem is NP-complete // Theoret. Comput. Sci. – 1977. – No. 4. – pp. 237-244.

[Reinelt, 1994] G. Reinelt. The Traveling Salesman Problem: Computational Solutions for TSP Applications // Lecture Notes in Computer Science. – 840, Springer-Verlag. – Berlin. – 1994.

Authors' Information

Roman Bazylevych – Full Professor, Ph.D., D.Sc, Mathematics and Computer Science Foundations, University of Information Technology and Management in Rzeszow, Poland and Software Engineering Department, Lviv Polytechnic National University, Ukraine; e-mail: *rbaz@polynet.lviv.ua*

Major Fields of Scientific Research: Computer Science, Design Automation, Algorithms, Combinatorial Optimization

Marek Pałasiński – Prof. nadzw. dr.hab., Mathematics and Computer Science Foundations, University of Information Technology and Management in Rzeszow, Poland e-mail: *mpalasinski@wsiz.rzeszow.pl*

Major Fields of Scientific Research: Theoretical computer science, Theory of algorithms, Graph theory, Data mining and Algebraic logic

Roman Kutelmakh – Assistant Professor, Ph.D., Software Engineering Department, Lviv Polytechnic National University, Ukraine; e-mail: *rkutelmakh@polynet.lviv.ua*

Major Fields of Scientific Research: Software technologies, Combinatorial Optimization, Algorithm design, Vehicle Routing Problems

Bohdan Kuz – Assistant Professor, Software Engineering Department, Lviv Polytechnic National University, Ukraine; e-mail: *bohdankuz@gmail.com*

Major Fields of Scientific Research: Software technologies, Combinatorial Optimization

QUEUING BASED SIMULATION MODELS FOR ANALYZING RUNWAY CAPACITY AND MANAGING SLOTS AT THE AIRPORTS

Sumeer Chakuu, Michał Nędza

***Abstract**: The scarcity of the runway slots is the crucial factor, which has a drastic influence on the aviation market worldwide. Both airports and airlines suffer because of the lack of the slots in the periods they are the mostly desired and when they would yield the highest profits. The biggest problem is the infrastructure or being more precisely the lack of the airport runway capacity or utilization and resulting attributes which lead to the difficulties in traffic operations. The problem is necessary to solve as it will reduce the costs involved and increase the profits generated. The main aim of this article revolves around the justification of the fact that the various simulation methodologies can be incorporated to solve this issue. The simulation models, which are used in this article, are the queuing based models as they precisely determine the behavior exhibited by the runway systems.*

***Keywords**: Slot Management, Runway Capacity analysis, Air Traffic Controllers, Queuing Theory and Models, Probabilistic Distribution, Simulation and Optimization*

***ACM Classification Keywords**: B.2.2 Performance Analysis and Design Aids Simulation, B.4.4 Performance Analysis and Design Aids, B.5.2 Design Aids, F.1.2 Modes of Computation, G.3 Probability and Statistics, G.m Miscellaneous. .*

Introduction

This article provides the theoretical concept for examining the standard day of operation at the airports and their evaluation from the airside perspective. In the process of the evaluation, the models of Queuing theory has been used which will lead to the design of the automated system. The simulation has been applied with adoption of three different models of queue. Each model is characterized by the different probability distributions of the time while examining the aircraft's occupancy of the runway.

To realize the above stated objective following sub-goals are to be taken into account:

- Selection of mathematical models of examining the queues for the decision situations given by the issue.
- Implementation of the chosen models
- Derivation of the hypothesis to check their correctness with the usage of applied models.
- Modeling the simulation process with application of M/M/1 , M/G/1 and M/Er/1 models.

The absolute necessity in the projection of runway operations is to perform a simulation. The model, which will be designed, will show the most important features of operational activities on the runway. The model should consist of the expected time of each aircraft movement that has been taken from the historical data. The model has to also include the slot time intervals and separation times. The major part of the simulation is to calculate the inter-arrival time and the service time (the time each scheduled aircraft occupies the runway in the sense of landing or taking-off). All these actions have to be been taken into account to check if the airport, in the examining period of the day, suffers from the waiting lines or all the operations are performed without any delay and in an optimal way. Three different models stated above are based on the similar concept and are key in simulating the runway operations. The article describes two types of the simulation – short and long run. Its purpose is to check if the number of cases influences the received results. The simulation is needful to put into practice the theoretical concepts of functioning of the airport on its airside.

After the simulation models will be planned and implemented, it will be important to form a particular hypothesis that will be examined at the later stage of the research.

The article covers the Queuing Theory modeling concepts and its implementation environment in the field of slot management in detail. Slot management is the most important theoretical part as it describes the importance of the issue. It starts with basic problem scope, followed by the various concepts that regulate the process of allocating and coordinating the slots and number of movements on the runway.

Problem Scope

Most of the international airports suffer due to the lack of the capacity. In Europe, the situation is complex, as the airports built decades ago are not planned for huge flow of passengers and cargo. The European Commission confers the statuses of coordinated and uncoordinated to each airport on the territory of the Community. The major problems are insufficient infrastructure to handle the demand and the problem of scarcity of slots, limited by the infrastructure and international regulations. While the background is constrained often by the small area occupied by the airport and the situation is not expected to improve. Only reasonable improvement proposed by many experts is application of market mechanisms [Doganis, 1991].

This article presents the approach to the problem with the application of the queuing theory. Besides the mathematical studies over the queues appearing on the runway at an airport during everyday activities, the models designed in this article can be also used to simulate the amount of money the airlines loose globally when their aircraft misses a slot and has to wait for the next available.

Slot Management

The main purpose of this part is to present the concept of the slot management, which is the backbone of the runway operations. The aviation area distinguishes two types of slots, namely, the airway slot and the airport slot (which is also known as a landing slot or runway slot). The latter is the right, which is allocated to the specific entity (like commercial airline) by the airport and which allows the owner of the slot to perform landing or departure on the runway in the determined period of time [Doganis, 2001]. Airport slot is mandatory at coordinated airports for each movement (arrival and departure) and is valid for a specific time and specific weekday for the complete planning season (summer/winter season). The airport slot is used to plan the runway capacity (and/or other constraints) for a period of half a year to minimize airport congestion and potential cancellation or delays.

On the other hand, Airway slot or the Air Traffic Control Slot is needed in case of traffic limitations in the airspace. It is provided for a departing flight and is only valid for this specific light, for a specific departure time window (15 minutes) during a specific day. According to European Regulation EEC, “the airport slot means the permission given by a coordinator, in accordance with the Regulation to use the full range of airport infrastructure necessary to operate an air service at a coordinated airport on a specific date and time for the purpose of landing or take-off, as allocated by a coordinator in accordance with this Regulation” [Official Journal L 138, 2004].

The scarcity of the runway slots is the crucial factor that influence the aviation market worldwide. Both airports and airlines suffer because of the lack of the slots in the periods they are the mostly desired. This problem turns the clock back and limits the development of this extremely important arm of global economy.

Due to an imbalance between the demand for worldwide air transport and the availability of adequate airport facilities/infrastructure and airspace systems to meet such demand, the number of congested airports is growing. As a result, the airlines industry is increasingly subjected to serious operational disruptions, with a significant number of delayed departures and arrivals, which results in significant economic penalties.

Runway Capacity

The capacity of the runway is simply the number of movements (counted on hourly or daily basis) that the airport is able to serve or is allowed to serve by the international regulations [Simpson, Belobaba, 1992]. Capacity of the runway is the crucial constructive factor considering the group of restrictions affecting the number of slots that airports offer. It presents a performance of runway system and depends on many elements. The major element being the number of runways and their independent utilization. At some airports, there are more than four runways but they are not allowed to use them in parallel. The great impact on the system is also exerted by the surrounding area (the topography of the landscape) and obviously approach and departure routes. All aircrafts are taking-off opposite to the direction of the wind and if it changes, it automatically limits runway capacity. If the airport is located in the neighborhood of water region or in wooded area it can possible affect the landing approach what eventually could translate into regularity of the operation on the runway [Madas, 2007].

Additionally the ICAO regulations apply at all airports, which limit the number of movements in time. The authorities also put these regulations into practice with compliance of their own runway capacity.

Peak period problem

The peak period is time where the majority of airlines are willing to make their movements at the airports. However, the number of slots limits it. As air traffic grows, demand can exceed capacity at key points of the air transportation network and at critical times. These local overloads create delays, which propagate to other parts of the air network, amplifying congestion as increasing numbers of local capacity constraints come into play. Moreover, the average delay generally increases faster than linearly with traffic.

As it is observable, the peaks occur during the specific period of the day and create the waves of the periods where the number of flights are greater and smaller. The insufficient number of slots or the physical constrains create delays in the system. When such a queue begins to create, the ATC controllers and the airport authorities cannot simply utilize First-In-First-Out rule, because next flights are supposed to start or land on time to avoid further delays. Delayed flights have to wait even for tens of minutes for their movements. Moreover the situation on the destination airport has to be taken into consideration as well, if the aircraft starts but will not be able to land on target airport it will create even more delay and perturbation at the other airport. It creates the absurd situation where the flight despite the fact it possess the right to start and available slot will not be able to do so due to lack of permission to land at its destination [ACI, 2007] .

The peak period problem will be also taken care of by the models developed in this article. The historical data will serve to investigate the queue that is created during the delays in the system and finally the model will examine the stability and intensity of these queues.

Queuing Theory for determining runway utilization and managing slots

The main purpose of this section is a presentation of concepts of the sphere of knowledge known as queuing theory (or traffic theory) [Bose, 2002]. The subject of this conception, studies the waiting lines from the mathematical point of view. The mathematical model of functioning of mass service systems are based on the stochastic theory [Gajowniczek, 2008]. The theory provides an opportunity to investigate a number of processes, which are related to each other. It could include arrival of the customers, the process of waiting in the queue (that is known also as a storage process) and the service at one or more servers. The term of traffic theory is often applied to the theories of telephone and communication traffic; however, it finds adoptions in various areas of life like designing of hospitals, shops, or factories. The main aim of this article is to apply these mathematical models to study the waiting lines in the utilization of airport runway. As it has been mentioned before, the optimization of runway capacity is a key success factor in functioning of the airport at its airside as well in avoiding possible delays resulting from queues. The theory permits the differentiation and calculation of several performance measures like the average waiting time in queue or in the system, expected number of clients in the beginning of queue or in the service station or generally investigation if the whole system is stable or there are delays in it. The application of the queuing theory will not solve the capacity problems, but instead it can be used to provide some suggestions about the use of runway. If the aircraft is scheduled to land or take-off at the certain time and it does not find the available runway, it must take some specified action such as waiting or flying away. Whereas the latter is not very likely, we can consider the model of waiting in the line and define it in terms of three characteristics: input process, the service mechanism and queue discipline [Cooper, 1981]. The basic queuing model at the runway is shown in figure 1.

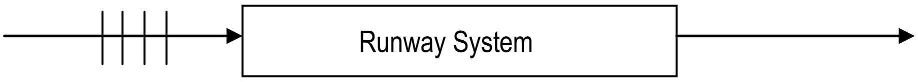


Fig.1. Basic runway queuing model.

If the time of the service is greater than the average separation between the arriving customers, the queue can become longer in infinity. The input process or the arrival process of the customers describes the sequence of request of service. Often, the arrival process is specified under condition of the distribution of lengths of time between consecutive arrival instants. Usually there is an assumption, which provides inter-arrival times in independent way and gives them common distribution. In the most cases, the researchers assume that customers arrive according to the exponential inter-arrival times (Poisson stream). The arrival time counts the time that last between appearance of the customer, which is the last in queue, and the customer, which has just arrived.

The service mechanism includes such characteristics as the number of servers and the period of time the client holds the server. For example, clients may be processed in a single server and each client holds the station for the same length of time. In this particular investigating case, the focus us on the single server model (the airport with only one runway).

The queue discipline gives the disposition of blocked customers (the ones who find all station busy). For example, the system might employ the assumption that blocked customers leave the system immediately or wait for server in a queue and are served in arrival order. The arriving customers might be served in groups or one-by-one. There are number of possibilities for the order in which they enter the service.

Some examples are:

• First-come-first-served, so in arrival order what (in common use in runway utilization)

• Random order,

• Last-come-last-served,

• Priorities, like the urgent service firstly or the shortest processing time first (setting priorities on the largest aircrafts in peak periods) [Gross, Donald, Harris, 1998].

Among above characteristics of queues we can also mention some others like the behavior of the customers, service time and capacity and the waiting rooms. The behavior of the service is not very topical in given problem because a pilot of aircraft will not leave the waiting line due to impatience, which can happen in other types of queues. The waiting rooms are also not applicable in our case.

Whereas the service capacity simply refers to the number of servers, the very important characteristic is service time. Usually it is assumed that the times spent for service is independent of each other and inter-arrival times and identically distributed. The service times can be for example exponentially distributed or deterministic and depend on the length of the queue.

For the purpose of this article, the runway capacity is analyzed by the characteristic of M/M/1, M/G/1 and M/Er/1 queues. In all of the considered cases, the Poisson arrival process is applicable. It differs in terms of service process, which will be respected as Poisson, General, or Erlang distributions [Cooper, 1981].

In M/M/1 Queue, each queue could be described by the arrival rate λ (which is the number of customers arriving to the system of service in the accepted unit of time) and the service rate (known as the number of application to the system in the accepted unit of time) [Jędrzejczyk, Skrzypek, Kukuła, Walkosz, 1997]. In this model with single server, exponential inter-arrival times with the mean 1/λ (which is also the intensity of arrival) and exponential service times and the mean 1/µ (known also as service intensity will be analyzed. The customers will be served in the arrival order. The occupation rate is p is given by the formula 1:

$$p = \frac{\lambda}{\mu} \tag{1}$$

The value p is the fraction of time when the server is working. The occupation rate is considered by the three possible values. If p is equal 1, the system is able to serve one unit per one time unit and such situation causes the system during its work will serve the exact number of customers that are willing to use the station. If the p > 1 the length of the queue could explode, the queue will become longer with every arriving customer and considering infinite time of working will never be totally served. Finally, in the last case when p <1 the system in long run is stable, however some waiting lines might appear in some periods of the day depending on the density of the interarrivals. The exponential distribution gives opportunity to very simply describe the system at the time t state or the number of clients in the system (like customers which are served or the one which wait in the queue). It is neither necessary to remember when the arrival of the last customer occurred nor to register entering the service by the customer [Adan, Resing, 2002] . The other formulas considered in the model as given below:

- The mean number of the customers in the waiting line E(Lq) is given by the formula 2:

$$E(L^q) = \frac{p^2}{1-p} \tag{2}$$

- The average waiting time E (W) is represented by the formula 3:

$$E(W) = \frac{p/\mu}{1-p} \tag{3}$$

- an expected number of users in queuing system is calculated by formula 4:

$$L = \frac{\lambda}{\mu - \lambda} \tag{4}$$

- the expected time spent in the system by the one customer is calculated by formula 5:

$$W = \frac{L}{\lambda} \tag{5}$$

Formula 1,2,3,4 and 5 refers to the queue M/M/1[Cooper, 1981] can be used to present the situation on the runway in examining period of the day.

In M/G/1 Queue, it is reasonable to give some basic characteristics and differences from other types of queue. In M/G/1, the arrival of the customers happens according to the Poisson process and the rate λ. This queue is not a continuous-time Markov chain because the service time need not have the exponential distribution. Similar to the previous investigated queue, they are served in arrival order. The times of the service are independent and distributed in identical way with including distribution function FB (·) and Fb (·). For the stability the occupation rate (p= λE (B) is required to be less than one.

In this queue the state (a summary of its prior history that suffices to evaluate current and future actions [Denardo, 2002]) of the system normally consist the number of the customer in the system and the time that has elapsed since the customer who is being served entered service. As mentioned before, the arrival process is Poisson (memory less) so the state does not include the time that has elapsed since the last arrival. The service distribution is not assumed as exponential. In order to determine the distribution of the time that remains until service is complete, the state must include the time that has elapsed since service began. The formulas considered in the model as follows:

- The mean number of the customers in the waiting line E (Lq) is given by the formula 6:

E (All) = λ E (W) (6)

- The average waiting time E (W) is represented by the formula 7:

$$E(W) = \frac{\overline{\lambda}\,\frac{}{}\lambda - \sigma}{p-1} \tag{7}$$

- Formula 8 gives an expected number of users in queuing system:

$$L = \frac{p^2 + \lambda^2\sigma^2}{2(1-p)} \tag{8}$$

- And finally the expected time spent in the system by the one customer is represented by formula 9:

$$W = \frac{L}{\lambda} \tag{9}$$

Formulas 6, 7, 8 and 9 refer to the basic characteristics of M/G/1 queues [Cooper, 1981].

In M/Er/1 Queue is the last from queue types. The Erlang distribution in service station part of the system could be used to model times of the service which has low coefficient of variation(low means less than one), but it can also arise in natural way. The example might be the job, which has to be proceeding systematically, through the series of r steps that are independent where every stage takes time with exponential distribution. While analyzing the M/Er/1 queue it easy to observe that is it alike to the M/M/1 queue. The analyzing waiting line possess single server and the arrival of the clients is according to the Poisson process where the rate is λ. They are treated in arrival order and the queue has service time is Erlang-r distributed where the rate is r/μ. The occupation rate (p = λr/μ) as in the previous examples is required to be less than one [Adan, Resing, 2002].

The natural way of describing the state of the system which is nonempty is by the pair (k,l), where k indicates the number of clients in the system and l the remaining number of service phases for the client who is in service [Adan, Resing, 2002] . The formulas considered in the model as follows:

- The average waiting time E (W) is represented by the formula 10:

$$E(W) = \frac{p}{1-p}\frac{r+1}{2}\frac{1}{\mu} \tag{10}$$

- The mean number of the customers in the waiting line E (L^q) is given by the formula 11:

$$E(L^q) = \lambda E(W) \tag{11}$$

- formula 12 gives an expected number of users in queuing system:

$$L = \frac{\lambda}{\frac{r}{\mu} - p} \tag{12}$$

- And finally the expected time spent in the system by the one customer is represented by formula 13:

$$W = \frac{L}{\lambda} \tag{13}$$

Formula 10, 11, 12, and 13 refers to the characteristics of the M/Er/1 queue [Tijms, 2003]. The considering queue differs from the previous two types by the fact it allows to divide the service process on stages or steps. This characteristic is included in the simulation.

Simulation

The simulation, which will be used for the above stated purpose, is limited to discrete event digital simulation. Digital simulation means that the simulation occurs entirely inside a digital computer, and discrete event means that the studying system can be viewed intermittently, not continuously.

A discrete event digital simulation can emulate systems that evolve in continuous period and the laws of this evolution are governed by discrete events. In addition, queuing systems illustrates this point well. A queue evolves in continuous time, but it can be described in terms of a sequence of customer arrivals, service initiations, and service completions. The simulation of the runway queue keeps all of those features that why it is reasonable. The simulation uses the generator of pseudo-random number what makes it stochastic model. In its result, it will produce the set of data what is typical characteristic of static simulation. Moreover, only one computer is used to perform the simulation so the model will be local.

A simulation imitates the behavior of a system, but it does not show how the system should behave. It neither tells which settings of the system's parameters cause the best performances. It is important to remember that in nearly every digital simulation the central role is played by the uncertainty. Sometimes, when a simulation samples uncertain quantities, its result produces the behavior of a system in a particular instance, what in consequence delivers the situation in which simulation has to be run many times in order to get the average of the system behaviors. In the same case, it is required to perform repeatable simulation and to experiment with a variety of parameter settings in order to verify how the system reacts to different settings of its parameters [Denardo, 2002].

There are three areas of computer simulation, which are interconnected on each level of simulation: model design, model execution, and model analysis [Denardo, 2002]. The simulation model presented by this work is a simple approach to the problem of optimizing the runway at the airport and slot management. It is designed to reproduce a phenomena or its behavior by creating a model, which comprises dependencies that occur in it.

Implementation environment

This topic focuses on the implementation environment, which is the VBA language in Excel environment. Visual Basic for Applications (VBA) is based on Visual Basic (VB) programming language, implemented in Microsoft Office applications and several others, like popular in designer and architecture' environments AutoCAD. This simplified version of Visual Basic is used primarily for automation of work with documents, for example through macros [Walkenbach, 2002]. A macro is a set of commands that can be run with just one click or press of the button. Macros can automate almost anything that can be done in preparing program or spreadsheet and sometimes even allow increasing the usability of it by fastening the basic work with the program and make some calculation more automatic. Macros are a kind of programming, but the applied in the programming language is logic and by the numeric specialist considered as one of the simplest. Often there is not even required to be a professional programmer to successfully create or use them. Most of the macros that can be created in Office programs are written in Microsoft Visual Basic for Applications, usually abbreviated VBA. Macros save time and extend the programs usability during every day work with it. They could be to automate repetitive tasks during creating documents, simplify hard tasks, or create solutions, for example, to automate document creation in often-used formula. The VBA language allows using macros to create custom add-ons, such as templates, dialog boxes, and even to store frequently used information. There is no debate that the macros are mostly harmless and helpful, they are however an important issue from the perspective of system security. For the purpose of the simulation models explained in this article, the macros are mainly present the results of the simulation in a better way and give the user who browse those results the opportunity to choose the model of the simulation.

Simulation structure and Decisional situation

Using the capabilities of the program MS Excel and standing behind it language Visual Basic for Application, we can perform a simulation of the problem of runway capacity. Then a user interface can be built and the application stands behind it, which allows examining the utilization of airport runway during a given the period of the day. The Chapter's structure starts from the defining the characteristic of the problem scope and the processes of decision making in the investigating situations. Firstly, the presentation of the decision problems of the subject and the process of simulation and building the application is to be done.

For the purpose of the simulation, any airport, which fits the given criteria can be chosen. Firstly, it is necessary to perform a simulation on the airport with one runway, because it is simplest runway configuration and acts as a single server model. More than one runway at an airport builds a great complexity. In reality, it is usually dynamic decision of Air Traffic Control Tower, which allocates the runway for aircraft landing, or aircraft take off. For testing the model it is advisable to select the aerodrome that has the busiest one-runway system and possesses capacity comparable with some two-runway equivalents. This will make sure that if the model provides exact results for a complicated single runway system airport then the final application will be able to examine any single-runway airport in world.

