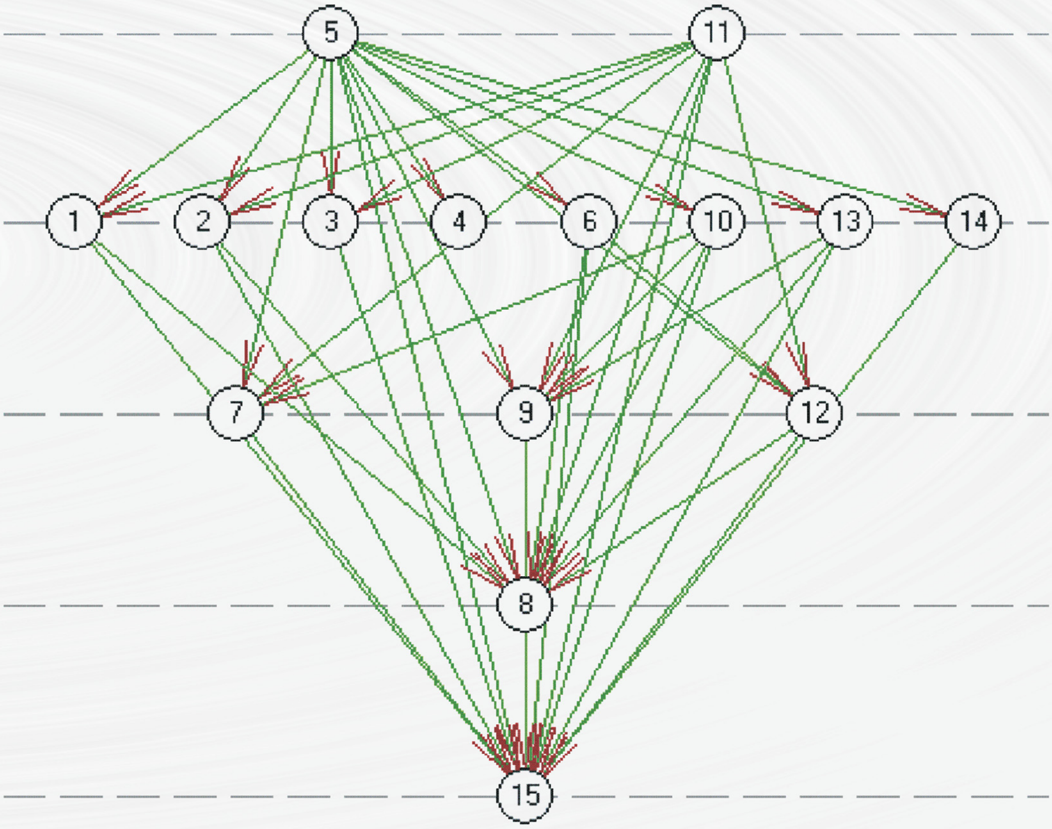
Galina Setlak, Krassimir Markov
(editors)

Transactions on Business and Engineering Intelligent Applications



ITHEA

2014

Galina Setlak, Krassimir Markov
(editors)

Transactions on Business and Engineering Intelligent Applications

I T H E A®

Rzeszow - Sofia

2014

Galina Setlak, Krassimir Markov (ed.)

Transactions on Business and Engineering Intelligent Applications

ITHEA®

2014, Rzeszow, Poland; Sofia, Bulgaria,

ISBN: 978-954-16-0068-9 (printed)
ISBN: 978-954-16-0069-6 (online)

ITHEA IBS ISC No.: 29

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 collection of papers that concern actual problems of research and application of information technologies, especially the new approaches, models, algorithms and methods of artificial intelligence to be used in business and engineering intelligent applications.

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

ISBN: 978-954-16-0068-9 (printed)
ISBN: 978-954-16-0069-6 (online)

C\o Jusautor, Sofia, 2014

PREFACE

ITHEA Publishing House is official publisher of the works of the members of the ITHEA International Scientific Society. The scope of the ITHEA International Book Series "Information Science and Computing" (**IBS ISC**) covers the area of Informatics and Computer Science. It is aimed to support growing collaboration between scientists from all over the world.

IBS ISC welcomes scientific papers and books connected with any information theory or its application.

IBS ISC rules for preparing the manuscripts are compulsory. The rules for the papers and books for IBS ISC are given on www.ithea.org. Responsibility for papers and books published in IBS ISC belongs to authors.

This issue contains a collection of papers that concern actual problems of research and application of information technologies, especially the new approaches, models, algorithms and methods of artificial intelligence to be used in business and engineering applications. Main topics which are included in the issue are:

- Decision Support and Expert Systems.
- Linguistic Technologies.
- Computer Aided Engineering and Simulation.