Next step is to choose the period to examine. The selection of the appropriate period directly depends on the historical data taken into account. As it was mentioned before, the slots are allocating for the particular flight on the seasonal basis, which implies that for half a year the flight schedules for peculiar day of the week will reiterate for the whole six-month season. For the purpose of the simulation, the time period can selected randomly, for example selected period can be between 7.00 pm and 9.00 pm. To select time period, which will adequately test the creditability of the model, it should be kept in mind that the time period with maximum number of movements can be taken into account. The created application will give an opportunity to its user to unrestrictedly choosing the period of the day and the study over it, what is also helpful is that in examining the periods of the day hour after hour and choosing the peak periods in simpler way will make the analysis and optimization clear.

After characterizing the scope of the problem and the decision taken to the simulation process, next step is launching the model. In the next sub section, the exact process of launching the model is presented.

Model Launching

This sub-section provides an important insight in launching the model. First step of launching the model is the selection of the appropriate simulation software, which is a computer program imitating a real life process using a set of mathematical formulas. In this case, as mentioned in the simulation environment, MS Excel can be used as it possess all the required tools and methods to perform the simulation. The simulation can be performed easily on the computer with the parameters appropriate to perform the simulation.

The steps for launching the model i.e. overall simulation is provided below. These steps also provide answer to the various issues defined previously.

Step 1: Firstly, it is reasonable to study the whole normal day of the operations on the runway and pick up the busiest period. It will simply identify the hour, when the number of movements is the largest or when the airport suffer due to the incapacity. Possessing such knowledge, it will be possible to advice how to unburden overloaded period by dynamic slot management that is normally conducted by the Air Traffic Controllers. Due to that need, the first part of the simulation model will focus on identifying the busiest period or using the appropriate nomenclature – peak period during the everyday operation. For that goal, the model of queue M/M/1 seems to be the most appropriate in its readability and that is why this part of work will use it.

Step 2: As the main goal of the simulation is to examine, using Queuing Theory analysis, whether the runway of an airport creates a waiting lines or the aircrafts landing on or taking-off from that runway make their movements without any delay. The three way approach already mentioned previously is applied to the given research problem due to the examination and consideration of the runway utilization from the point of view of three different queues – M/M/1, M/G/1 and M/Er/1. Each of them will produce different results; however, the major trends and indicated problems with capacity are expected to be similar.

Step 3. Each queue will be investigated in the long term. This can be done more precisely by analyzing the crucial parameters obtained from the simulation. The important parameter will be copied to the table where the further investigation can be conducted. This step also uses the model for clarifying if the queue is stable in the longer run. Testing if queue is stable in the longer run implies the number of attempts or the observations of inter-arrivals customers to the queuing system. This particular clarification can be conducted for all of the queuing models chosen to simulate the operations at the airport runway.

Step 4: After step 3, we can compare the results that are obtained by three different probability distributions in service station– Poisson process, General distribution, and Erlang distribution. It is also reasonable to compare the results of mean inter-arrival and service times got form the simulation.

Step 5: this particular step will test if the system of slots are planned in optimum way and if particular flights catch their slot or they more often miss them causing additional delays and costs for the both airlines and airports. The simulation counts up the aircrafts that have missed their slots what is another factor indicating if the single-runway system works in optimal way and is not queue creating. The missed slots are counted in the examined period of the day and both operational and economic efficiency are derived.

Step 6: during this step, we can estimate the economical aspect of the runway utilization and capacity. The major problem of the airlines, beside the insufficient number of slots that the airport offers in the most attractive periods of the day, is additional costs caused by standing in queue or missing the slot. The economic parameter, which can be estimated here, is the cost of fuel consumption considered by the type of movement and extra costs of missing the slot that will be summed up and can be presented in the results section.

Step 7. This step is the decision making step. From the knowledge attained from the previous steps of simulation, the airport managers can analyze their airport runway and based on the results judge if it is used in optimal way or some upgrades and changes are sought to be made.

Hypothesis Testing

On based of the results, several hypothesis can be tested . The following hypothesis can be investigated:

Hypothesis 1: Occurrence of the busiest period or the peak period

Hypothesis 2: Airport capacity's situation for stability and profitability.

Hypothesis 3: The stability of the situation on the researching airport.

Hypothesis 4: The influence of the number of events on the parameters used in main simulation.

Hypothesis 5: fixed maximum number of aircrafts that missed the slot in a particular investigating period.

Hypothesis 6: The airlines at the airport do not suffer due to additional costs caused by missing the slot or extra fuel consumption.

Conclusion

The simulation model presented in this article has a practical application. It can be used to analyze the various operational aspects of the runway system. The final simulation will equip the manager will a powerful tool to access the efficiency of the runway while keeping in view both operational and economic aspects of the business. Various conclusions which an airport manager can draw from the simulation are closely correlated with the hypothesis mentioned in the above section. In general, the simulation model developed in this article will provide the information related to various factors mentioned below:

- the air traffic behavior and customer segmentation from the peak periods;
- the bank structure at the airport;
- overall air side capacity utilization and its overall profitability;
- number of aircrafts waiting in the queue at a particular instant of time;
- additional costs which can be incurred by the airlines;
- and lastly the efficiency of various strategies incorporated to improve the air side operations.

This type of modeling will be used by all the airport in order to minimize the losses due to congestion of air side. The model is lost cost and can be initiated and used easy while providing the exact results.

Bibliography

[Doganis, 2001] Doganis R., The Airline Business in the 21st Century, Routledge, USA, 2001,

[Doganis, 1991] Doganis R., Flying Off Course: The Economics of International Airlines, Harper Collins Academic, USA, 1991

[Official Journal L 138, 2004] Regulation (EC) No 793/2004 of the European Parliament and of the Council of 21 April 2004 amending Council Regulation (EEC) No 95/93 on common rules for the allocation of slots at Community airports, Strasbourg, 2004

[Simpson, Belobaba, 1992] Simpson R., Belobaba P., The Demand for Air Transportation Services, Air Transport Economics, MIT, Cambridge, 1992

[Madas, 2007] Madas, M., A Critical Assessment of Airport Demand Management. Ph.D. thesis, Athens University of Economics and Business, Athens, 2007

[ACI, 2007] Airports Council International Worldwide Airport Traffic Statistics, Canada, 2008

[Bose, 2002] Bose S.J., An Introduction to Queuing Systems, Plenum Publishers, New York, 2002.

[Gajowniczek, 2008] Gajowniczek P. : Teoria kolejek. Instytut Telekomunikacji Politechniki Warszawskiej,Warsaw, 2008.

[Cooper, 1981] Cooper R.B., Introduction to Queuing Theory, New York, North Holland (Elsevier), 1981

[Gross, Donald, Harris, 1998] Gross M., Donald R., Harris C.M., Fundamentals of Queueing Theory. Wiley, USA, 1998 [Jędrzejczyk, Skrzypek, Kukuła, Walkosz, 1997] Jędrzejczyk Z., Skrzypek J., Kukuła K., Walkosz A. : Badania operacyjne w przykładach i zadaniach. PWN, Warsaw, 1997.

[Adan, Resing, 2002] Adan I., Resing J., Queuing Theory. Eindhoven University internal paper, Eindhoven, 2002

[Tijms, 2003] Tijms H.C, Algorithmic Analysis of Queues. Wiley, USA, 2003.

[Denardo, 2002] Denardo E.V., The Science of Decision Making: A Problem-Based Approach Using Excel, John Wiley & Sons Inc., USA, 2002.

[Walkenbach, 2002] Walkenbach J., Excel 2003 Biblia. Willey, Helion, 2003

[Walkenbach, 2003] Walkenbach J., Excel 2003 – programowanie w VBA. Willey, Helion, 2002.

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*

Major Fields of Scientific Research: Transport Economics, Operational research, Knowledge management, Econometric models in various sectors of transportation Industry, Air Transportation Knowledge Hub, Decision support systems and expert systems in various fields of aviation industry

Michał Nędza, M.Phil. –University of Information Technology and Management in Rzeszow, ul. Sucharskiego 2, 35-225, Rzeszow, Poland. ; e-mail: *mnedza@wsiz.rzeszow.pl*

Major Fields of Scientific Research: Operational Research, Application of optimization methods in airport and airline management, IT in Econometrics Models & Management Systems, Customer Relationship Management, Data Mining and Data Warehousing

EVALUATION OF RUNWAY CAPACITY AND SLOTS AT LONDON GATWICK AIRPORT USING QUEUING BASED SIMULATION

Sumeer Chakuu, Michał Nędza

***Abstract**: The evaluation of the runway capacity and its optimization is one of the core goals of the airports. Most of the time due to infrastructural and regulatory factors it is quite impossible to increase the capacity of the runway. Therefore, it is of utmost priority to optimize its usage. Nowadays, the decision support systems play a very crucial role in defining the threshold capacities at the runway to make it economical and operationally efficient. This fact makes them a crucial factor in the aviation market. The use of the support system for runway evaluation and assessing slots makes the business profitable for both airports and airlines, as they will highly get hurt economically if they do not use the runway as an airside infrastructure efficiently. The main aim of this article revolves around the design of a decision support system, which will help in providing the decisional support to the managers by evaluating the various scenarios for optimization of the runway usage. The evaluation models, which are used in this article, are the queuing based models and they accurately cope with the logic lying behind the runway capacity usage.*

***Keywords**: Runway Capacity analysis, Slot Management, Queuing Theory and Models, Probabilistic Distribution, Simulation and Optimization*

***ACM Classification Keywords**: B.2.2 Performance Analysis and Design Aids Simulation, B.4.4 Performance Analysis and Design Aids, F.1.2 Modes of Computation, G.3 Probability and Statistics, G.m Miscellaneous. .*

Introduction

The main aim of this article is to examine and evaluate the standard day of operation at the London Gatwick airport. It is important to evaluate the runway capacity as it provides the insight into the number of movements served at the airports [Simpson, Belobaba, 1992], there by directly complying with the slot management. The airport chosen for the evaluation is not by accident – this is the most overloaded single runway airport on the globe. In the process of this evaluation the models of Queuing Theory is used and a special system has been developed and implemented. The queuing theory is also referred as the traffic theory because of the characteristics it possess [Bose, 2002]. The simulation has been applied with adoption of three different models of queue. For this article, only two models will be elaborated. Each model is characterized by the different probability distributions of the time depending on the runway occupancy.

After the simulation models is planned and implemented, it is quite reasonable to form a particular hypothesis that will be examined during the research and will help in final fulfillment of the main goal. The following hypothesis will be investigated:

Hypothesis 1: The busiest period or the peak period takes place in the morning and in the early afternoon what is a consequence of business travel.

Hypothesis 2: The London Gatwick Airport capacity's situation, despite of the fact that it is the busiest single-runway airport, is stable and the probability that the aircraft misses its slot is less than 10%.

Hypothesis 3: Despite of the fact the situation on the researching airport is stable, the small waiting lines might occur. However, the utilization of the runway is optimal and number of aircraft waiting in the queue at a particular moment is smaller than 3 and the probability of that event not taking place is smaller than 20% during the whole period.

Hypothesis 4: The results (L, Lq, W, and Wq) from the main simulation will be very similar to the second long run simulation with more number of observations. The number of events do not influence on the mentioned parameters.

Hypothesis 5: Sum-up of the all aircrafts that missed the slot in investigating period will not be greater than 2 missed slots per hour.

Hypothesis 6: The airlines on the Gatwick airport do not suffer due to additional costs caused by missing the slot or extra fuel consumption.

Simulation

The aim of this section is to show the interface and other design aspects of the simulation.. Simulation directly provides reasonable improvement in the application of market mechanisms [Doganis, 1991].

The simulator is named as Runway Examiner, which goes precisely with the task it is accomplishing. All of the steps and actions that occur during the interaction are described here. The simulator uses a detailed scenario that indicates in exact way how the customer works on it. The basic strategy is to identify a so called path through the user case and then to write an exemplary scenario. All the simulations have the generalized characterisctics of having an input process, the service mechanism and queue discipline [Cooper, 1981]. Figure 1 shows the runway examiner interface design to evaluate runways.

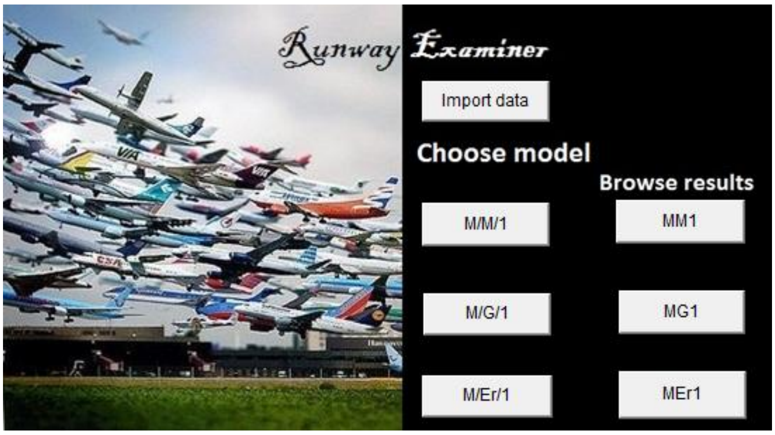
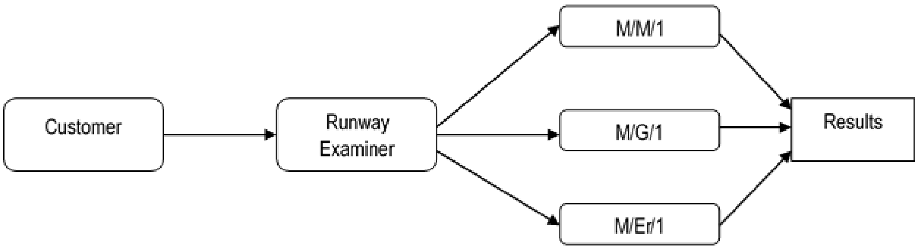


Fig. 1 Runway Examiner interface

The graphical presentation of simulator usage is provide in schema 1.



Schema 1 Simulator Usage

The Runway Examiner for the random customer works as follows:

1. The customer runs the file Runway Examinet.xls and the control panel presented above appears.
2. The user clicks on the button import the data in order to load the desired information about the operation that is scheduled during particular day and time. The imported file has to be a *.txt type and has to be prepared earlier by the form builder.
3. The next step is choosing the model of queue the customer wants to examine desired airport runway for. The choice has to be made between three considered types of queues: M/M/1, M/G/1, and M/Er/1. After moving the mouse over the button, the short comment including some basic information about the mathematical model of examining the waiting lines is printed.
4. The final step is to press the button Results in order to get the findings of the most important characteristic of the airport runway and browse the figure section.
5. Alternative way is to click the button Browse and observe the results straight from the table printed in the spreadsheet.

Setting the priorities is also an important aspect of the simulation which allows to allocate the appropriate service time [Gross, Donald, Harris, 1998].

Simulation results

This section highlights the results of the simulation. Though we can perform all the three types of simulations using the runway examiner, during this article only results of two simulations is discussed. The simulations, which are discussed in this article, are, namely, M/M/1 queue simulation and M/Er/1 queue simulation.

M/M/1 queue simulation:

The results of this simulation are shown in the table 1:

Table 1. Characteristics of M/M/1 queue simulation

Characteristic	Value
Total movements in examining period =	76
Arrival rate λ =	0.42
Service rate µ =	0.44
Occupational rate p=	0.95
Number of airplanes by weight class(light;heavy;massive) =	(6,47,23)
Number of movements by the type of movement(land;take-off)	(40,36)
Expected number of users in Queueing system L =	5
Expected time in Queueing system per user W =	12.16
Expected number of users in queue L_q=	2
Expected waiting time in queue per user W_q=	1.01

The formulas used to calculate the parameters are taken from M/M/1 queue simulation [Denardo, 2002].The total number of movements have not exceeded the 80 that is the maximum possible size of traffic that is allowed by the airport authorities and international regulators at the London Gatwick Airport. That means that, at least theoretically, the Air Traffic Controllers should be able to handle the number the movements that appeared in examining hour. Each queue is described by the arrival rate λ and the service rate μ [Jędrzejczyk, Skrzypek, Kukuła, Walkosz, 1997]. The arrival rate λ equals 0.4, whereas the service rate μ 0.45. That means that less that one aircraft appears on the runway every two minutes and respectively roughly 2 minutes is enough for the service station for providing service. The very important characteristic - the occupational rate is 0.89, inhibits that the queuing system in long run is stable, however some waiting lines might appear in some periods of the day depending on the density of the inter-arrivals. Such situation will be examined during the hypothesizes in later section.

Now, there is a high time to consider the profile of the customers (which are aircrafts in our case) by the weight class and the type of movement. The great majority of the runway system users, taking into consideration the historical data, are the heavy class aircrafts. Completing the profile – less than 10% of total number of aircrafts are light aircrafts flying on the regional lines mainly. The distribution of the traffic by the type of represents equilibrium. Almost the same number of planes land and start their journey at the London Gatwick.

The next step was to, using Queuing Theory formulas to get the expected number of users in queuing system L, expected time in queuing system per user W, expected number of users in queue (Average number of airplanes in the queue) Lq and expected waiting time in queue per user Wq. The values shown in the table indicate that the queuing system is rather stable. The average number of clients in the system is 6. That number may seem high, but it should be kept in mind that some flights have been scheduled at the same time, which is why the short queue may occur, (only one aircraft on average is expected to stay in waiting line). The total time is very likely resulting from this fact. The average number spend in queue per user is equal 6.73.

Finally, the distribution of service and arrival time per user is provided by the service time and it balances between 1 and 3 minutes, almost 98% of all examining movements are in this range. Considering the inter-arrival time of the aircraft it is between one and 5 minutes.

M/Er/1 queue simulation:

The results of this simulation are shown in the table 2:

Table 2. Characteristics of M/Er/1 queue simulation

Characteristic	Value
Total movements in examining period =	76
Arrival rate λ =	0.39
Service rate μ =	0.43
Occupational rate p=	0.91
Number of airplanes by weight class(light;heavy;massive) =	(6,47,23)
Number of movements by the type of movement(land;take-off)	(40,36)

Expected number of users in queuing system L =	5
Expected time in queuing system per user W =	12.16
Expected number of users in queue L_q=	2
Expected waiting time in queue per user W_q=	1.01

The formulas used to calculate the parameters are taken from M/Er/1 queue simulation [Tijms, 2003].The values that are different form the first sight are arrival rate λ and service rate μ. The distinction between them is the same as in M/M/1; however, their proportion, which is also the occupational rate, is the lowest from all the models. The expected number of users in queuing system L is equaled to 5, the value of expected number of users in queue – Lq is equaled to 2, what in the case of investigation insinuate that the system will face greater problems with the queues that form. Generally, the results are pretty close that might indicates congenital distribution of time General and Erlang. The time each aircraft on average spent in queue is around one minute and in system 12 minutes. Considering the arrival distribution of time, it is similar as in model M/M/1 – the distribution time in both cases is in Poisson process and it has been normal that they will differs only slightly. The occupation rate ($p = \lambda r/\mu$) which is required to be less than one [Adan, Resing, 2002] is also up to the mark. The more detailed interpretation of the results characterizing this model is presented with particular hypothesizes.

Hypothesizes testing

This section will analyze the hypothesizes defined at the beginning of the article. Hypothesizes provide more comprehensive treatment to increase the optimality of the results [Lehmann, Erich L., Romano, Joseph P.,2005]. The results are presented in the form of the table. After that, each outcome is interpreted in harmony with the mathematical and statistical formulas. Though we can test all the hypothesis based on the result, in this article on hypothesis 1,2,3 and 5 is tested.

Hypothesis 1 – Peak period

The first hypothesis has opened the issue of choosing peak period, because according to the literature this is the time when it is the most probable that the airport will be congested. The congestion will automatically create a waiting line that disturbs the flight schedule plan, often for many hours.

The most logical way of defending such a sentence is to take one randomly chosen day of the airport operation and investigate it hour by hour by the known methods. It is important to mention that the British authorities and international aviation institutions allow the airport due to its location to operate during the nighttime; however, the operations between midnight and 6.00 am are limited to 25 movements. In the regular hour of the operation, the airport is allowed to serve 40 arrivals or departures on its single-runway. The results of the event is presented below in table 3:

Table 3. Peak period investigation

Period	No. of movements	Possible movements	Landings	Taking-off	% of usage
6.00 - 7.00	35	40	14	21	88%
7.00 - 8.00	38	40	16	22	95%
8.00 - 9.00	38	40	18	20	95%
9.00 - 10.00	36	40	16	20	90%
10.00 - 11.00	38	40	14	24	95%
11.00 - 12.00	38	40	15	23	95%
12.00 - 13.00	35	40	15	20	88%
13.00 - 14.00	38	40	14	24	95%
14.00 - 15.00	37	40	16	21	93%
15.00 - 16.00	38	40	20	18	95%
16.00 - 17.00	38	40	16	22	95%
17.00 - 18.00	33	40	12	21	83%
18.00 - 19.00	30	40	18	12	75%
19.00 - 20.00	36	40	18	18	90%
20.00 - 21.00	40	40	22	18	100%
21.00 - 22.00	24	40	13	11	60%
22.00 - 23.00	15	40	7	8	38%
23.00 - 24.00	16	40	8	8	40%
24.00 - 1.00	18	25	6	12	72%
1.00 - 2.00	12	25	5	7	48%
2.00 - 3.00	15	25	4	11	60%
3.00 - 4.00	14	25	5	9	56%
4.00 - 5.00	16	25	8	8	64%
5.00 - 6.00	24	25	9	15	96%

The above table unambiguously shows that the distribution of the movements at the London Gatwick airport during its everyday operation. It indicates the total number of movements each hour and the contribution of arrivals and departures to that number. Additionally, there is a column showing the total allowed number of movement per hour and the percentage of its utilization by scheduled movements.

From the analysis it is quite clear to observe that the periods indicated in the hypothesis are one of the busiest, however the higher number of movements occurs between 7 pm and 8 pm. The first hypothesis was not completely correct so its status become disapproved.

Hypothesis 2 – probability of missing the slot

The second hypothesis highlights the problem of missing the assigned slots. The concept of this hypothesis has an operating approach. The exact formulation of the hypothesis is that the London Gatwick Airport capacity's situation, despite of the fact it is the busiest single-runway airport, is stable and the probability that the aircraft misses its slot is less than 10%.

For defending this hypothesis, the research outcome is presented below in the table 4.

Table 4. Percentage of missed slots by the models

Attempt	M/M/1 model	M/Er/1 model	Attempt	M/M/1 model	M/Er/1 model
1	7.5%	7.3%	26	13.5%	0.0%
2	9.7%	5.2%	27	14.2%	6.1%
3	1.4%	8.4%	28	2.6%	1.8%
4	11.0%	4.8%	29	9.2%	7.0%
5	6.1%	12.0%	30	12.3%	0.4%
6	14.2%	12.4%	31	2.1%	2.1%
7	7.4%	2.8%	32	3.3%	0.3%
8	3.3%	6.6%	33	2.2%	4.6%
9	9.1%	0.9%	34	13.8%	2.5%
10	13.6%	12.0%	35	0.5%	1.0%
11	11.3%	1.8%	36	14.8%	4.3%
12	7.8%	3.0%	37	8.0%	3.4%
13	7.0%	6.2%	38	9.9%	2.2%
14	3.0%	4.3%	39	3.6%	0.9%
15	14.8%	0.9%	40	13.1%	1.2%
16	9.1%	11.8%	41	4.6%	0.6%
17	0.9%	3.5%	42	9.4%	7.4%
18	14.8%	12.7%	43	8.9%	10.3%

19	0.0%	1.4%	44	11.2%	1.1%
20	8.0%	9.1%	45	0.6%	2.4%
21	11.7%	12.1%	46	2.4%	7.3%
22	7.4%	3.9%	47	1.5%	2.9%
23	2.1%	0.5%	48	5.1%	0.8%
24	14.8%	11.3%	49	10.2%	10.0%
25	8.3%	3.2%	50	4.2%	9.0%
Mean	8.2%	6.3%		7.2%	3.6%

The table compares the probabilities of missing the slot, as different models were considered; the different probability distributions of service time are taken into account. The results differ while taking into consideration each model. 10% is the threshold that should not be exceeded in any model; this will indicate the overall stable situation in the investigating period.

Based on the conducted research and obtained findings, it can be claimed that the raised hypothesis is correct. The London Gatwick, despite of the fact it is the busiest single runway airport, represents the stability of the operations in examined period. The number of aircrafts missing their slots is in each investigated model is smaller than 10%

Hypothesis 3 – likelihood of queue occurrence

The hypothesis number three actually has been answered by the data collected for the purpose of previous one. Despite of the fact the situation on the researching airport is stable, the small waiting lines might occur. The second part of the hypothesis gave specific numbers describing the queue and for those objectives the models has been tested. The second part of raised hypothesis standpoints "the utilization of the runway is optimal and number of aircrafts waiting in the queue at a particular moment is smaller than 3 and the probability of that events' absence is smaller than 20% during the whole period". To defend this statement the formulas from the Queuing Theory are quite adequate. Those formulas were Q, which calculates the number of aircrafts on average waiting in queue, and P (n) which as a result will give the probability that more than 3 aircrafts waits for the runway. The results of this analysis is provided in table 5. The calculations from the peak period have been reproduced, as they are based on the results that consists randomized number in itself. The standard deviation has had a small positive value that is the reason that the trail of 20 attempts is enough to perform this research. The table 5 is shown below:

Table 5. Results on Q and P (n=3) by the models

Attempt	Formula	M/M/1 model	M/Er/1 model	Attempt	Formula	M/M/1 model	M/Er/1 model
1	Q	0	0	11	Q	1	2
	P(n=3)	0.13	0.08		P(n=3)	0.06	0.18
2	Q	2	1	12	Q	1	1
	P(n=3)	0.19	0.20		P(n=3)	0.24	0.19

3	Q	3	1	13	Q	3	2
	P(n=3)	0.22	0.22		P(n=3)	0.13	0.10
4	Q	3	3	14	Q	1	1
	P(n=3)	0.14	0.19		P(n=3)	0.17	0.18
5	Q	1	2	15	Q	3	0
	P(n=3)	0.14	0.02		P(n=3)	0.03	0.23
6	Q	3	1	16	Q	0	1
	P(n=3)	0.16	0.15		P(n=3)	0.11	0.11
7	Q	3	0	17	Q	2	3
	P(n=3)	0.01	0.01		P(n=3)	0.15	0.20
8	Q	0	3	18	Q	0	3
	P(n=3)	0.11	0.00		P(n=3)	0.08	0.22
9	Q	1	0	19	Q	1	0
	P(n=3)	0.22	0.08		P(n=3)	0.14	0.19
10	Q	1	2	20	Q	0	1
	P(n=3)	0.03	0.01		P(n=3)	0.13	0.20
Average	Q	2	1	Average	Q	1	1
	P(n=3)	0.13	0.10		P(n=3)	0.12	0.18

The table calculates the quantity of the aircraft waiting on average in the line and it is rounded to the nearest integer. The second value in the each attempt calculates the probability that the number of aircraft waits in forming queue is greater than 3. Both formulas provide an appropriate view to check if the raised hypothesis has been proved or disproved.

The summary of above outcomes gives a clear answer for the raised hypothesis. The number of the aircraft waiting on average in forming waiting line in each model is lower than 3. The average from the M/M/1 model is equal to 2 aircrafts, whereas in the model with Erlang distribution model is even lower and just one airplane on average has to wait for its access to the runway. Only in 11 attempts the number of investigating flights has equaled 3 and there has been no observation of the number greater than 3

Hypothesis 5 – slots missed in total

The fifth hypothesis highlights the similar topic as a second one – the missed slots. The defending however takes a different approach – it sums-up the total number of aircrafts that missed the slot by the column of leaving time. It counts the time of leaving by summing the service time with the time of leaving of the predecessor. After that the next column compares that time with the slot range and prints the information “ok” for hit or “not ok” for missed one. The experiment counts and sum up the cells with the string “not ok”. The trail of 40 attempts is sufficient to conduct the test. After the test, it will be possible to compare the results with those attained from the second hypothesis. The formulated hypothesis is as, “Sum-up of the all aircrafts that missed the slot in investigating period will not be greater than 2 missed slots per hour” The table 6 presents the findings of the research.

Table 6. Number of missed slots by model

Attempt	M/M/1 model	M/Er/1 model	Attempt	M/M/1 model	M/Er/1 model
1	2	0	21	1	2
2	1	1	22	0	2
3	3	3	23	2	3
4	1	0	24	3	2
5	2	3	25	0	0
6	0	1	26	2	1
7	2	3	27	3	3
8	2	3	28	1	3
9	3	0	29	2	2
10	2	2	30	1	3
11	2	3	31	2	5
12	1	1	32	0	4
13	1	4	33	1	1
14	3	4	34	1	2
15	0	3	35	2	4
16	2	3	36	0	3
17	1	2	37	2	3
18	1	3	38	1	3
19	3	1	39	2	4
20	3	1	40	0	2
Mean/h	2	2		1	3
Percent	4.2%	3.8%		3.2%	6.3%
Hypothesis 2	4.9 %	7.7 %		4.9 %	7.7 %

The investigating period considered here is two hours. From the mathematical point of view, the results from the simulation and chosen simulation models show unambiguously that the London Gatwick Airport deals with the runway operations in satisfactory way. The aircraft appearing on the runway in the great majority catch the slots and even if they have to wait, the waiting time is not very long. The outcomes are rounded to the nearest integer. The results in M/M/1 presents the range between 0 and 3 that gives the mean 2 missed slots per one hour of operation during peak period. In reality, such a score is considered close to perfect and highlights the good runway organization at London Gatwick Airport. The model with the general distribution of service time has a wider range, affecting the mean – the number of aircrafts missing the slot in 40 attempts during the period of two hours is equal to 3. The last model – M/Er/1 range is from 0 and 5 missed slots that gives a mean 3 in 40 attempts. The table additionally includes the percentage of the number of missed slots to the total number of slots and compares the results from the second hypothesis. The results are very close and the trend attained from the Hypothesis 2 results' is maintained. Backing the hypothesis, in some attempts, the number of missed slots has been greater than 2 but on average in two models the statement is proved.

Conclusion

In relation to the conducted research, the following conclusions are formulated:

- The Queuing theory has its application in runway investigation and simulation concept.
- The busiest period at the London Gatwick, while examining the normal day of operation, occurs between 7 pm and 9 pm. It is not the expected peak period that is formed in the hypothesis, which was based on business traffic and nominated around 8 am and 4 pm.
- The slot situation on the London Gatwick airport is stable even in the peak period. The authorities do not exceeds the regulated number of hourly slots and this number is sufficient to face the demand.
- Despite of the fact that situation is stable small waiting lines have occurred in the simulation. However, the number of aircrafts staying in the queue at a particular moment has been lower than 3 and the probability that it will be greater was less than 15%.
- The results from the simulations in short or long run do not possess large differences. All of examining parameters have acted similar way with no heeding to number of observations.
- The number of aircrafts that miss the slot every hour, accordingly to the simulation, is relatively low and do not affect the stability of operations on the runway.
- The airlines using London Gatwick Airport for their operations, do not suffer a significant financial penalties from delays in the operation.
- Comparing all models, if the arrival rate λ and the service rate μ are constant the Markovian distribution of time gives the largest values for examining total time and number of customers in queue and in the whole system.
- The distribution of movements on investigating airport is balanced by the type of movement, which is characterized by the majority of heavy aircrafts and little number of light planes, while considering the weight classes.
- Visual Basic Applications for this particular simulation has been found as simple, user friendly and sufficient programming language for building the user interface for the purpose of presenting the results of the research.

Bibliography

[Simpson, Belobaba, 1992] Simpson R., Belobaba P., The Demand for Air Transportation Services, Air Transport Economics, MIT, Cambridge, 1992

[Bose, 2002] Bose S.J., An Introduction to Queuing Systems, Plenum Publishers, New York, 2002.

[Doganis, 1991] Doganis R., Flying Off Course: The Economics of International Airlines, Harper Collins Academic, USA, 1991

[Cooper, 1981] Cooper R.B., Introduction to Queuing Theory, New York, North Holland (Elsevier), 1981

[Gross, Donald, Harris, 1998] Gross M., Donald R., Harris C.M., Fundamentals of Queueing Theory. Wiley, USA, 1998 [Denardo, 2002] Denardo E.V., The Science of Decision Making: A Problem-Based Approach Using Excel, John Wiley & Sons Inc., USA, 2002.

[Jędrzejczyk, Skrzypek, Kukuła, Walkosz, 1997] Jędrzejczyk Z., Skrzypek J., Kukuła K., Walkosz A. : Badania operacyjne w przykładach i zadaniach. PWN, Warsaw, 1997.

[Tijms, 2003] Tijms H.C, Algorithmic Analysis of Queues. Wiley, USA, 2003.

[Adan, Resing, 2002] Adan I., Resing J., Queuing Theory. Eindhoven University internal paper, Eindhoven, 2002

[Lehmann, Erich, Romano, Joseph, 2005] Lehmann, Erich L., Romano, Joseph P., Testing Statistical Hypotheses, Springer, USA, 2005

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

Major Fields of Scientific Research: Transport Economics, Operational research, Knowledge management, Econometric models in various sectors of transportation Industry, Air Transportation Knowledge Hub, Decision support systems and expert systems in various fields of aviation industry

Michał Nędza, M.Phil. –University of Information Technology and Management in Rzeszow, ul. Sucharskiego 2, 35-225, Rzeszow, Poland. ; e-mail: mnedza@wsiz.rzeszow.pl

Major Fields of Scientific Research: Operational Research, Application of optimization methods in airport and airline management, IT in Econometrics Models & Management Systems, Customer Relationship Management, Data Mining and Data Warehousing

THE FAST FOURIER TRANSFORM AND CEPSTROGRAM BASED APPROACH TO THE ASSESSMENT OF HUMAN VOICE STABILITY

Krzesimowski Damian

Abstract*: The topic of the paper is to assess the stability of the human voice on the basic of results of Fast Fourier Transform and cepstrogram, which are subjected to statistical analysis. The presented results are the first part of the study on the usefulness of these analyses in voice quality assessment and identification of persons and voice commands in a noisy environment. The purpose of the study is to select the indicators, characteristic for the voice of the one particular person. In the paper is described the data selection algorithm for testing purposes from a single recording of a human voice, and the results of statistical analysis of the stability of expression of voiced vowel. Proposed algorithm allows extracting voiced elements of a desired length from the recordings with the exception of noise and silence. Then, to assess stability of waveforms, the recording is divided into several to several tens of fragments a length of few tens milliseconds. Each fragment is analysed independently of the other, and the result is a measure of the error of inference algorithms for identifying the person and voice commands. Particular attention is given to comparing the results of statistical analyses, after the splitting into blocks of the same duration and the same number of micro phonemes in the waveform. Due to fact, that in the studies have used a frequency analysis, it is possible to determine the stability of both the fundamental frequency and formants using the same statistic apparatus.*

Keywords*: voice, FFT, cepstrogram, data selection, signal processing.*

ACM Classification Keywords*: F.2.1 Analysis of Algorithms and Problem Complexity – Numerical Algorithms and Problems – Computation of transforms, G.3 Mathematics of Computing – Probability and Statistics – Statistical computing*

Introduction

The topic of the presented research is the usefulness of Fast Fourier Transform and cepstrogram for assessing voice quality of a healthy person, and identify the persons and voice commands. The purpose is to define the vectors, which are characteristic for the individual human voice using signal analysis and statistics. Research should be used to identify people, voice commands and voice quality assessment such people as opera singers or speakers. It will be possible to apply the developed algorithms in speech therapy and the voice training for speakers or singers. The basis for the research will be own database of recordings at least 100 persons of both sexes at different ages.

The survey plan is established as follows:

1. development of algorithms for the selection of voiced recordings of the human voice with the elimination of noise components;
2. execution of Fast Fourier Transform, spectrographic and cepstrographic analyses on selected parts of recordings;
3. statistical study of the results of the mentioned above numerical analysis for a single recorded signal divided into sections of predetermined length;

4. selecting from the resulting figures coefficients which for one recording (one person) will not deviate beyond a preset threshold error;
5. comparison of numerical results with the results for step 4 for at least 100 people for verification of found coefficients.

In the paper is presented:

1. data selection algorithm for testing from a single recordings of a human voice;
2. Fourier analysis, spectrographic and cepstrographic for one recording;
3. statistical analysis of the mentioned above numeric result studies for one recording.

The subject of the described stage of research is to determine the stability of the human voice for the one person. Unmodulated voiced sound, typically pronounced a few seconds by a person with a healthy speech executive and decision-making apparatus, the listener perceives as stable. That is, changes cannot be detected at the lowest level by the listening [Lombardi et al, 2009, Larrouy-Maestri et al, 2012]. For the research purposes was introduced the concept of micro-phoneme, as the smallest indivisible entity, repeated periodically while generating voice. Duration of the micro-phoneme is possible to calculate on the basis of the fundamental frequency and ranges from 4ms to 10ms. The shape of micro-phoneme includes both fundamental frequency sine wave, formants, and noise associated with the imperfection of the speech executive apparatus. The question posed by the author, on which tries to answer in this paper, is as follows: how stable during several microseconds is the human voice and how it affects the results of the analysis of the signals?

Data preparation

It was decided, that to develop of voice pattern would be conducted recording of voiced vowel, as in many similar researches [Bala et al, 2010, Fang and Gowdy, 2013]. A vowel used in the study is the vowel "a" pronounced a few seconds. On the next stages of research will be allowed the opportunity to change the frequency of voice generated by a recorded person. This is a very important statement because sound in the process of speaking is modulated, and an object of the present stage of research is to obtain a vector identifying a person, regardless of the frequency of sound generated by it. Work began with the recording of sound "a" pronounced by 1944ms by a man at the age of 31 years. The voice was recorded using the built-in condenser microphones device Tascam DR-40 with parameters: sampling frequency of 96kHz, the accuracy of 24 bits per sample, the format of uncompressed WAV/BWF. Recording formed the basis for the development of algorithms for selecting input signal for analysis. The waveform of the recording is shown in Fig. 1.

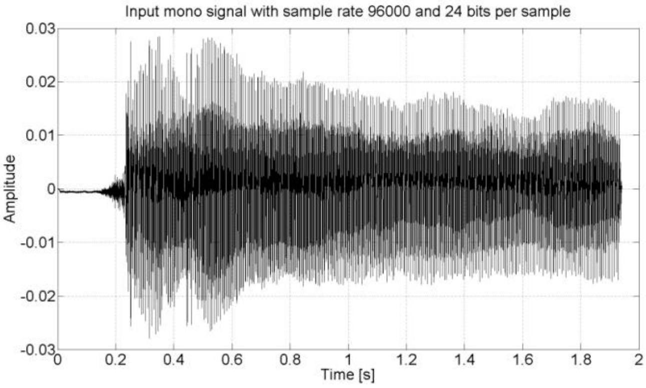


Fig. 1. Voice recording as a basis for the development of algorithms for data selection.

Subjectively established recording length, subjected to analysis, is 1000ms. It was also decided to eliminate the elements of silence and noise at the beginning and end of the recording. Also take into account the possibility of discontinuous speaking, that is, recording containing voiced elements divided by with silence or noise. Partly based on the published results of the author [Krzesimowski, 2012].

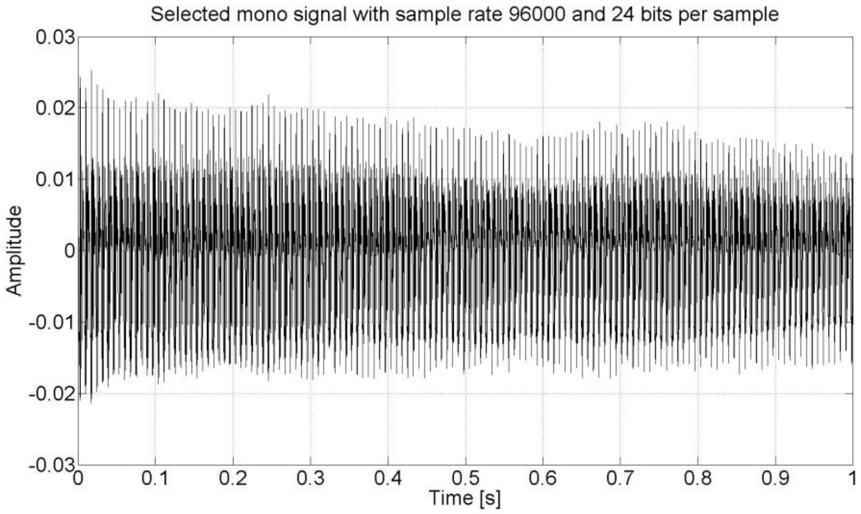


Fig. 2. A fragment of recording with a length of 1000ms selected from the recording shown in Fig. 1.

Material selection is based on the determination of the maximum value of the recording and division recordings on blocks with a length of 6400 samples. For the sound recorded at a frequency of 96000Hz are obtained sections of 67ms length. Then, the maximum value of the whole recording is compared to the registered in block. If this value is less than the 16.7%, block is rejected, otherwise the block is appended to the new vector. The result is the recording devoid of elements of silence and low-amplitude noise. The length of the recording thus obtained may be greater than expected 1000ms, that is why occurs the cutting of a predetermined length. For this purpose, it is determined the midpoint of data, and next, intervals of length 500ms counted from that point. Selection is relative to the centre to eliminate the modulation at the beginning and end of the recording. The waveform of the recording after the described selection is shown in Fig. 2.

Studies so far of the voice [Reilly et al, 2004, Grimaldi and Cummings, 2008], including studies of the author [Krzesimowski and Ciota, 2010], were based on analysis of voice in the form exactly as shown in Fig. 2. This time, it was decided to approach the analysis in a different way, not by analysing data from the whole recording, but by analysing the variability in the data based on the same recording treated as many independent recordings. This means, that waveform of signal analysis is not important at this stage, but its variability described in a statistical manner. This variability is then compared to all the recordings in order to extract the characteristic vectors of the person. The first step to getting so understood result is the division of the current data for a few to several dozen. At this stage of the study the recording was divided into 25 parts, the length of which is determined by the expression (1).

$$PL = \frac{(RL - 1)}{NP} \tag{1}$$

where: PL is the length of the current part, RL is the length of the recording and NP is amount of parts. Samples are cut down with an added margin with length dependent on the length of recording after the elimination of silence, and the number of the desired fragments in accordance with the expression (2).

$$ML = \frac{PL}{20} \tag{2}$$

where: ML is the length of the margin and PL is the length of the current part. Example charts of two sections of the recording are shown in Fig. 3 and Fig. 4. Each fragment is 44ms length.

In Fig. 3 are mapped the 5 full micro-phonemes, two micro-phonemes at the end and the beginning of the recordings are cut off. In Fig. 4 are mapped 6 full micro-phonemes, but the latter is cut off.

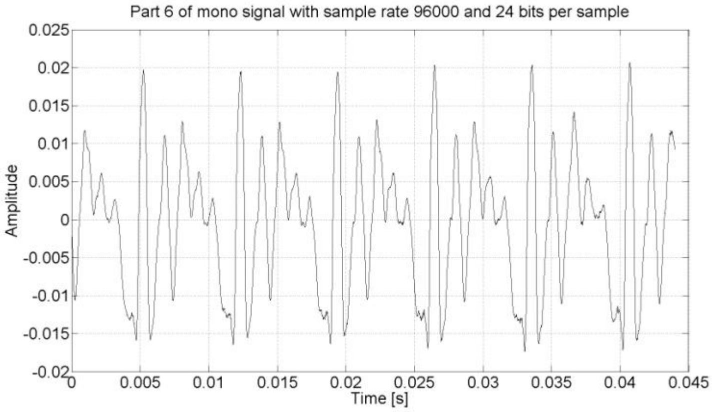


Fig. 3. Fragment number 6 after the split into blocks.

Given the fact, that a different person generates voiced sounds with different frequencies, the number of full micro-phonemes in so-characterized blocks may be different. In addition, because studies have used the frequency characteristics, it is possible that unacceptable errors can occur. Therefore it was decided to extract the blocks in such a way, that they contain only full micro-phonemes without cropping. For this purpose, the algorithm of trimming blocks was developed with respect to the local maximum value, and the counting of occurrences of that value. The fact is used of occurrence of one clear maximum in each micro-phoneme.

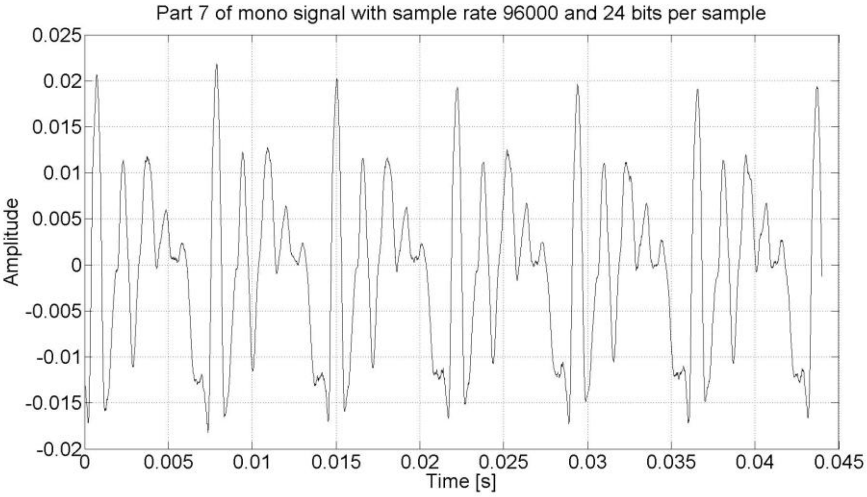


Fig. 4. Fragment number 7 after the split into blocks.

In the first step of cutting the maximum value is determined and the minimum value of block, and then is calculated the counter limit according to the expression (3).

$$CL = \frac{\left(|Max| + |Min|\right)}{10} \tag{3}$$

where: CL is limit for counter, Max is the maximum value of the current block and Min is the minimum value of the current block. Position the first values, which will satisfy the condition *max – limit*, is saved as the initial flag cut. Then all the positions of values are stored that meet condition, far from the position of the last value not less than 100 samples. After checking the entire block, as the final flag, is taken the position of value that starts the last section of maximum. In this way, ensured is sufficiently accuracy to extracting the full micro-phonemes, with designated a maximum error the peaks of the amplitude of 0.0005. So obtained examples of two blocks are shown in Fig. 5 and Fig. 6.

So obtained data blocks have different lengths related to the modulation of voice while recording. Moreover, in this example blocks contains a different number of micro-phonemes, 5 in Fig. 5 and 6 in Fig. 6. Therefore, the algorithm has been expanded so that the number of micro-phonemes was the same in the blocks. For this purpose, after cutting the recording in accordance with the previous point, its length is measured. This number is a reference value for the next block, and it is necessary to determine the 5% margin of error in estimation of length. If the next segment of recording is in this range, then is saved to the file, and another segment is collected. If one of the following sections of the recording is longer than the designated margin, it is trimmed in accordance with the criterion of maximum signal to the desired range. Alternatively, if a fragment length is shorter than the designated range, the length becomes a reference value. Thus, the appointed interval time is new, and splitting the entire recording begins again.

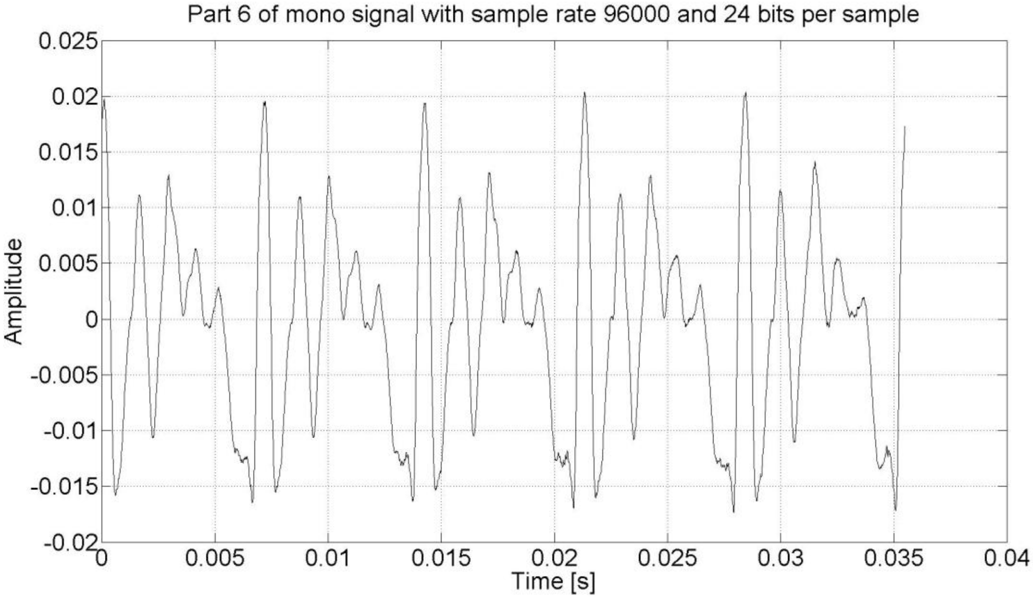


Fig. 5. Example section number 6 consisting only full micro-phonemes.

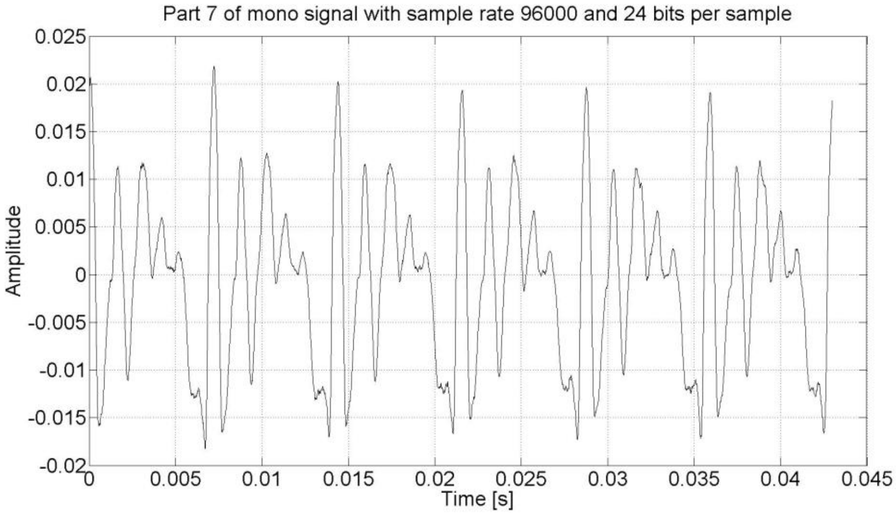


Fig. 6. Example section number 7 consisting of only full micro-phonemes.

Defined in this way algorithm allows to obtain fragments of recordings of the same number of micro-phonemes with an accuracy of 5% of their duration. In addition, in this way are eliminated the errors caused by trimming the voice recording in the first stage, if there were breaks in the sounds voiced pronouncing. Sample parts of recording after applying the final stage of selection are presented in Fig. 7 and Fig. 8. It is worth emphasizing, that the input and output data, at each stage of the division, are the sounds – the recorded voice that can be played. There is no interference in the waveform, operations affect only the duration of the recordings.

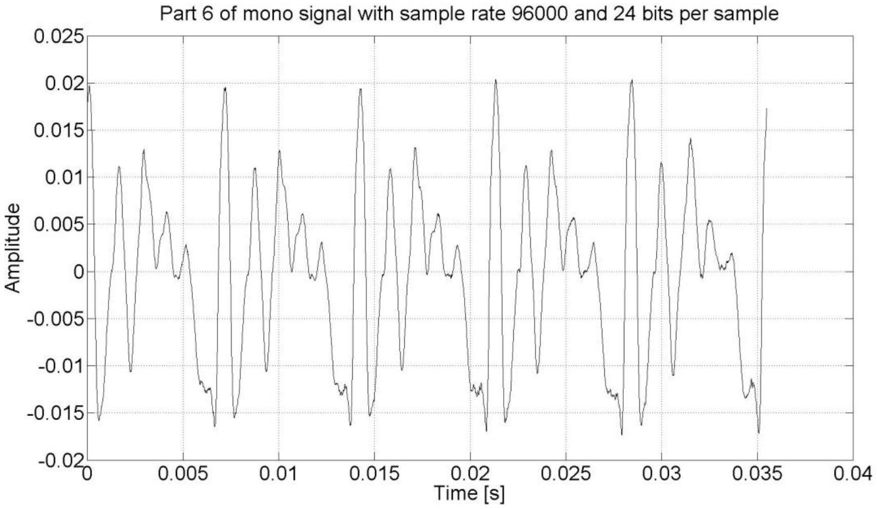


Fig. 7. Section number 6 composed exclusively of full micro-phonemes with a length within the limits of tolerance relative to the remaining blocks.

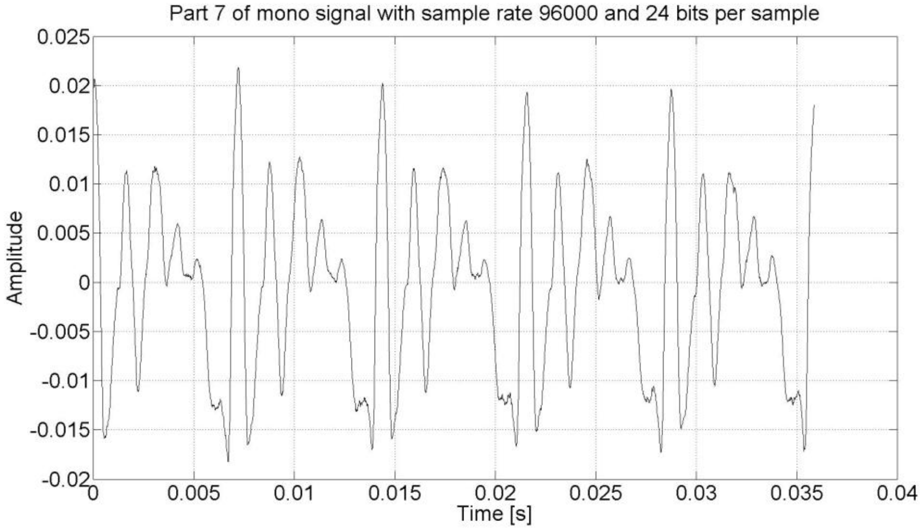


Fig. 8. Section number 7 composed exclusively of full micro-phonemes with a length within the limits of tolerance relative to the remaining blocks.

Analysis

The next step is to perform, for each block, signal analyses, mentioned above Fast Fourier Transform, spectrographic and cepstrographic. To get a complete view of the changes, the analysis were made of segments from each of the stages of division. It was noted small changes in signal waveform analysis.

These changes are undoubtedly related to the way in which was selected research material. Thus proved, that the method of selecting of the test material has an influence on the final results, regardless of the nature of an analysis. It was decided to estimate the size of the observed changes in the waveform using the statistical apparatus. Number of analysis of signals is equal to the number of blocks for which the input signal is divided. The values are stored as vectors in external files, possible to open in most computational programs.

In the same way are stored results for other analyses. Statement of vectors for each piece of recording in the matrix represents the starting point for statistical analysis. For the analysed sample of voice was determined:

1. standard deviation;
2. median;
3. arithmetic mean;
4. maximum value;
5. covariance;
6. correlation coefficient.

Two recent analyses were not carried out for a spectrogram, which results from the final vectors of different lengths for different blocks. In Fig. 9, Fig. 10 and Fig. 11 are presented as an example of the correlation coefficient for the Fast Fourier Transform relative to each of the sections of the recording of all three stages of the selection of material.

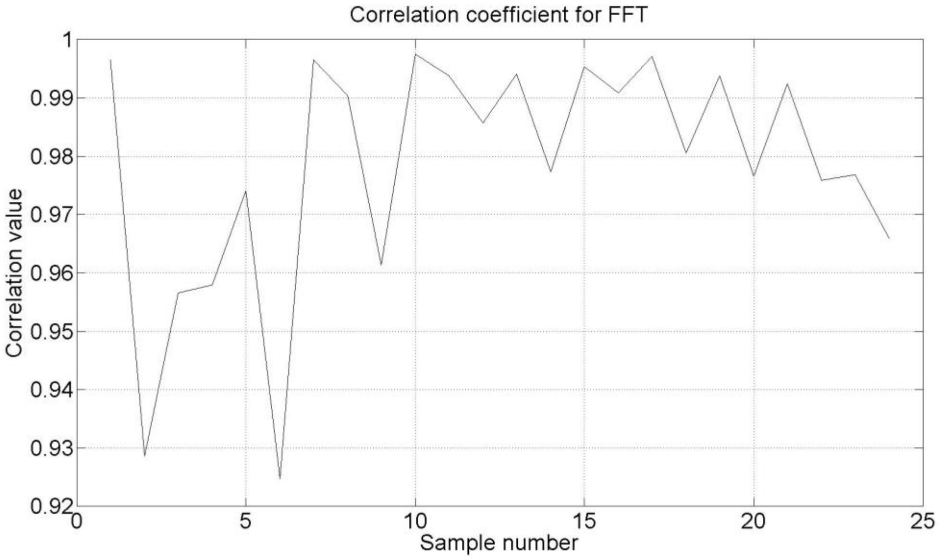


Fig. 9. Chart of the correlation coefficient for the Fourier transform performed for first stage of the selection of material.

At one stage the selection of data achieved are 16 characteristics of a statistical analysis of signal analysis. For described audio samples were obtained in a total of 48 plots, which are compared with each other. In the paper are described observations related to changes in the shape of the characteristics and values for each analysed separately fragments. There is no description of changes associated with the statistical values obtained from the analysis of signals.

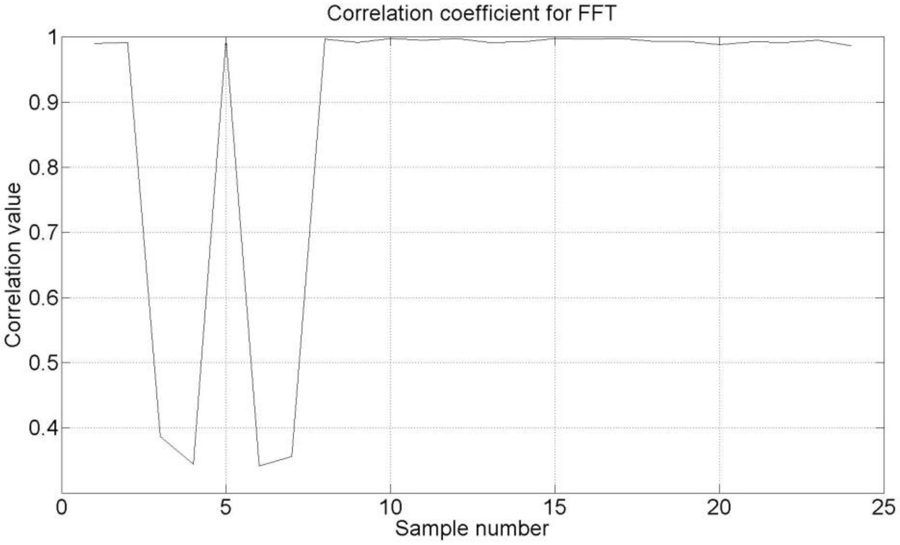


Fig. 10. Chart of the correlation coefficient for the Fourier transform performed for second stage of the selection of material.

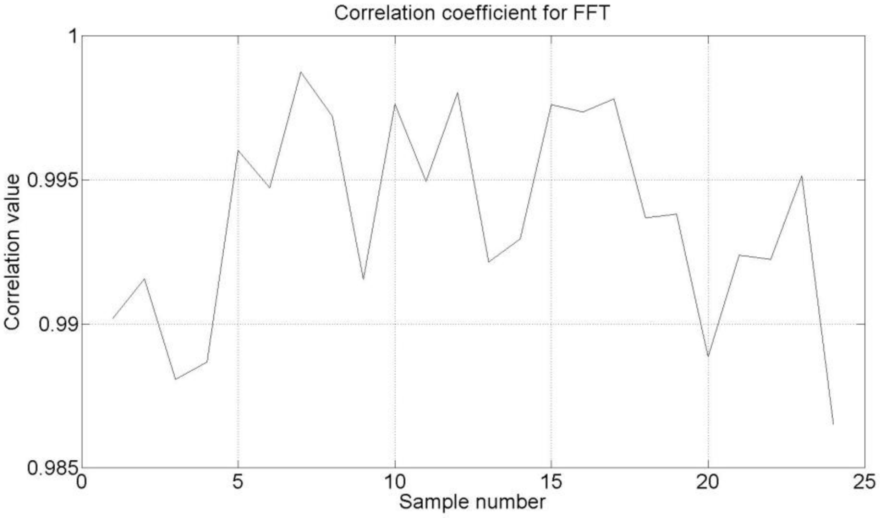


Fig. 11. Chart of the correlation coefficient for the Fourier transform performed for third stage of the selection of material.

In the case of the arithmetic mean and standard deviation was noted, that the values are 20% greater for the blocks after the first cutting step as compared to other steps. The shape of the waveform after the first and third stage of selection is on a close approximation the same. However, significant differences were noted in the waveform after the second selection phase, for the samples with number 4 and 7. These samples contain a different number of micro-phonemes than the other; the waveform of block number 7 has already been presented in Fig. 6. These same differences between blocks caused similar changes in the waveforms of the arithmetic mean and standard deviation for cepstrogram. Here as well noted a significant boost to values for samples number 4 and 7 after the second stage of selection. The shape of the waveform for the first and third cutting step, however, is different. In the case of the graph of the arithmetic mean and standard deviation for a spectrogram not be noticed significant changes associated with in the length of the analysed fragments, whereas the waveform after the second and third stages are almost identical.

There has been noticed a significant differences in the waveforms of the maximum and median Fourier transform between the first and second and third selection step. In addition, the lengths of the analysed blocks affect the value of the maximum for sample number 4 and 7. Apart from these two exceptions, the waveform of results of the second and the third stage is the same. An identical relationship has been observed for the waveform of maximum and median of cepstrogram and spectrogram. Here too the biggest differences are related to the different number of full micro-phonemes in blocks after the selection. The most interesting from the point of view of the usefulness of the described research have proved waveforms of covariance and correlation coefficient. For a chart of covariance for Fourier transform strong correlation was observed depending on the length of the analysed sections of the data. Again, the most abrupt changes have been observed in the waveform on the border of fragments with number 4 and 7. In addition, the change between the blocks of the same length and blocks having the same number of micro-phonemes proved to be negligible. On the other hand the differences between the first, the second and the third cutting steps has been observed for the waveform of covariance for cepstrogram. After the first stage, the amplitude of the waveform is characterized by significant changes. Waveforms in the second and third stage of selection are almost identical, and much more smooth.

The correlation coefficient has been treated in a particular way, as the most important parameter associated with the stability of the voice. In the case of the waveform for Fourier transform for the first stage of the selection values are in the range from 0.925 to 0.997 (approximately 7.5% difference). After the second stage, a significant influence the amount of micro-phonemes on the waveform has been observed, which can be seen in Fig. 10. Because they are compared with each other the current and the next block, the difference in their lengths has to affect the final result. After the third selection step values of the correlation coefficient ranges from 0.987 to 0.997, the difference between these values is approximately 1%. In the case of the cepstrogram no significant changes had been noticed in relation to the amount of micro-phonemes in the sample. However the differences have been noted in the waveforms for blocks after the first and subsequent stages of selection. Values of the correlation coefficient after the first cutting step are from 0.25 to 0.49, after the second step from 0.31 to 0.47, while the third stage from 0.31 to 0.5.

Conclusion and future work

The graphical results allowed formulating conclusions regarding the extracting methods for research material and sensitivity to changes in blocks for signal analyses. On the basis of the correlation coefficient has been found, that the material cut into equal blocks is a solution introducing an error. It results from the fact, that a block may include various amount of full and fragments of other micro-phonemes. From the parts of statistical analyses can be concluded, that the material selection can be terminated at this stage, but only in the case of the Fourier transform. The results of spectrographic and cepstrographic analyses proved to be sensitive to inaccurate trimming the research material. Such data should not be subjected to detailed analysis. Error appointed on the basis of the correlation coefficient for the Fourier transform is about 7.5%, there is need for further steps of selection of material. The second stage cuts allow storing only full micro-phonemes without their fragments. The result is the difference in the duration of the analysed fragments, which have significantly influenced the results of all analyses related to the Fourier transform. However, has not been noted the influence the length of the analysed data in the waveform of the spectrogram. To obtain meaningful final results, characterized by as little as possible error associated with the selection of data, has been developed third stage of cuts. This results in data, that contains the same number of micro-phonemes in each new block, and thus for the unmodulated voice is the time of each piece almost the same. Error determined on the basis of the correlation coefficient for the Fourier transform of the thus-obtained fragment of approximately 1%. This number can be considered as the value of volatility of the human unmodulated voice. This is a numeric value of changes, that may be seen in the waveforms of analysed fragments (compare Fig. 3 and Fig. 4) and which is associated with the stability of the speech executive apparatus. Moreover, thanks to the presented method obtaining the results of signal analysis is possible to use described frequency analysis, which are sensitive to different parameters related to the input data.

As the article explains how to evaluate the stability of the voice for one person, the next stage of research is to repeat the same steps for at least 100 people of all ages and gender. Still recorded will be voiced vowel, pronounced a few seconds without modulation. After estimating the stability of the voice for each person, on the basis of analysis of signals used here will be extracted factors, that do not change with the modulation of the voice of the person. Therefore will be referred to the similarity waveforms of micro-phonemes without their duration. This will allow the develop a model of the waveform of voiced sounds for the person, which will allow for the unambiguous identification of the person taking into account the error determined at the first, described in the paper, the research stage.

Bibliography

[Bala et al, 2010] A.Bala, A.Kumar, N.Birla. Voice command recognition system based on MFCC and DTW. In: International Journal of Engineering Science and Technology. Vol. 2 (12). 2010

[Fang and Gowdy, 2013] E.Fang, J.N.Gowdy, New algorithms for improved speaker identification. In: International Journal of Biometrics. Vol. 5, No 3-4. 2013

[Grimaldi and Cummings, 2008] M.Grimaldi, F.Cummings. Speaker identification using instantaneous frequencies. In: IEEE Transactions On Audio, Speech And Language Processing, Vol. 16 No 6. 2008

[Krzesimowski and Ciota, 2010] D.Krzesimowski, Z.Ciota. Signal processing of voice in case of patients after stroke. In: Electrical Review. R. 86 No 11a. 2010

[Krzesimowski, 2012] D.Krzesimowski. Preliminary processing of the human voice recordings. In: Conference Archives PTETiS. Vol. 31. 2012

[Larrouy-Maestri et al, 2012] P.Larrouy-Maestri, Y.Lévêque, D.Schön, A.Giovanni, D.Morsomme. The evaluation of singing voice accuracy: a comparison between subjective and objective methods. In: Journal of Voice. Vol. 27. 2012

[Lombardi et al, 2009] C.P.Lombardi, M.Raffaelli, C.DeCrea, L.D’Alatri, D.Maccora, M.R.Marchese, G.Paludetti, R.Bellantone, Long-term outcome of functional post-thyroidectomy voice and swallowing symptoms. In: Surgery. Vol. 146. 2009

[Reilly et al, 2004] R.B.Reilly, R.Moran, P.Lacy. Voice Pathology Assessment based on a Dialogue System and Speech Analysis. In: Proceedings of American association for artificial intelligence fall symposium on dialog systems for health communication. 2004.

Authors' Information

Krzesimowski Damian – Kielce University of Technology, Department of Applied Computer Science, al. 1000-lecia PP 7, Kielce 25-314, Poland; e-mail: damiank@tu.kielce.pl

Major Fields of Scientific Research: signal processing and numerical methods, mobile systems, automated systems

Business Intelligence Models

INFORMATIONAL SUPPORT OF MANAGERIAL DECISIONS AS A NEW KIND OF BUSINESS INTELLIGENT SYSTEMS

Volodymyr Stepashko, Oleksandr Samoilenko, Roman Voloschuk

Abstract*:* The paper considers main aspects of developing a combined system for informational support of operative managerial decisions. Such Managerial Decisions Informational Support System (MDISS) should contain data and models storage and the following three main components: subsystem for current analysis and visualization of operational management information; subsystem for modelling and forecasting of processes involved; subsystem for integral evaluation of interdependent indicators of a complex system state. Functional capabilities and features of the subsystems are presented.

A new type of sorting-out GMDH algorithm based on the principle of backward successive selection (BSS) of the most informative variables was used to perform modeling and prediction of economic indicators of Ukraine. Multiple dynamic autoregression models were built in form of systems of multidimensional difference equations of interdependent indicators. Construction of the models was based on statistical data of Ukraine economy for 13 years (1996 to 2008). Such models were built for 4 demographic and 5 investment indicators demonstrating good accuracy on validation part of the sample as well as on prediction period of 2009 and 2010 years.

An example of informational support task solution in the field of economic safety by the developed MDIS System is presented using the built predictive model for the area of Ukraine investment activity. Estimations of future evolution of the activity are based on predictions of this model as useful information for decision making.

Keywords: *business intelligence, decision making, informational support, GMDH, dynamic models, investment activity, demographic indicators.*

ACM Classification Keywords: *H.3.4 Systems and Software – Information networks; H.4.2 Types of Systems – Decision support (e.g., MIS); J.1 Administrative Data Processing – Government; I.6.5 Model Development.*

Introduction

Efficiency and quality of managerial or administrative decisions essentially depends on the timely supply of management process by necessary reliable information which describes these processes and phenomena that occur at a particular management object. Taking this into consideration, the informational support of managerial decisions is a vital problem. To solve this problem, it is proposed to develop appropriate tools based on inductive algorithms for analysis, modeling and prediction of complex processes.

To enhance the effectiveness of administrative decision making support, for instance, in the state economic safety field, it is necessary to monitor the main safety indicators statistics, to quantitatively evaluate the safety level, predict the indicators taking into account its dynamic interdependence and to visualize all the monitored and predicted information in the human-transparent form being easy-to-use by decision-makers. This approach leads to the necessity to analyze and solve the problem of construction a system for informational support of managerial decisions in the area.

Also the informational support concept refers to the new type of decision-making process organization that takes into account not only the traditional tasks of data storage, processing and visualization, but also providing full support for this process based on solving the analysis, modeling and forecasting tasks, presenting the results in informational and advisory form under conditions of constant changing the managerial situation.

Managerial Decisions Informational Support System as a kind of information system used to support managerial decisions

A Managerial Decisions Informational Support System (MDISS) being introduced here is a kind of Information Systems (IS) of a general type and at the same time it has some number of properties inherent to Executive Information Systems (EIS) and Decision Support Systems (DSS).

In general, an information system can be defined as an automated man-machine system that provides information to users from different organizations [DeSanctis, 1987].

Current DSSs that have arisen by the merger of management information systems and database management systems are the systems most adapted to solve problems of daily management activities and are tools that aim to help persons and/or authorities to make decision [Little, 1970]. Choice-making in complex problems, including those based on many criteria, may be carried out by DSS features [Power, 2000].

According to Turben [Turban, 1995], DSS has the following four main features: 1) uses data as well as models, 2) is designed to assist managers in making decisions for slightly structured and unstructured tasks, 3) supports, rather than replaces, decision made by managers 4) is designed to improve decisions.

An Executive Information System (EIS) or Information System for Managers is a specialized DSS that helps implementers to analyze important information and use appropriate tools to guide it in forming strategic decisions within a specific organization [Edwards, 1992].

We consider Managerial Decisions Informational Support Systems (MDISS) [Samoilenko, 2008] as systems that combine main characteristics of EIS and DSS. However, in contrast to DSS, they have no means of generating and actual making decision. In other words, tools for generating possible solutions and choosing the optimal one of them are not present in MDISS; user generates and makes decision with help of appropriate system features. But inherent for EIS means for data handling as well as visual representation and analysis are widely used here. This provides great opportunities for a user or decision maker to orient oneself in the current state of a problem and find the most appropriate course of action to resolve it. Modeling methods and forecasting tools aimed to help increase decision making effectiveness should be presented in MDISS as well.

General structure of the MDIS System

To solve the stated problem, the task of developing a software system should be considered. Such a system should contain data and models storage and the following three main blocks/components (Fig. 1):

- subsystem for current analysis and visualization of operational management information;

- subsystem for modelling and forecasting of processes involved;
- subsystem for integral evaluation of interdependent indicators of a complex system state.

Subsystem for current analysis and visualization of operative management statistics

This component provides support to the following functional tasks:

- collecting and storage of primary statistical data;
- pre-processing the row initial data;
- checking the correlation dependences of the primary indicators;
- tracking the status of every indicator;
- evaluating the integral state of a system;
- visualization and documentation of all the results.

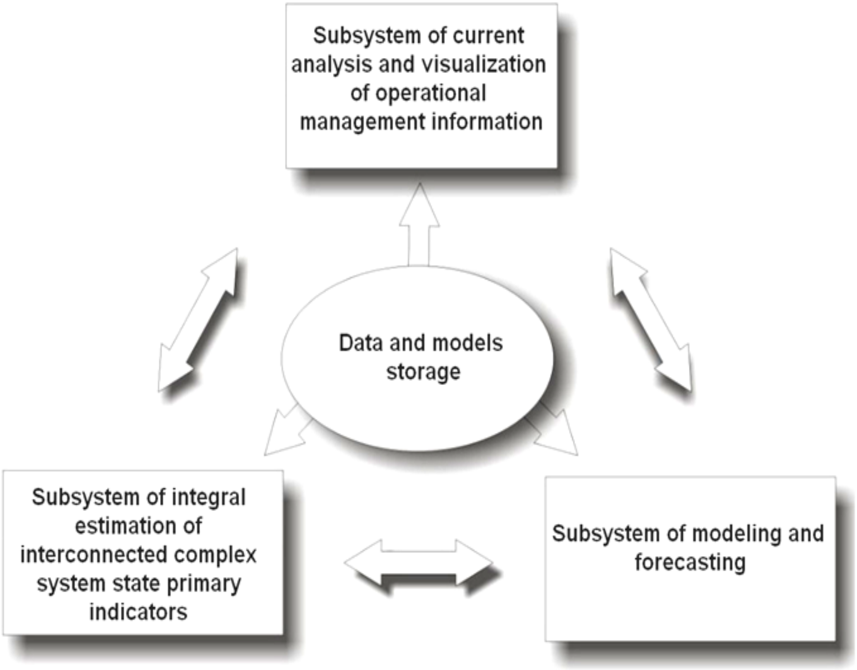

Subsystem of current analysis and visualization of operational management information
Data and models storage
Subsystem of integral estimation of interconnected complex system state primary indicators
Subsystem of modeling and forecasting

Fig. 1. Main components of Managerial Decisions Informational Support System MDISS

Subsystem for modeling and forecasting

It is based on the inductive approach to building a system model of a controlled multidimensional process. Methodology of this approach is based on maximum "extraction" of all necessary information from the data sample and focused on the inclusion to the model only the most significant/informative factors under specific conditions, rather than all factors which may affect the target value [Ivakhnenko, 1985]. The main functional characteristics of this subsystem:

- building models (manual and automatic modes);
- selecting the optimal model for information support;

- determining significance of each indicator (factor);
- visualization and documentation of the results.

Modeling and forecasting subsystem is intended to build models of optimal complexity using inductive modeling algorithms. Visual analytical tools are implemented for models analyzing to help users to choose best models. Based on the obtained models, approximations and predictions may be calculated here and for other system components. There are tools to analyze significance level of indicators regarding their influence on the final result.

Sorting-out GMDH algorithms based on known Combinatorial COMBI algorithm [Stepashko, 1981] are used for the modeling. Directed successive selection algorithm [Samoilenko, 2008] is realized in this component making it possible to effectively solve problems with large number of arguments.

Subsystem for integral evaluation of the state of a complex system based on interdependent primary indicators

To comprehensively analyze the performance of an economic system it is necessary to construct a generalized integral index for a group of interdependent primary indicators which jointly describe the state of such multidimensional system. This subsystem implements a new approach to quantitative calculating such integral index of a system state. This approach is based on non-linear normalization of the primary indicators taking into consideration certain expedient constraints on their optimal, satisfactory and unacceptable values [Stepashko, 2006].

Tools for dealing with complex data structures typical for governmental decision-makers are implemented. The data structures are displayed in a tree view. For example, sectors and sub-sectors of the economy may be represented as leaves and branches of this tree. Each structural element of the tree is associated with a set of panels based on a number of features specific for an appropriate type of this element. For example, integrated evaluation of process is done for any element and for the elements-leaves it is also possible to analyze the data using the implemented methods and conduct additional processing and analysis in the modeling subsystem.

The proposed subsystem provides support for such basic functional tasks:

- tracking the current state dynamics of the controlled process;
- data normalization by the developed technique;
- integrated and detailed evaluation of ongoing changes;
- analyzing the detected changes and finding main factors affecting them;
- identifying potentially dangerous phenomena and trends;
- visualization and documentation of the results.

Prediction of demographic indicators dynamics

The population size and age structure are fundamental data in definition of the state perspective revenues and expenditures including such important components as the pensions financing, social benefits, education and health facilities and so on. Without deep demographic substantiation it is impossible to determine the budget revenues amount, which depends on the labor force, the level of economic activity, education and qualification. Population is a major productive force as well as consumer of material goods. Rates and proportions of economic development, including production and consumption, and their changes significantly depends on the population, its age, educational, professional and social structures [Presidium, 2007].

Dynamics of population size and composition is characterized by a large degree of uncertainty. Processes of fertility, mortality and migration are stochastic in nature.

To build prediction model of demographic sphere in Ukraine, we used data from the Ministry of Economy for 1996-2010 years (15 points). The following 4 parameters were used:

x_1 – life expectancy at birth, years;

x_2 – conditional depopulation rate (the ratio of mortality to fertility factor), times;

x_3 – the proportion of the elderly population in the total population, %;

x_4 – the demographic burden of disabled on the working-age population, %;

The modeling of dynamics is executed in the class of multidimensional difference models when interdependence of the indicators is taken into account. Accordingly, structural and parametric identification is performed for models of the following type:

$$x_i(t) = \sum_{j=1}^{4} \theta_{ij} x_j(t-1) + \sum_{k=1}^{4} \theta_{ik} x_k(t-2) + \sum_{p=1}^{4} \theta_{ip} x_p(t-3), \quad i = \overline{1, 4} \tag{1}$$

or in matrix form

$$X(t) = \Theta_1 X(t-1) + \Theta_2 X(t-2) + \Theta_3 X(t-3) \tag{2}$$

where X is the vector of 4 elements, Θ_1, Θ_2 i Θ_3 – matrices of model coefficients (3) with 4×4 dimension.

As noted above, we have a sample of 15 points (1996-2010 years). Three points (1996-1998 years) are used as lag (lag L=3). Thus, for each $x_i(t)$ as an output variable we have a sample with 12 arguments and 12 points ($t = \overline{1999, 2010}$).

For each indicator, linear models were constructed using combinatorial algorithm BSS (Backward Successive Selection) [Samoilenko, 2008]. The main goal of the BSS algorithm is to select the most informative arguments by sorting out the relatively small group of built models as compared to the exhausted search. This approach significantly reduces the number of models to be built in order to find the optimum model.

Models have complexity s = 1,…,12 and were built on 6 training points (1999-2004 years). The used lag is 3 (initial conditions 1996-1998 years).

Constructed models for one-step forward prediction were estimated by regularity criterion AR on 4 checking points (2005-2008 years). Prediction qualities of constructed models were tested on 2 points (2009 and 2010 years). Successive calculation of predictive values for all parameters and use of them to build the next prediction enables to obtain predictions for a given number of steps.

To construct the models normalized values are used. For each normalized indicator following models was built:

$$x_1(t) = 1.23x_1(t-1) + 1.67x_1(t-2) + 1.79x_2(t-2) - 1.29x_3(t-2),$$

$$x_2(t) = x_1(t-1) + 0.45x_4(t-2) + 0.17x_2(t-3) - 0.39x_3(t-3) + 0.97x_4(t-3),$$

$$x_3(t) = 1.02x_2(t-2) - 5.52x_4(t-2) + 2.24x_2(t-3) + 5.74x_4(t-3) + 0.37x_3(t-3),$$

$$x_4(t) = -0.07x_4(t-1) + 4.16x_2(t-2) - 8.03x_3(t-2) + 6.98x_2(t-3) + 4.57x_3(t-3) - 0.64x_4(t-3),$$

Figures 2 – 5 show predicted values compared to actual.

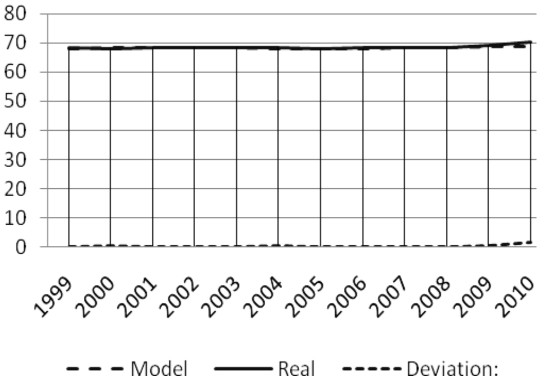


Fig. 2. Life expectancy at birth, years

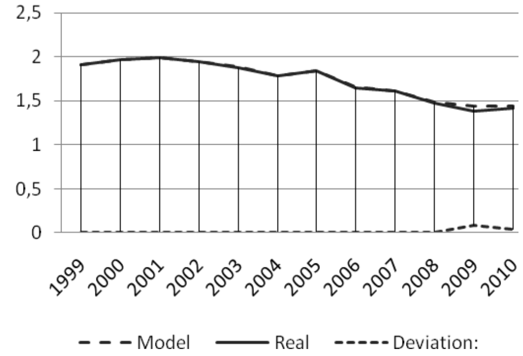


Fig. 3. Conditional depopulation rate (the ratio of mortality to fertility factor), times

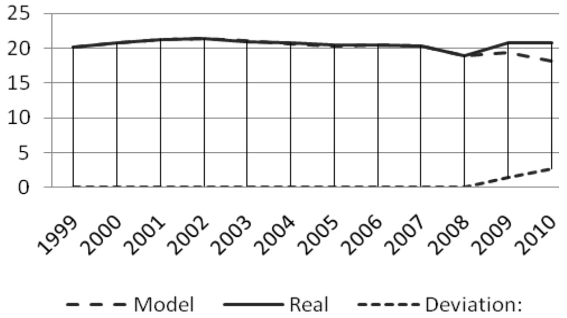


Fig. 4. The proportion of the elderly population in the total population, %

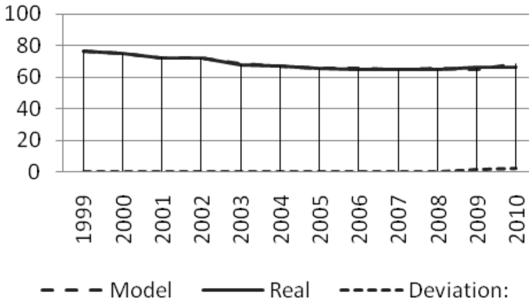


Fig. 5. The demographic burden of disabled on the working-age population, %

It is obvious that constructed models mostly show good prediction results even with a small learning sample.

The largest differences of predicted values to actual can be seen in the model constructed for the indicator reflecting the proportion of older people in the total population. Perhaps this is due to some external factors that are not taken into account in the model.

Construction of dynamic system models for investment activity indicators

The main aim of the state economic safety is to guarantee its stable and effective functioning now and high potential of development in the future. The investment component is a special subsystem of economic safety.

World experience indicates that countries with transition economies are unable to come out of economic recession without involvement of foreign investments, because in such a way they gain access to modern techniques, management technologies and so on. In addition, foreign investments make an important contribution to the macroeconomic stabilizing. Difficult economic state of a country as an independent state also constrains the country to apply to foreign investments.

Ukraine investment activity is characterized by a range of economic indicators. Ministry of Economy of Ukraine uses them to analyze the economic safety level. There are the following indicators in this area: accumulated depreciation level, share of direct foreign investments in general investments amount, ratio of investments amount to the fixed assets cost, ratio of investments amount in the capital asset to the Gross Domestic Product (GDP), direct foreign investments to GDP ratio.

To enhance the effectiveness of administrative decision making support in the investment activity field, it is necessary to predict the investment indicators taking into account its dynamic interdependence. This approach leads to the necessity to analyze and solve the problem of modeling and prediction of the given set of economic indicators in this area.

Similarly to the above, the modeling of Ukraine investment activity indicators was carried out with the use of multiple autoregression models in the form of multidimensional difference equations of interdependent indicators [Stepashko, 2010]. The structure and parametric identification was executed for model variants in the same form as (1) or (2) with limitation of model complexity by exhaustive search using combinatorial algorithm COMBI GMDH. Evidently that Θ_1 and Θ_2 in (2) are 5 by 5 matrices of parameters of the model of the type (1).

The models of the limited complexity from s=2 to s=6 (according to observations number of the training subset) for every index were built with the use of the combinatorial algorithm for models structure optimization. After building models being optimal by the regularity criterion [Ivakhnenko, 1985] for all indicators, the one step forward system predictions for every indicator were obtained. The sequential computation of all indicators values enables to get predictions for some steps forward.

Real data for the 5 mentioned above indicators of investment activity for years 1996 to 2008 was used for modeling. We have built the following dynamic system models:

$$x_1(t) = 0.524x_1(t-1) + 0.434x_1(t-2) + 0.174x_2(t-1) + 0.236x_2(t-2) - 0.281x_3(t-1)$$

$$x_2(t) = 0.408x_2(t-1) + 0.676x_2(t-2)\,,$$

$$\begin{aligned} x_3(t) = 0.725x_1(t-1) - 1.581x_2(t-1) - 0.302x_2(t-2) - 0.345x_3(t-2) - \\ -1.417x_4(t-1) + 8.954x_5(t-1) \end{aligned}\,,$$

$$x_4(t) = 0.453x_1(t-1) - 0.186x_2(t-1) - 0.482x_2(t-2) + 1.356x_3(t-1)\,,$$

$$x_5(t) = -0.007x_2(t-1) + 1.144x_5(t-1)\,.$$

Real values of the indicators are illustrated in Fig. 6 – 10 by a firm line and computed values by a dotted line. The last 4 marked values are prediction for 1 to 4 steps, respectively. Errors of developed models in examination points (2007 and 2008 years) are presented in Table 1.

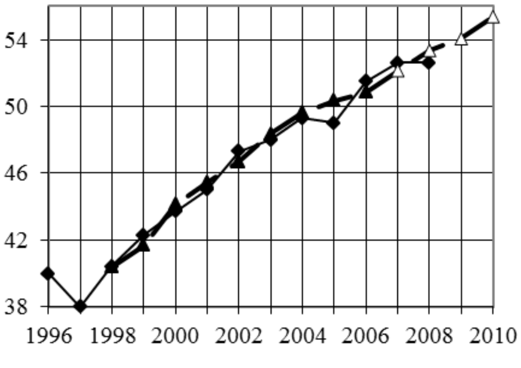


Fig.6. Accumulated depreciation level

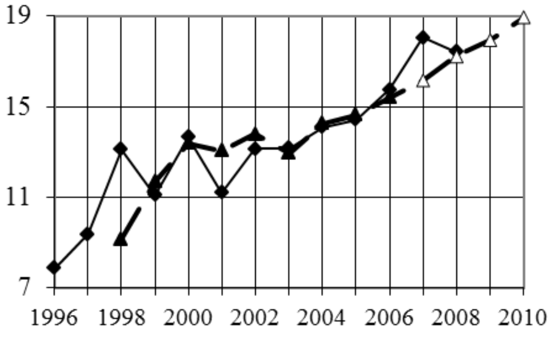


Fig.7. Share of direct foreign investments in general investments amount

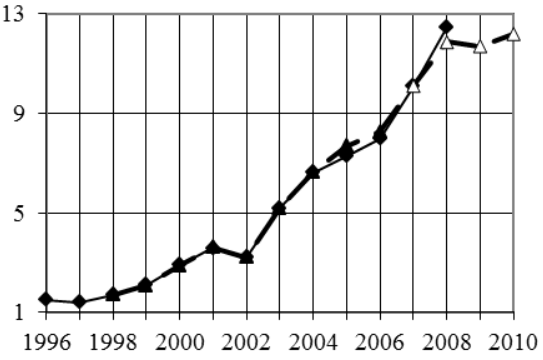


Fig.8. Ratio of investments amount to the fixed assets cost

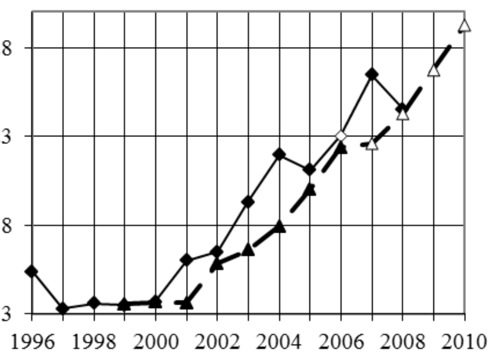


Fig.9. Ratio of investments amount in the capital asset to GDP

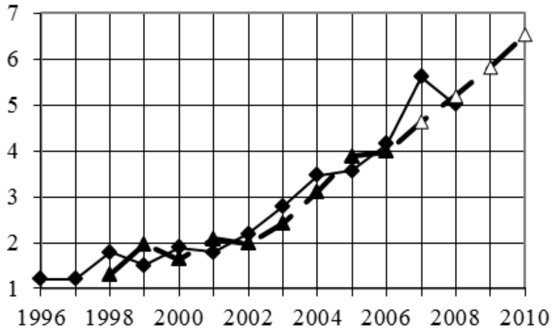


Fig. 10. Direct foreign investments to GDP ratio

Table 1. Errors of developed models for 5 indicators in examination points, %.

Indicator	1	2	3	4	5	Average
Accuracy	1.2	5.8	2.5	8.6	10.6	5.7

Investment activity index prediction

These models and calculated by them predictions were used to get the so-called *inertial prediction* of the integral index for investment area in Ukraine for 2011-2012 years. The respective results are shown in Fig. 2.

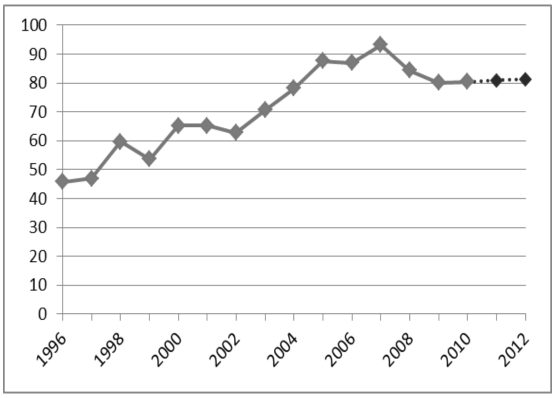


Fig. 11. Evolution and inertial prediction of the integral index dynamics for the investment area

This information, received using tools of the MDIS system, on predictive behavior of the integral index for investment area may be used to facilitate the managerial or administrative decision making process in relevant governmental bodies.

Conclusion

The proposed system for informational and analytical support of managerial decisions is aimed to evaluate, analyze and forecast complex economic processes with respect to their interdependence. It supports the following main tasks: import of statistic data; evaluation and analysis of ongoing changes; construction, analysis and correction of models; operative forecasting of processes in real time. This system contains data and models storage and consists of three main components: subsystem for current analysis and visualization of operational management information; subsystem for modeling and forecasting the interdependent primary indicators; subsystem for integral evaluation of the complex system state.

The system has an interface for tabular and graphical data analysis. The results of modeling and forecasting are presented graphically and analytically. Models and forecasts are corrected in real time. The system functioning is demonstrated on the example of current analysis of the investment activity as an important component of economic safety of Ukraine.

The developed information technology is intended for a comprehensive analysis of socio-economic processes to improve efficiency and quality of managerial decisions due to identifying hidden regularities of socio-economic processes and respective reduction of inefficient decisions at different levels of economical and political governance.

The experiments show the effectiveness of GMDH algorithms even at small size of the given data, when the number of statistical points is less than the number of variables that affect the output value. During the research, dynamic system models were built for demographic and investment areas using the data of the Ministry of Economy of Ukraine. Most of the built models in class of the multiple autoregression equations demonstrate good forecasting results confirming their effectiveness.

This paper presents project of an integrated environment for Managerial Decisions Informational Support System. Such kind of informational system can be applied to solve typical tasks of business intelligence including analysis, evaluation, modeling, forecasting, classification, visualisation and others. Applying GMDH algorithms in business intelligence systems gives particularly promising opportunities towards building complex models for business data analysis.

An illustrative example of informational support task solution in the field of economic safety by the developed MDIS System was presented using the built predictive model for the area of Ukraine investment activity. Estimations of future evolution of the activity are based on predictions of this model as useful information for managerial decision making.

Bibliography

[DeSanctis, 1987] DeSanctis G., Gallupe R. A Foundation for the Study of Group Decision Support Systems // Management Science, 33, no. 5, May 1987. — P. 589—609.

[Little, 1970] Little I.D.C. Models and Managers: The Concept of a Decision Calculus // Management Science, 1970. - V. 16. – N 8.

[Power, 2000] Power D.J. Web-based and model-driven decision support systems: concepts and issues. Americas Conference on Information Systems, Long Beach, California, 2000.

[Turban, 1995] Turban, E. Decision support and expert systems: management support systems. Englewood Cliffs, N.J.: Prentice Hall, 1995.

[Edwards, 1992] Edwards J.S. Expert Systems in Management and Administration - Are they really different from Decision Support Systems? // European Journal of Operational Research, 1992. - Vol. 61. - P. 114-121.

[Samoilenko, 2008] Samoilenko O.A., and Stepashko V.S. A system for informational support of operative managerial decisions making // Modelling and state control of ecological and economical systems of a region. Collection of research papers, No. 5. – Kyiv: IRTC ITS NASU, 2008. – P. 211-219. (In Ukrainian).

[Ivakhnenko, 1985] Ivakhnenko A.G., Stepashko V.S.: Pomekhoustoichivost modelirovania (Noise-immunity of modeling). Kiev: Naukova Dumka, 216 p, 1985. (In Russian).

[Stepashko, 1981] Stepashko V.S. A Combinatorial Algorithm of the Group Method of Data Handling with Optimal Model Scanning Scheme, Soviet Automatic Control, 14, No. 3, (1981), P. 24-28.

[Samoilenko, 2008] Samoilenko O., and Stepashko V. A method of Successive Elimination of Spurious Arguments for Effective Solution of the Search-Based Modelling Tasks. – Proc. of the II Internat. Conf. on Inductive Modelling ICIM-2008, 15-19 Sept. 2008, Kyiv, Ukraine. – Kyiv: IRTC ITS NANU, 2008. – P. 36-39.

[Stepashko, 2006] Stepashko V.S., Melnyk I.M., Voloshuk R.V. Models for Synthesis of Integral Evaluation of the Status of a Complex System of Interdependent Primary Indicators // Modelling and state control of ecological and economical systems of a region. Collection of research papers, No. 3. – Kyiv: IRTC ITS NASU, 2006. – P. 275-284. (In Ukrainian).

[Stepashko, 2010] Stepashko V., Yefimenko S., Voloshuk R. Investment Activity Prediction with the Use of Multiple Autoregression Models / Proceedings of the III International Conference on Inductive Modelling ICIM-2010, 16-21 May 2010, Yevpatoria, Crimea, Ukraine. – Kherson: KNTU, 2010. – P. 149-151.

[Yefimenko, 2009] Yefimenko S. M., Kvasha T. K., Stepashko V. S. Systemne prohnozuvannya dynamiky vzajemozalezhnyh pokaznykiv enerhetychnoi sfery (System forecasting the dynamics of interdependent indices of Ukraine energy sector) // Induktyvne modeluvannia skladnyh system. Zbirnyk naukovyh prac. (Inductive modeling of complex systems. Collected articles). Kyiv: IRTC ITS, pp. 54-59, 2009. (In Ukrainian).

[Presidium, 2007] About demographic development forecasts for Ukraine to 2050 year // Resolution of the Presidium of the National Academy of Sciences of Ukraine № 313, November 21, 2007

[Samoilenko, 2007] Samoilenko O.A., Stepashko V.S. Combinatorial GMDH algorithm with successive selection of arguments. // Proceedings of the II International Workshop on Inductive Modelling IWIM-2007, September 19-23, 2007, Prague, Czech Republic. – Prague: Czech Technical University, 2007. – P. 139-143.

Authors' Information

Volodymyr Stepashko – Head of Department for Information Technologies of Inductive Modeling of IRTC ITS, Professor, Dr Sci, P.A.: 40, Akademik Glushkov Prospect, Kyiv, Ukraine, 03680; e-mail: stepashko@irtc.org.ua

Main Fields of Scientific Research: Data analysis methods and systems, Knowledge discovery, Information technologies of inductive modelling, Group method of data handling (GMDH)

Oleksandr Samoilenko – researcher of IRTC ITS NASU, P.A.: 40, Akademik Glushkov Prospect, Kyiv, Ukraine, 03680; e-mail: soa0pga@gmail.com

Main Fields of Scientific Research: Information technologies of inductive modelling, Business Intelligence solutions

Roman Voloschuk – researcher of IRTC ITS NASU, P.A.: 40, Akademik Glushkov Prospect, Kyiv, Ukraine, 03680; e-mail: volrom@bigmir.net

Main Fields of Scientific Research: Information technologies of inductive modelling, Business Intelligence solutions

STATISTICAL MODELS FOR THE SUPPORT OF OVERBOOKING IN TRANSPORT SERVICE[5]

Vladimir Averkiev, Mikhail Alexandrov, Javier Tejada

***Abstract**: Overbooking is a strategy of extra ticket reservation, which needs precise models concerning the probability of ticket acquisition. Without such models a company will have losses related to unused seats or compensation for the absence of these seats. In the paper we consider several models for such a forecast and demonstrate their functionality on real data of one Peruvian railway company. The best model includes an original algorithm of revealing seasonal prevalence.*

***Keywords**: overbooking, transport service, statistics*

***ACM Classification Keywords**: 1.6.4. Model validation and analysis*

Introduction

Companies associated with transport services, often use overbooking strategy. This strategy consists in an redundant reservation of tickets when some passengers are expected to refuse their trip in future. Such a strategy must be supported by accurate predictive models, which could answer the question about the share of purchased tickets among all reserved tickets for a given train (or flight) in a given day. Here:

- pessimistic forecasts may lead to losses related to unused seats;
- optimistic forecasts may lead to losses related to compensation for passengers, which will not have seats.

In the latter case the company being responsible for ticket reservation may have a reputation loss.

The problem of determining virtual capacity of vehicles were considered in [Barnhart, 2003] and [Talluri, 2005]. In the paper [Mozgovaya, 2011] virtual capacity of transport were estimated on the basis of forecasts for the number of purchased tickets and the number of return tickets. In this paper we consider two forecasting models. The first model does not take into account any characteristics related to a given trip.. It is considered as a basic model. The second model takes into account the travel date and the interval of reservation. This is two-parameter model. It can be built on the basis of regularities concerning seasonality and reliability of preliminary reservation. To reveal these regularities we use two original algorithms. The first algorithm realizes time series decomposition. It was described in [Averkiev, 2012]. The second algorithm determines the intervals of constancy for the mean values of time series. Here the method of discrete dynamic programming is used [Bellman, 1962]. In both cases the visual presentations of ticket sales are good helpers for decision making.

The paper is structured by the following way. In section 2 we describe source data. In section 3 we present algorithms and results of experiments. Section 4 contains conclusions.

[5]The work done under support of the British Petroleum grant (BP-RANEPA 2013).

Data and models

Data description

Initial information for modeling is data of one Peruvian railway company. It is results of daily ticket sales during 2012 from January 1 to December 31. Table 1 shows parameters of various reservations.

Table 1. Initial data (example).

N	*D1*	*D2*	*D3*	*D4*
1000567	16.12.2011	06.04.2012	1	1
1033112	22.03.2012	25.04.2012	4	0
2034566	22.03.2012	25.04.2012	2	2
………	…	…	…	…

Here: *N* is the order number, *D1* is the date of reservation, *D2* is the travel date, *D3* is the number of reserved seats, *D4* is the number of used seats.

Conditions of booking and purchase of tickets as the follows:

- Several reservations can be made in the same day and for the same date of trip;
- Reservation interval is from 0 to 365 days (one year);
- Ticket sales can be done in any day including the day of the trip.

Models under consideration

The aim of the work is to forecast ticket sales on a given day. For the simplicity we assume that there is only one train per day related to a given route. This forecast concerns the share of purchased tickets among reserved tickets. *Hereinafter we will associate this share with the probability that a given reserved ticket will be really bought.* So, if we know this probability *p* and the number of reserved tickets Q_R then the number of purchased (used) tickets Q_U can be calculated by a simple formula $Q_U = p\ Q_R$. We consider several models. Each model is based on one of the hypotheses:

Hypothesis 0. There are no any regularities related to probability *p*

Hypothesis 1a. There is a stable dependence between the probability *p* and the date of trip

Hypothesis 1b. There is a stable dependence between the probability *p* and the interval of reservation.

Hypothesis 2. There are stable dependences between the probability *p*, the date of trip and the reservation interval.

List of all models presented in Table 2:

Table 2. Forecast models with their parameters

Model	***Day of trip***	***Interval of Reservation***
Basic model	-	-
Intermediate decomposition model	+	-
Intermediate interval model	-	+
Two-parametric model	+	+

The more detailed models can be considered. For example, a three-parametric model can take into account additionally the number of reserved tickets. Usually, the more number of tickets are reserved the less probability of purchased tickets is. In this paper we consider only the models mentioned in the table.

To build models we use data of the first 270 days for training models and the last 90 days for testing models. To evaluate model quality MAPE index is used. MAPE stands for Mean Absolute Percent Error.It is calculated by the formula:

$$MAPE = \frac{1}{90}\sum_{t=271}^{360}\left|\frac{\widehat{p_t}-p_t}{p_t}\right| * 100\%$$

where $\widehat{p_t}, p_t$ are data of modeling and data of experiments.

Preprocessing

To construct the models listed in the table, preprocessing is performed. This preprocessing consists in 2 operations:

1) Check of values. All data should be positive; the number of purchased tickets must be no more than the number of reserved tickets, etc.

2) Compression of data. All reservations with the same date of trip and the same date of reservations are combined: the number of reserved tickets and the number of purchased tickets are summed. Besides the reservation interval is calculated. Table 3 is a result of preprocessing:

Table 3. Data after preprocessing (example).

D	***T***	***R***
97	110	0.38
97	112	0.32
116	33	0.29
…	…	…

Here: *D* is a day of trip (from January 1, 2012), *T* is a reservation interval measured in days; *R* is a share of purchased tickets.

Algorithms and experiments

Basic model

The basic model does not consider any regularities in ticket reservation concerning time. It is assume that the ratio of the number of purchased tickets to the number of reserved tickets is a constant. According to the data of Table 3 we have:

$$p_D = \bar{p} = \frac{\sum_{i=1}^{270} R_i}{270}$$

Dependence R_D on day D is shown in Figure 1.

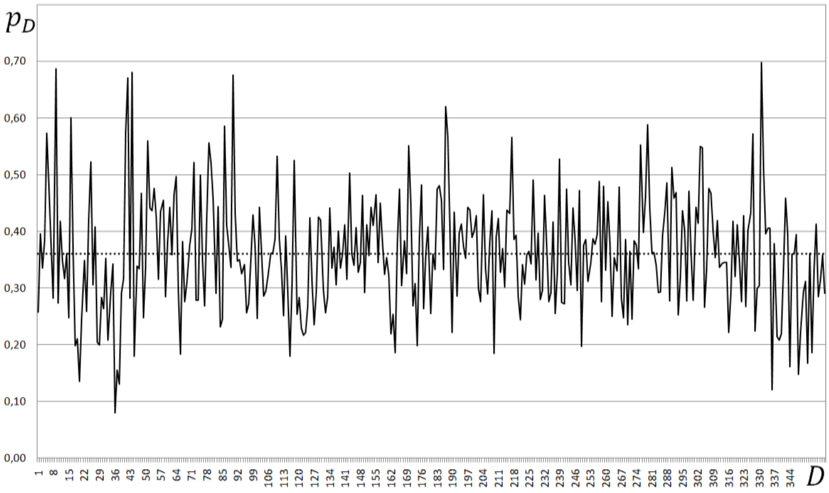


Fig.1. Dependence of ratio R_D on day D

The experimental results are as follows:

- The probability of purchasing ticket $p_D = 0.36;$
- The accuracy of forecast MAPE = *26,7%.*

Seasonality identification

Seasonality allows to take into account the date of the trip. To reveal the seasonality we need to decompose a given time series into the components, i.e. we need to build such a model $p_t = T_t + S_t + N_t$. Here:

- T_t is the trend, the main component;
- S_t is the seasonal component, which gives information about the periodic oscillations in time series;
- N_t is the random component of time series.

The method of curvature minimization is used to evaluate the trend and seasonality. The corresponding algorithm is described in [Averkiev, 2012]. The forecast model based on the decomposition is built by the following way:

1) First, the local decomposition in a current step t of the time series is implemented. Thus, we find seasonal coefficients k_t.

2) Second, we calculate the local time series average value using a moving average:

$$p_{t+1} = \frac{\sum_{i=1}^{16} p_{t+1-i}}{16}$$

3) And finally, it is determined the forecast value of probability like the sum of time series average value and seasonal coefficients k_t:

$$\hat{p}_{t+1} = \frac{\sum_{i=1}^{16} p_{t+1-i}}{16} + k_t$$

Identification of patterns in reservation period

To reveal the regularities related to the length of reservation period let see on Figure 2 Axis X is the reservation period, and axis Y is the ratio of the number of purchased tickets to the number of reserved tickets. As have mentioned above we associate this value with a probability.

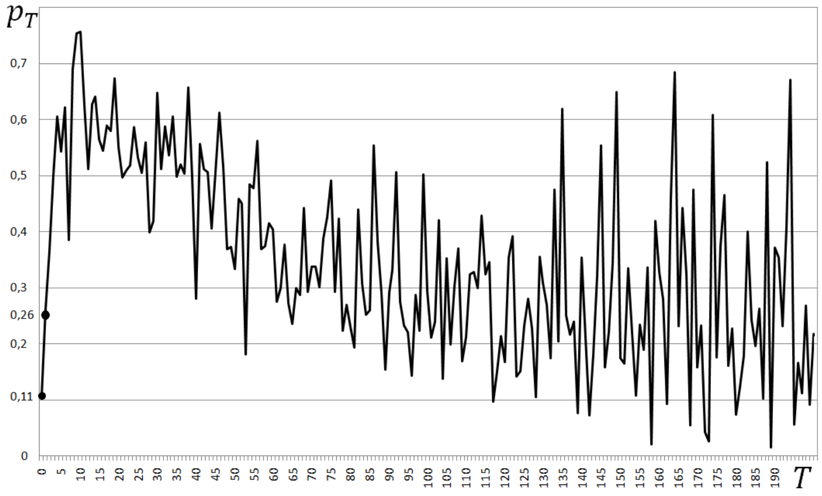


Fig. 2. Dependence of ratio R_T on interval of reservation T.

The figure clearly shows that a) the values p_T for bookings in the day of departure and one day before the departure (i.e. T=0 and T=1) are significantly smaller than p_T for other days (we mean here the stable values) b) there are no data on graph concerning intervals $T \geq 200$ because that data are not representative. With the method of discrete dynamic programming [Bellman, 1962] we revealed two intervals related to mean value of p_T on these intervals (the number of intervals were fixed but the boundary between them were unknown). Just the first two points proved to belong to the first interval. The forecast model based on the selected interval is built by the following way:

1) Each booking gets its weight. Weight of booking is equal to the probability of its realization. Bookings made in the day of departure and one day before the departure get weights 0.11 and 0.26, respectively. All other bookings get weight 0.36.

2) All weighted bookings for the departure day are summarized. As a result, we obtain the desired ratio $\hat{p}_t$.

The values 0.11 and 0.26 are taken from the Figure 2. The value 0.36 corresponds to the basic model. One should say that the selection of two intervals is enough subjective. In future it is necessary to consider more objective procedures.

Two-parametric model

In two-parametric model we use decomposition of a given time series together with taking into account interval of reservation. The model is built by the following way:

1) Trend and seasonal component are estimated

2) Correction is made using interval of reservation.

3) The final probability is the composition of the results of previous steps. We calculate the probability by the formula:

$$\hat{p}_D = \frac{\sum_{i=1}^{N} I_D R_i}{\sum_{i=1}^{N} I_D} + \frac{\sum_{j=1}^{16} p_{D-j}}{16} + k_D^{seasonal}, where\ I_D = \begin{cases} 1, when\ D_i = D \\ 0, when\ D_i \neq D \end{cases}$$

Here *D* is the day of departure, and *N* is the total number of bookings related to day *D*. In this expression the first term is the correction related to the interval of reservation, the second term and the third terms take into account the local mean value and seasonality.

Testing on real data gives MAPE = 23.7%. Figure 3 shows the experimental data (thick grey line) and the forecast data (thin black line). One should remind that we use the interval 270 days for learning model and the last 90 days for testing model.

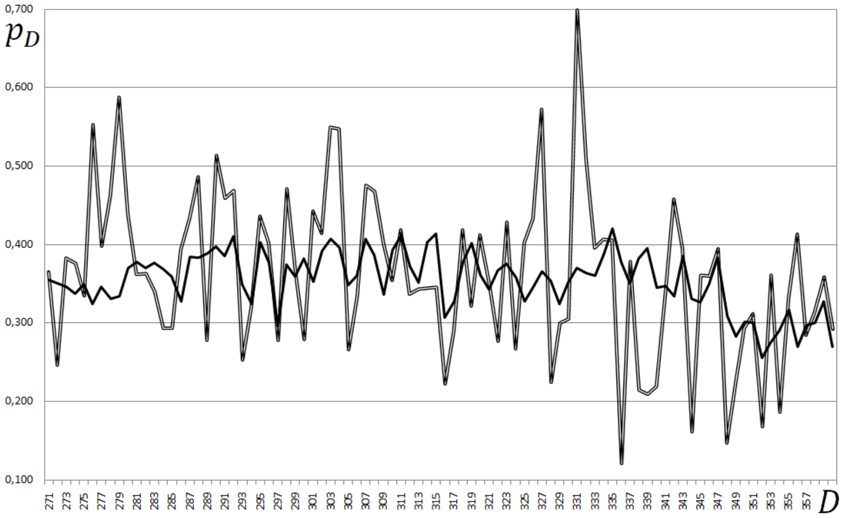


Fig. 3. Real data and forecast on the examination period

This figure shows that the series p_D is still strongly volatile but the most of variations are explained by an irregular component. It is also seen that the forecast well predicts general fluctuations of the trend

Conclusions

In the paper we build several models to predict the probability for purchase of reserved tickets. The best two-parametric model takes into account the day of departure and the interval of reservation.

We tested the basic and two-parametric models on real data related to the activity of one Peruvian railway company. The two-parametric model provides the accuracy 23.7%. In the comparison with the basic model with its accuracy 26.7% we have the absolute improvement 3% and the relative improvement 11,5%. This result shows the possibility to increase the efficiency of overbooking strategy.

Bibliography

[Averkiev, 2012] A. Kovaldji, V. Averkiev, M. Sarkissyan. Smoothing and prognosis of multi-factor time series of economical data by means of local procedures (regression and curvature evaluation) // Artificial intelligence driven solutions to business and engineering problems. ITHEA Publ, vol. 27, 2012, pp. 27-31

[Barnhart, 2003] Barnhart C., Belobaba P., Odoni A. Transportation Science // Applications of Operations Research in the Air Transport Industry, No.4, vol.37, 2003 , pp. 368-391

[Bellman, 1962] Bellman, R., Dreyfus, S. Applied dynamic programming // Amazon, 1962

[Mozgovaya, 2011] Mozgovaya K., Yablochkina M., Friedman G. Numerical analysis of the influence of predictive accuracy of the passenger demand on the efficiency of ticket sales with the overbooking // Scientific and Technical Bulletin Saint Petersburg National Research University of Information Technologies, Mechanics and Optics, №6

[Talluri, 2005] Talluri K., van Ryzin G. The Theory and Practice of Revenue Management // Springer, 2005, pp. 129-160

Authors' Information

Vladimir Averkiev – M.Sc. student, Russian Presidential Academy of national economy and public administration; Prosp. Vernadskogo 82, bld. 1, Moscow, 119571, Russia; Moscow Institute of Physics and Technology (State Research University); Institutskii per 9., Dolgoprudny, MoscowRegion, 141700, Russia e-mail: *vlaverkiev@gmail.com*

Major Fields of Scientific Research: mathematical modeling, system analysis.

Mikhail Alexandrov – Professor, Russian Presidential Academy of national economy and public administration; Prosp. Vernadskogo 82, bld. 1, Moscow, 119571, Russia; fLexSem Research Group, Autonomous University of Barcelona, 08193 Bellaterra (Barcelona), Spain; e-mail: malexandrov@ mail.ru

Major Fields of Scientific Research: data mining, text mining, mathematical modelling

Javier Tejada – *Professor of Computer Science Department, San Pablo Catholic University; Campus Campiña Paisajista s/n Quinta Vivanco, Barrio de San Lázaro, Arequipa, Perú; e-mail:* jtejada@ itgrupo.net

Major Fields of Scientific Research: Natural Language Processing, Business Intelligence

HOW TO MASTER BIG DATA

Justyna Stasieńko

Abstract. *Information is the value that is now of critical importance in building competitive advantage. However, the amount and variety of data becomes a challenge for business and IT. Both possess a large amount of information, its accelerating growth and expectations of more effective management of these data still give birth to new technical problems and system. When BI proved to be insufficient to process the data in the shortest possible time, there is Big Data, which became the best current method of sourcing, collection and analysis of data. Using Big Data can be found as an answer for questions: How to respond to the changing expectations of customers and employees? How to improve models and business strategies? How to keep up with the next innovation, emerging at a dizzying pace? And finally, how to change your business from the inside?*

Keywords: *Business Intelligence, Business Discovery, information, analysis, QlickView, Natural Analytics*

ACM Classification Keywords: *K.6 MANAGEMENT OF COMPUTING AND INFORMATION SYSTEMS - K.6.0 General* Economics

Introduction

Data analysis, or even so called data analytics, is not a new idea. Technological development caused the increase of data amount and the ways of analysing them. It is available for these organizations that have a suitable amount of data. In the past the knowledge of organization activity was enough because new solutions, such as ERP systems or operational reporting, were used in order to have an advantage over the competition. After that the emphasis was put on processing the data by analysing them what resulted in development of Business Intelligence (BI) systems. They provide an easy access to information, its analysis and sharing within the whole organization and business surrounding [Stasieńko, 2011a)],[Stasieńko, 2011b)]. They make it possible to see the whole business of the organization. Their aim is to support the effective management of the enterprise and planning the business by delivering the proper information. The group of such tools includes the systems of information resources management, reporting and analysing tools and the solutions which help to manage the efficiency. Analysing the non-structural data is a new trend. IT appears in 'text-mining' and in the analysis of the image of social media which expanded their knowledge, which is essential for making right decisions.

Personalizing the products and services is an additional element causing the increase of the amount of information. The environment created in this way includes a big amount of data resulting from processes complexity, personalizing offers and the tendency to the offers to the groups of clients (even the smallest ones). The structure and the contents of the data which is going to be analysed come from the unknown source. Another characteristic feature is a great variability of economic models, in which many organizations function. The existing analytical and reporting data environments are able to support this process but their dynamics is not enough for business workers. That is how 'Big Data' started to exist.

Big Data – the evolution of unlimited data processing

Data and its aims were always essential for the organization. A lot of tools were created which made it easier to gather and process the data. 'Big Data' made it possible to make analysis quicker and more accurately by using the data coming from different sources. 'Big Data' is a new and large sector which helps the business to find the point in large data collections. The data may come from different sources e.g. purchase and sale transactions, public data bases, video files digital photos GPS signals from mobile phones, social networks etc. The aim of 'Big Data' is to process different types of data. Data collections such as 'Big Data' are based on four 'V': volume, variety, velocity and veracity [Płoszajski, 2013]. Volume which stands for the amount is characterised by large data beginning from petabytes collections. It is estimated that about 2,5 eksabytes of information are created every day. It means that 90% of all data in the world have been created during last two years. Variety refers to different types of data and files for which the traditional data bases are not suitably adapted e.g. sound and video files, documents, text links, network loggings and geolocation data. Velocity refers to the speed of an update and the use of data essential for creating its value. Veracity – the credibility of the information used for making decisions. Using 'Big Data' is most profitable for telecommunication companies, banks, insurance companies, Internet services: Google, Facebook as well as administration and medicine. Thanks to Big Data the bank is able to predict if the particular client divorces or has children or if they listen to rap. It all may have an influence on their credibility. Insurance companies can check if their client likes extreme sports and they can change their offers. It is up to the particular company if it uses Big Data properly or not, or if it decides to improve the cooperation with the client and minimise unnecessary costs and mistakes. Internet services, which offer various services, may also use the data. It is crucial to pay attention to the regulation and to what the user allows accepting the regulation. It is worth to be conscious that the law does not allow to be free with the clients' data. It is still being discussed if it is possible to collect and protect the clients' data and at the same time not to limit the process of its analysis. The new model gathers all data that are available and its processing is not expensive and it is used for building large data bases. After that it is possible to ask questions but not necessarily. The existing methods allow to analyse and to search for various data bases in order to find an unexpected correlation. The method used by Google is a good example [Anderson, 2008]. Google algorithms offer the word that are written the most often. Google does not use a dictionary for that but it uses the recorded answers for the previous questions asked in the past. Google suggests the word that was used the most often. The same method is used to translate texts. It searches for the sentences that have already been translated. It is an example of the machine being able to learn. In Big Data the mechanism of learning machines will become the main constituent of the business models. Learning algorithms allow the companies to follow the changing market conditions and to keep the clients. They make it also possible to find new trends. It seems that 'the revolution of infinite computing' has just started. It is the result of three trends: exponential increase of measures of computer performance, the access to them and their decrease in prices. Nowadays, the data processing is the cheapest resource used for solving the management problems. Thanks to the scalability of clouds computing it is possible to connect the powers of many computers in order to face different challenges. Processing large amount of data makes value by making the information clear and more available, creating and gathering more information about transactions in digital form for better examining all activities' effectiveness, creating more precise clients niches as well as products and services which are better adjusted to them, supporting the development of products and services of other generations and carrying out controlled experiments [Płoszajski, 2013].

As the result of this revolution the company's resources became the part of its informational system. They are able to collect and process the data, to communicate, to cooperate with others, and even to adapt or react automatically to the changes in the surrounding. They may be called 'intelligent' resources, which improve the quality of processes and will create new business models. The aim is to analyse all transactions, the clients' interactions in order to shorten the waiting time for the data and to make decisions in the real time.

In 2012 Google company introduced a free product called Analytics Content Experiments. Its role is to check the content of websites by measuring, testing and optimizing them at the same time. It makes it possible to test each version of the particular website and to decide which of them is the best. Internet giants (Amazon, Google, eBay) have already used it for some time. Nowadays, technology is available also for small companies. What is more, it is also possible now to automate completely the process of making decisions without the participation of a human being.

The problem of privacy in Big Data

Using data is well-known and important in economy, nevertheless Big Data gives more possibilities. Many companies process, analyse and visualize the data taken legally from various resource. Most often the data come from own legal resources of the company. For example, banks use all the information about their clients' accounts. They possess all the information about their payments, their shopping and their transfers to the account etc. Bank gathers all this data straight away and make it available to the clients after logging on. The organizations draw conclusions and create an outline of a particular situation, or a person who helps in their activity. It can be said that Big Data is a process based on using the data and not only on collecting it. The way of collecting it seems illegal for the clients. The world of technology causes that an average Internet user leaves a lot of their traces. On one hand it is dangerous but on the other hand it gives many possibilities of personalizing the clients, or the whole systems of managing the relations with them.

Thanks to the availability of more and more data about the clients' behavior, the activities concentrated on the client make it possible to deliver the value for the client with increasing the effectiveness of the company at the same time. Better analytics can change the assessment of activities and show how to avoid these with the smallest cots returns. The proper implementation of the clients' segmentation will constitute a tool for changing the processes within the organization in order to reach the expected goals.

Big Data allows to forecast the clients' behavior and as a result to adjust the offer to their needs, optimization of logistic and marketing activities. It also make it possible to gain knowledge about the competition. It requires the analysis of data coming from various resources e.g. the rival's prices, the amount of sold products, the clients' preferences and customer loyalty program[6].

Using Big Data, the bank workers may estimate the credibility of borrowers. They collect and analyse the information about the clients from social services, any systems of marketing information, the clients' data bases or by installing cameras and microphones in institutions. It makes the clients afraid and they associate Big Data with the invigilation and collecting data which are used for their use without any scruples. If some organizations, such as employment services, ZUS, IRS, worked with each other in correlation, it would be possible to gain information who was unemployed, who worked, who paid the taxes and how many times they have changed their employee. This knowledge would help to manage the human and financial resources better. There are also some units in big cities (e.g. New York) which work on Big Data. They improve the effective management of the city: the traffic, security services, reactions in critical situations and

[6] Amazon- the biggest market's player - using these solutions.

make it easier for the habitants to make decisions connected for example with buying the flat. They can check the buildings in every respect: redecoration, technical state etc. before they buy it or start to live in it. Checking the frequency of choosing the given words on a particular topic (e.g. a flu epidemic) it is easier to forecast something than by gathering the data from other sources (e.g. the data coming from hospital reports).

The constituents of Big Data

Gathering the data only does not guarantee a success. The data must be used skillfully. This process is supported by BI tools, which are able to aggregate the data from different sources (Oracle) or the programs analysing the data structures (MongoDB, Cassandra, Hadoop). Many of new services, which allow to collect and analyse the data, work in cloud computing. It allows to save the costs connected with creating own facility. Nevertheless, a well-qualified computing staff is indispensable here to operate the software and analysts who will create the models of analysis. Building the models and algorithms is the last part of Big Data analytics. Big Data systems used for analysing large volumes enrich the existing databases with new functions available for the users. Their germs exist in organizations in form of scattered repositories created directly by analysts. They should be provide with high performance computing solutions in order to be profitable. They can help to utilize their capability to create new business solutions. IT may successfully deliver this quality creating an unique bridge between technology and business by extending the existing data bases.

Nowadays, IT plays an important role in the organization because it takes part in building a competitive advantage by creating and adjusting the computer systems. Creating the computer system based on Big Data conception should be done taking into account the following rules. The analytic system should:

- response quickly to the questions and make analysis improving the analysts and designers' job,
- be effective and flexible as far as delivering large amount of data goes as well as enriching and making connections between them (data explorer),
- possess some analytical functions used by business analysts,
- make it possible to create interactive analyses delivered also to portable devices (designer analysts).

Taking into account technological solutions, there are tools available on the market which possess the data in different way:

- in-memory providing better quality and shorter time of calculating;
- in-database optimizing the use of the equipment power for analysing and processing the data;
- grid computing – processing the data by many computers at the same time.

The solutions described below are based on processing the data in-memory.

Big data is focused on complicated algorithms used for processing the data in huge processing centers by using highly efficient network servers. Their main role is to solve complicated calculation problems of academia, public administration as well as companies and corporations. People creating such algorithms are called the data researcher. The classic algorithmic model search through all the data in order to find connections between them. Business users prefer asking questions adhoc in order to take proper business decisions.

Fig.1 shows the process of data flow from the source to the form prepared on the angle of analyses and presentations. The raw data comes from various sources. Big Data possess the technical data from devices (e.g. web logs, server logs), transactional data (data coming from sales systems) as well as data from cloud computing (data from social services). Such information often is unstructured (series of pictures or signs) or partly structured (logs with timestamps, IP addresses or other more detailed information). Big Data definition suggests that this data is high volume (given in terabytes or petabytes), high growth (given in gigabytes or petabytes) and it has a very high level of localization dispersion (many different databases and applications generating the data in their own formats).

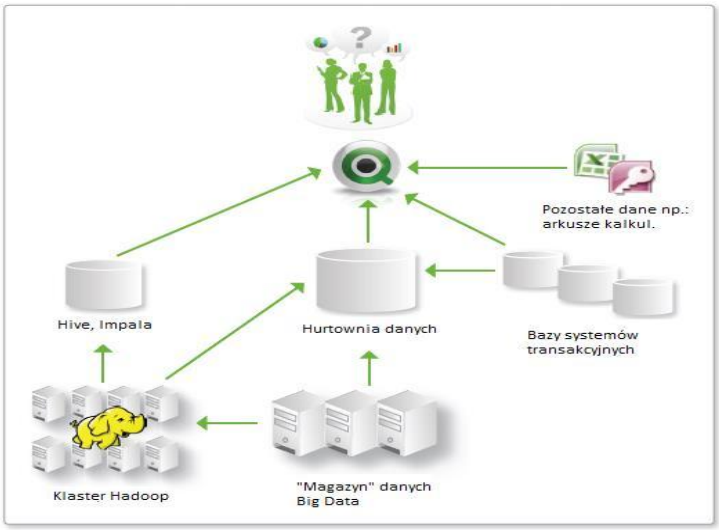


Fig.1. The data flow from the source to the analytical system using QlikView system.
Source: [Pastuszek, 2013]

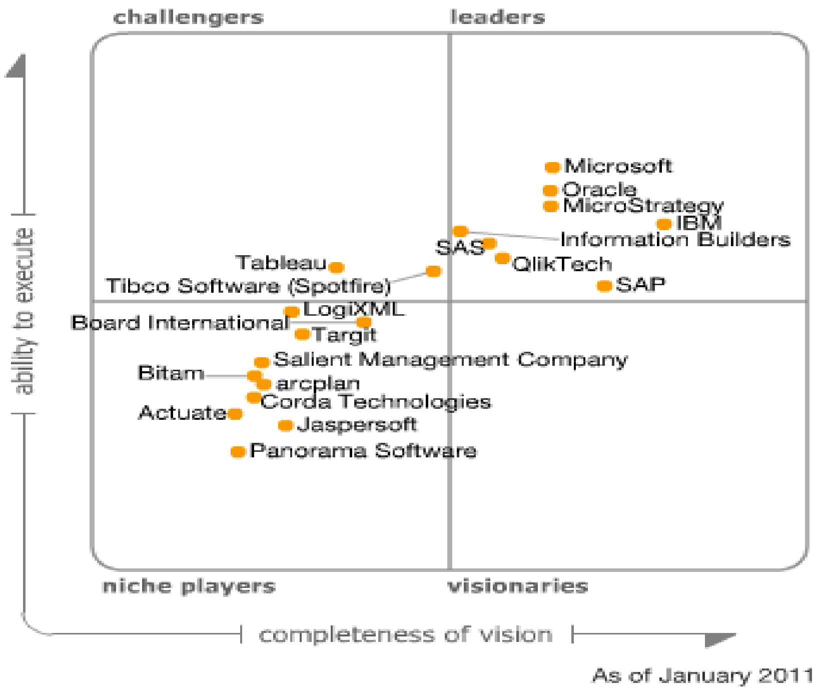
During the first stage of processing, if the costs of storage are high, the data is copied to Hadoop. It makes it possible to process, manipulate and aggregate the data at the same time. It is the first stage of interpreting the raw data. In the following processes the organizations use data warehouses as a central repository of unstructured data used in analyses. The data in warehouses come from Storage Area Network or Network Attached Storage) as well as from Hadoop clusters. The data in warehouses is organized and as a result it is easier to use it. Data analysis is the last stage while taking the right decisions. The current tools used for analysing should integrate the data from different sources (e.g. QlikView) regardless of what their format and origin are.

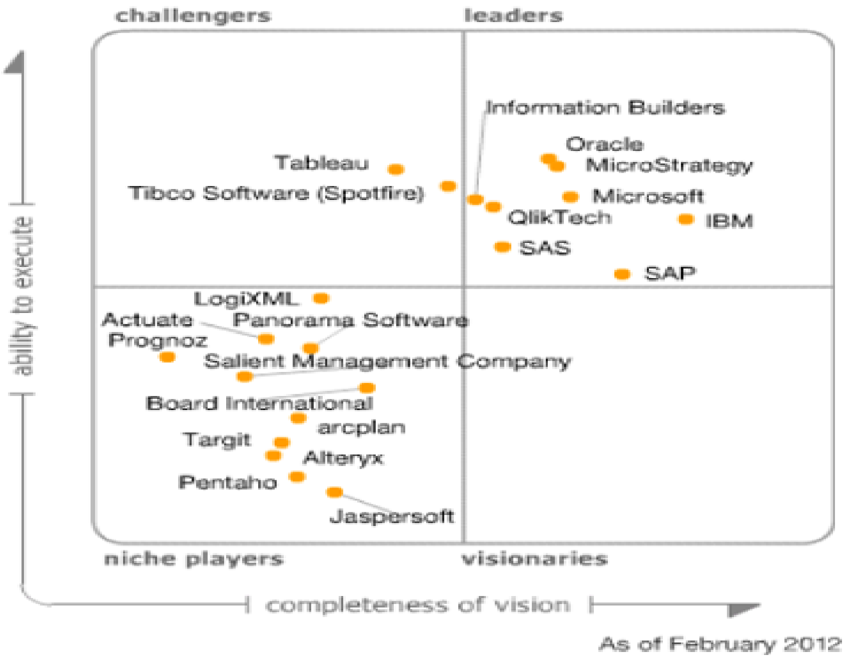
The business users require more and more as far as the access to the data, its search and analysis go without using complicated tools. Two main ways of using Big Data sources can be distinguished. The first on is when the most important data is in Big Data repository or in one repository.

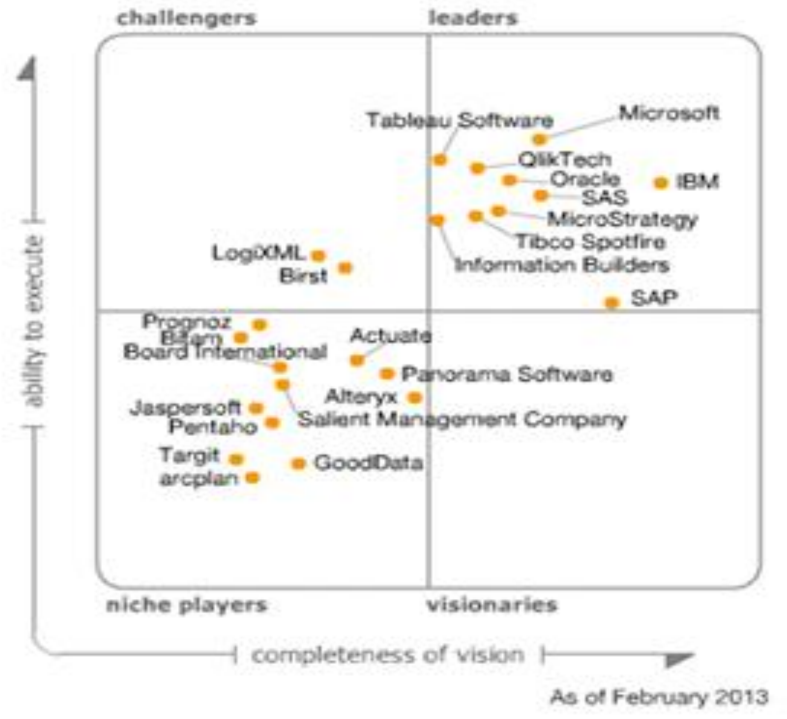
Big Data as a BI tool

Most of software potent have such a software which is used for carrying out the business analyses [Stasieńko, 2012]. So far, the market has been dominated by tools based on traditional solutions. Nowadays, the potent in IT (Microsoft, IBM, SAP, Oracle) try to include in-memory solutions in their offers. They usually create the hybrid tools which are to combination of BI and BI in-memory. On the other hand, BI products offered by these potent are getting less important as the solutions treated as the standards of analytic platforms in organizations.

The evolution of computer systems is strictly connected with the changes of consumers' behaviors e.g. these connected with using portable devices, Internet services, online shops etc. That is why, there is a need to use BI solutions in everyday life. A good example is a Social Media BI platform which process the data left in the Internet by the Internet users. Fig.2 shows that for the last 4 years some of the companies such as Microsoft lost their leader position, and others such as QlikView and Tableau improved their situation. QlikView takes the first place among other suppliers responsible for carrying out the business analyses on their own. It has been four years since it is in the bottom square (Fig.2) among the leaders creating the software Business Analytic. Tableau has moved from one quarter of the BI magic square to the other.







Source: Gartner (February 2013) ★

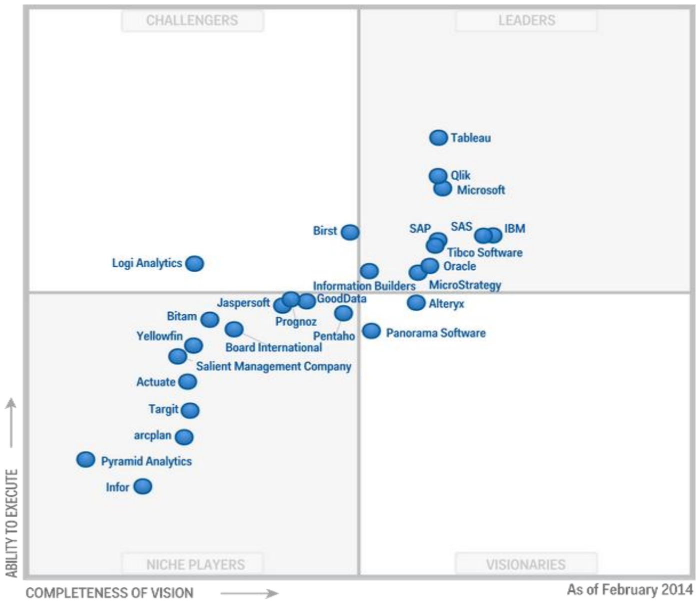


Fig.2. The magic square for BI and the analytical platforms according to Gartner[7] *:*

rok 2011 [http://bi.pl/publications/art/59-magic-quadrant-gartnera-dla-platform-business-intelligence-na-2011-rok]

rok 2011 [http://bi.pl/publications/art/59-magic-quadrant-gartnera-dla-platform-business-intelligence-na-2012-rok]

rok 2013 [source: Gartner (February 2013); http://www.sybico.pl/index.php/raport-gartnera-magiczny-kwadrant-dla-business-intelligence-i-platform-analitycznych/]

rok 2014 [source: Gartner (February 2014); http://eliasgagas.com/2014/02/25/business-intelligence-tools-magic-quadrant-by-gartner/

[7] Gartner's report shows a cross-sectional picture of the activities of suppliers in a given market segment to help end users make the best decision when choosing a partner or supplier of services or products.

QlikView – the pioneer of business analyses

QlikView is a program used to analyse data easier. It also plays an important role in Big Data implementation. It provides a quick and elastic system of presenting the data and the ability to integrate the data from different sources (Hadoop repositories, data warehouses, local data bases, spreadsheet) in one integrated solution. The most important thing is that at proper time a particular person receives proper information [Stasieńko, 2012].

QlikView offers the servicing of large amount of data. Thanks to that the client may get the best factor cost in relation to data volumes and the speed of processing it. Despite of a huge progress in technology of creating hard discs, the capacity and the access time of RAM memory are still better than other discs. That is why, if it is necessary to have 'on demand' analysis the ideal situation is to load the data into the memory. Analytical application QlikView may work on a particular data volume with information granulation which is indispensable for providing a proper level of analyses accuracy [Stasieńko, 2011], [Stasieńko, 2012]. It is possible thanks to the data compression equipment, multi-tiered architecture, the servers service and incremental loading of the data. The data loaded into the memory by QlikView is compressed up to 90% and thanks to the advanced algorithms of combining and searching the data, its processing is very effective. The solutions used recently make it possible to use great amount of memory in order to develop in-memory systems. Multi-core processors and servers make it possible to increase the computational power in a very cheap way. The servers may be divided according to their tasks. The server of the lowest level is responsible for extracting the data from the source systems. The server responsible for processing the data use the extracted data, load and process it. In QlikView there is a possibility to use the cluster of servers in order to process the data. It can also be configured in such a way which allows to take only the data which has changed from the last time. The second variant, as far as the amount of data goes, is QlikView Direct Discovery. It is a hybrid solution which combines processing the data in memory with the data loaded in currently. That is why it is intended to process large amount of data. It makes it possible to ask the BI sources without the complicated process of extracting, transforming and loading the data. The hybrid QlikView Direct Discovery attitude, makes it easier for the users to have an access to the information at the same time they need it. The users may look through the information and do not notice the difference between the data in-memory and in Big Data repository. Direct Discovery efficiency is strictly connected with the efficiency of the repository.

Another product of QlikView – QlikView.Next is an implementation of BI future. It is a complete platform of Business Analytic surrounded by a full ecosystem of people. The software and services which will satisfy the requirements of the modern organization in the realization of its strategic vision. The substructure of the platform is the conception Natural Analytic. It is based on a natural ability of human's brain to process the complex information. It uses the intuitive way of processing the information. It allows to examine the complicated data and discovering the relations between them. It is based on pairing and comparing. Natural Analytics helps to move from one data element to creating the connections with other elements. It gives potential answers intuitively, discovering new and unexpected connections between the data thanks to the cooperation and Data Dialogs. It initiates interactive discussions which refer to the data and processes in the real time. Data Dialogs helps to reach an agreement between in situations when many people work together. Thanks to Natural Analytic, BI is a real social tool. It also helps to show the connections between the data and to discover the opposed points of view while making decisions.

QlikView.Next is a new platform with a huge, associative mechanism of searching. The first clients may use the platform since 2013 but with the limited availability.

Tableau – the platform of the future

Tableau is a quite new tool which seems to be the leader on Business Intelligence market nowadays. The tools, created and developed by Tableau Software - the American company, are often called the precursor of new trends in BI solutions.

Tableau allows to quick reporting and visualizing the data from many sources. It is done by using 'in-memory' technology, which gives the possibility to analyse huge amount of data very quickly with the access to other 'on-line' sources. Tableau allows to create the data visualization automatically which are helpful while searching interesting information with the use of many dimensions (Fig. 3 and Fig. 4). This attitude, many areas of data can be analysed at the same time.

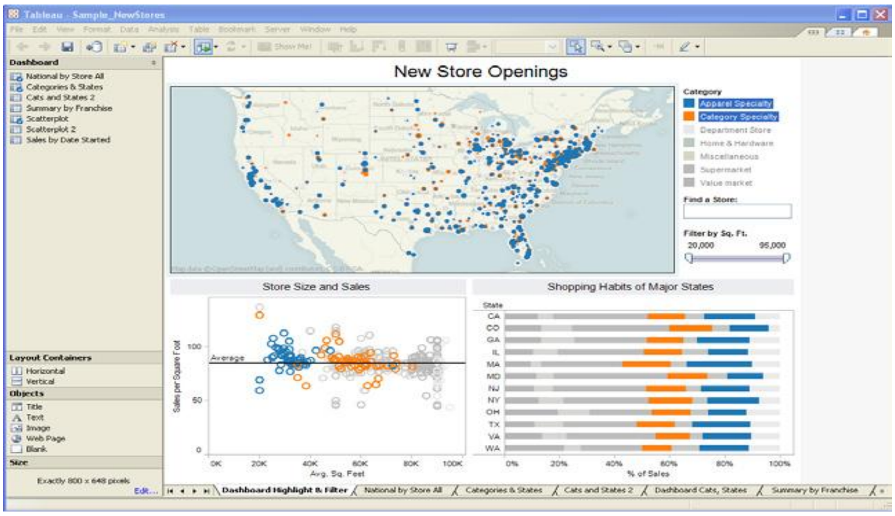


Fig. 3. Data analysis in Tableau application. Source: Own elaboration.

Tableau Business Intelligence took an advantage over other BI solutions as it is presented in Gartner's Report Fig. 2. It was done because of the intuitive solution for the user, not for the programmer, functionality adjusted to the changing needs of the users, easier Ad-hoc analyses with the use of a mouse, an easy data gathering and publishing it in the Internet.

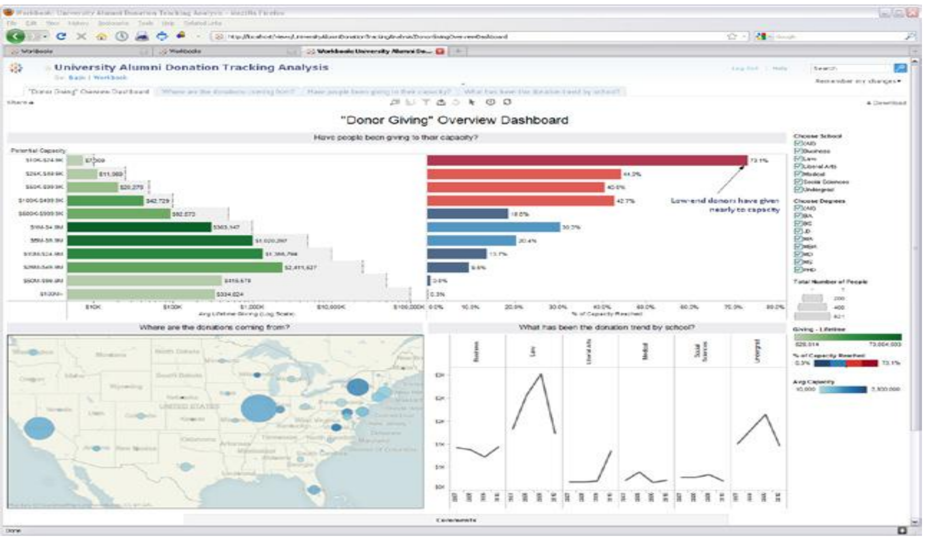
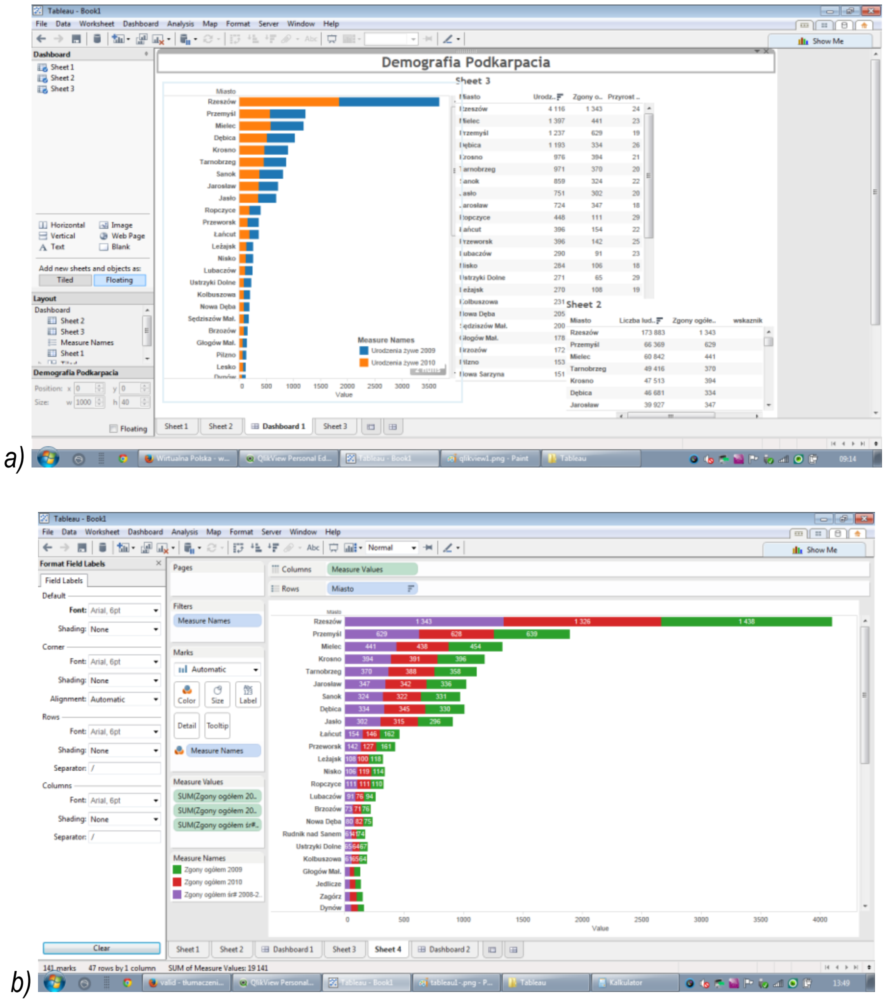

University Alumni Donation Tracking Analysis
"Donor Giving" Overview Dashboard
Have people been giving to their capacity?
Potential Capacity
Low-end donors have given nearly to capacity
Where are the donations coming from?
What has been the donation trend by school?
Choose School
Choose Degrees
Total Number of People
Giving - Lifetime
% of Capacity Reached
Avg Capacity

Fig. 4. Publication of data analysis from Tableau application into web browser. Source: Own elaboration.

There are many benefits for the company because of the implementation of Tableau Business Intelligence. First of all, the most important thing is the coherent and comprehensive information about what is going on in the company and about many ways of presenting the information from various data sources as well as from the scattered systems in the company. The identification of the previous unknown data is done in a very short time. All users are able to create and make available the analyses. The information is presented in a visual form (graphics, navigation desktops, reports). What is more, Tableau BI shortens the time and resources needed for reports and analytic works. It reduces the costs of supporting and implementing and it also minimize the dependence from IT. This application has low requirements as far as the equipment goes but it has a very high ROI coefficient.

Both applications are comparable as far as their use goes. They use in-memory to process large amounts of data. They are intuitive in the use. Both applications show only this data which are interesting for the client. In Tableau the Business Intelligence platform is based on web browser. Both applications do not require data warehouses. Fig.5 and Fig.6 show the visualization of the same data according to the same question. They load the data, which will be used by the client, intuitively into the application. The document format does not matter here. The interesting data is marked easily and quickly. The advantage of QlikView is the way of starting off the application. Those who use the spreadsheets may find Tableau even better. What is more, there is also the possibility to present the analyses in the Internet via dashboards.



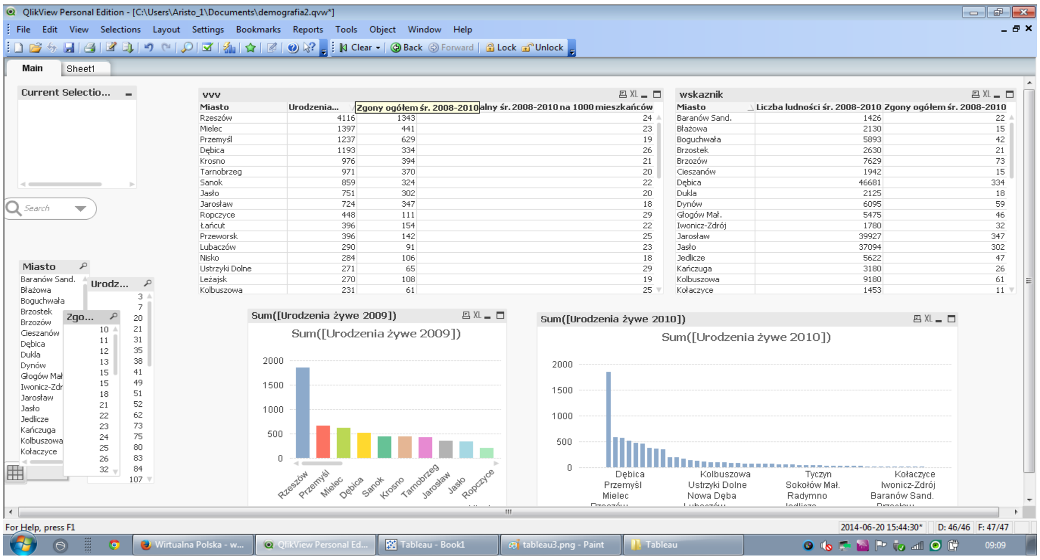
Sheet 3

Miasto	Urodz..	Zgony o..	Przyrost ..
Rzeszów	4 116	1 343	24
Mielec	1 397	441	23
Przemyśl	1 237	629	19
Dębica	1 193	334	26
Krosno	976	394	21
Tarnobrzeg	971	370	20
Sanok	859	324	22
Jasło	751	302	20
Jarosław	724	347	18
Ropczyce	448	111	29
Łańcut	396	154	22
Przeworsk	396	142	25
Lubaczów	290	91	23
Nisko	284	106	18
Ustrzyki Dolne	271	65	29
Leżajsk	270	108	19
Kolbuszowa	231		
Nowa Dęba	205		
Sędziszów Mał.	200		
Głogów Mał.	178		
Brzozów	172		
Pilzno	153		
Nowa Sarzyna	151		

Sheet 2

Miasto	Liczba lud..	Zgony ogóle..	wskaznik
Rzeszów	173 883	1 343	
Przemyśl	66 369	629	
Mielec	60 842	441	
Tarnobrzeg	49 416	370	
Krosno	47 513	394	
Dębica	46 681	334	
Jarosław	39 927	347	

Fig.5. Presentation of data analysis in a dashboard (a) spreadsheet (b) Tableau application Source: Own elaboration.



vvv

Miasto	Urodzenia...	Zgony ogółem śr. 2008-2010	alny śr. 2008-2010 na 1000 mieszkańców
Rzeszów	4116	1343	24
Mielec	1397	441	23
Przemyśl	1237	629	19
Dębica	1193	334	26
Krosno	976	394	21
Tarnobrzeg	971	370	20
Sanok	859	324	22
Jasło	751	302	20
Jarosław	724	347	18
Ropczyce	448	111	29
Łańcut	396	154	22
Przeworsk	396	142	25
Lubaczów	290	91	23
Nisko	284	106	18
Ustrzyki Dolne	271	65	29
Leżajsk	270	108	19
Kolbuszowa	231	61	25

wskaznik

Miasto	Liczba ludności śr. 2008-2010	Zgony ogółem śr. 2008-2010
Baranów Sand.	1426	22
Błażowa	2130	15
Boguchwała	5893	42
Brzostek	2630	21
Brzozów	7629	73
Cieszanów	1942	15
Dębica	46681	334
Dukla	2125	18
Dynów	6095	59
Głogów Mał.	5475	46
Iwonicz-Zdrój	1780	32
Jarosław	39927	347
Jasło	37094	302
Jedlicze	5622	47
Kańczuga	3180	26
Kolbuszowa	9180	61
Kołaczyce	1453	11

Fig.6. The presentation of data in one window of QlikView application. Source: Own elaboration.

Conclusion

Development of new tools and analytical methods is connected with the need to analyse the large amount of data which is being created all the time. Nowadays, each organization is obliged to look for new meaning and unexpected correlation in large data bases.

Big Data is not a cure for all the problems. First of all, the business needs must be defined and the first attempts should be made on a small scale. Many organization try to gather large amounts of data while most of them will not be used at all. Money spend on analytical tools and technologies do not guarantee the success. The key to it is the time of answering, asking the right business questions and the proper selection of data for the analysis. Big Data has just started to develop. There are still not enough books about this topic and no one has been translated into Polish. There are no effective ways of overcoming V4 and automatic structuring the files coming from different sources. Big Data is supposed to become a tool which can fulfill all the clients' needs.

The advanced methods of analysis will be used in the future to control the procedures. All institutions will control their workers in the future no matter what kind of organization they will be. It will be done with the use of Business Intelligence systems and the tools for thoroughgoing analysis of large amount of data. Big Data is going to be the most radical change that is going to take place in the future. This software is able to detect the economic abuse, corrupt activities and other things which are not accurate.

Bibliography

[Anderson, 2008] Anderson Ch.: The End of Theory, http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory [05.2014].

[Januszewski, 2008] Januszewski A. Funkcjonalność Informatycznych Systemów Zarządzania, t.2 Systemy Business Intelligence, PWN, Warszawa, 2008

[Morejjon, 2012] Morejjon M.: Qliktech, IBM provide newview of OLAP. http://www.crn.com/news/applications-os/18839582/qliktech-ibm-provide-new-view-of-olap.htm;jsessionid=dczRI7rxU7AMP3DdEVVM+g**.ecappj02 [07.2012]

[Nycz, 2008] Nycz M., Smok B. Busienss Intelligence w zarządzaniu. Materiały konferencyjne SWO, Katowice, 2008. http://www.swo.ae.katowice.pl/_pdf/421.pdf [05.2012]

[Pastuszek, 2013] Pastuszek B.: Big Data w QlikView, 2013, http://www.inmemory.bpx.pl/newsy/46-big-data-w-qlikview [04.2014]

[Płoszajski, 2013] Płoszajski P.: Big Data: nowe źródło przewag i wzrostu firm, E-mentor 3(50)/2013; http://www.e-mentor.edu.pl/artykul/index/numer/50/id/1016 [03.2014]

[Stasieńko, 2010] Stasieńko J.: BI in-memory – nowa jakoś systemów Business Intelligence. V Konferencja Naukowa Information Systems in Management – Information Systems in Management IX – Business Intelligence and Knowledge Management, Warszawa 2011, s.88-98

[Stasieńko, 2011 a)] Stasieńko J.: BI in-memory – nowa generacja narzędzi analitycznych, VII Krajowa Konferencja Bazy Danych: Aplikacje i Systemy, Zeszyty Naukowe Politechniki Śląskiej, seria INFORMATYKA, Wydawnictwo Politechniki Śląskiej, Gliwice, s.317-328.

[Stasieńko, 2011 b)] Stasieńko J.: Business Discovery – A new dimension of Business Intelligence. Methods and Instruments of Artificial Intelligence, ITHEA, Rzeszów-Sofia, 2011. s. 141-148

[Stasieńko, 2012] Stasieńko J.: BI – supporting the processes of the organization's knowledge management; 5th International Conference on Intelligent Information and Engineering Systems INFOS, Methods and Instruments of Artificial Intelligence, ITHEA, Rzeszów-Sofia 2012

[Surma 2009] Surma J. Business Intelligence – Systemy Wspomagania Decyzji Biznesowych, PWN, Warszawa, 2009

Netografia

[1] http://www.qlikview.com

[2] http://www.qlikviewaddict.com

[3] http://community.qlikview.com

[4] http://www.businessintelligence.pl

[5]http://www.comarch.pl/centrum-prasowe/aktualnosci/erp/comarch-cdn-xl-bi-start-nowosc-w-ofercie-comarch-erp

[6] http://www.sas.com/offices/europe/poland/actual/press/news2_01_13.html [27.04.2014]

[7] http://www.deloitte.com

[8] http://www.sybico.pl/?p=15 [28.04.2014]

[9] http://www.sybico.pl/ [27.04.2014]

[10] http://bi.pl/publications/art [25.04.2014]

Authors' Information

Justyna Stasieńko – 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

Major Fields of Scientific Research: Management Information Systems, Business information technology