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

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

INDEX OF AUTHORS

TABLE OF CONTENTS

DECISION SUPPORT AND EXPERT SYSTEMS

LINGUISTIC TECHNOLOGIES

COMPUTER AIDED ENGINEERING AND SIMULATION

DECISION SUPPORT AND EXPERT SYSTEMS

THE MUCHNIK'S DIAGRAM 'SUCCESS-SUPPORT' FOR SELECTION OF MULTIPARAMETRIC OBJECTS

Anna Pashkovskaya, Mikhail Alexandrov, Yuriy Pervishin

Abstract: *The man-machine graphic procedure of multiparametric object selection is presented. The principal approach is based on reducing multidimensional problem to 2D problem, which is resolved by means of visualization. The technology includes 4 steps: (1) an expert builds a criterion ('Success') as a composition of object parameters; (2) parameters having a positive correlation with Success ('positive parameters') are selected; (3) these parameters are substituted by one generalized parameter ('Support'); (4) new coordinates of objects in 2D space of Success-Support are determined that allows to build the diagram. To calculate Support we use the first principal component of the correlation matrix related to positive parameters. The best objects are those having the largest values of Success and Support simultaneously. We demonstrate the proposed technology with the data of 784 Russian companies of mobile communication.*

Keywords: *decision making support, Success-Support diagram.*

ACM Classification Keywords: *I.2 Artificial Intelligence.*

Introduction

In this paper we consider multiparametric object selection with respect to the problem of investments. However the proposed method can be used for other applications where multiparametric objects are considered. The principal properties of object selection for investments consist in high dimensionality of the problem and in high level of responsibility of decision-makers. The former defines the necessity to use automatic methods of object selection [EC, 2008; Gotze, 2008]. The latter defines the necessity to use manual or semi-automatic methods when expert is involved in the process of solution. However this is possible if the problem dimension is not high and if visualization is convenient for experts.

These requests are resolved by means of building the diagram 'Success-Support' proposed several years ago by I. Muchnik from the Rotgers University (USA). Here: 'Success' is a scalar variable being a composition of parameters an expert selects to express his/her preferences. This variable can be considered as an integral criterion of object quality. 'Support' is a scalar variable that expresses the influence of all parameters positively related to 'Success'. This variable can be considered as an additional criterion of object quality. The principal element here is using the technique of principal components to form the criterion Support. It is supposed that the large values of both criteria correspond to the best objects. Therefore, if we build a diagram of objects using the axes Success and Support then the best objects are concentrated on the upper-right corner.

This method was firstly realized and tested in one Russian company of mobile communication. The second version of the program was developed in B.Sc. thesis [Pashkovskaya, 2013]. In this work the stability of results was explored. In this paper we improve the last version by means of additional procedures of preprocessing and post processing.

The paper is built by the following way. In section 2 we describe the algorithm in details. In section 3 we build the diagram and study stability of results. Section 4 contains the conclusions.

Algorithm

Software platform

The source information is supposed to be presented in Excel sheets. Just for this reason we developed our program on the platform Excel-VBA [Walkenbach, 2010]. We use Excel sheets only for data storage but all calculations and all interfaces are programmed in VBA. The program contains 5 separated modules presented in Figure 1:

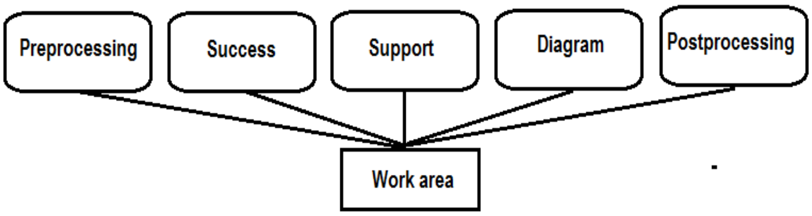


Figure 1. Program architecture

Each module is related to one Excel sheet. The interchange of information between these modules is realized by means of one work sheet. Such a way allows to modify modules independently one from the others and to provide easy intermediate control of module functionality.

Source information and module ‘Preprocessing’

In our experiments we use data from the Data Base (DB) SPARK. The data include information about annual activity of 784 Russian companies of mobile communication. Each company is presented by its 51 parameters. Therefore we have a matrix 784x51. Here is the example of parameters presented in SPARK: *Fixed assets, Gross profit, Net profit, Capital and reserves, Cash, Stock, Nominal capital,* etc.

The first a user does on the stage of preprocessing is the selection of parameters, which could be used for building Success and Support. It is a manual procedure: the user only sets indicators located near the titles of parameters. Then the user tests these parameters: whether they are suitable for further processing. For this the user implements 4 procedures of preprocessing: testing completeness, revealing outliers, testing variability, and implements normalization.

- **Testing completeness**

Completeness is measured by portion of existing data for a given parameter. A user sets a threshold for this portion. If this condition is satisfied then the parameter is marked by ‘1’, otherwise by ‘0’. Here 1 means that this parameter can be used without any correction (when all cells contain any data) and ´0´means that this parameter is not ready for further processing.

In order to recover omitted data it is necessary first of all to complete all other procedures: to reveal outliers, to test variability, and to normalize data. After thatone can recover data. The simplest solution is:

- to calculate the average value for each selected paramater having marked as the good one;
- to calculate ratios of these average values between themselves;
- to use these ratios in order to recover given parameters taking into account the other parameters of the same object.

- **Revealing outliers**

Speaking ‘outliers’ we mean parameter values, which are outside the acceptable interval. The acceptable interval is described by means of the formula [m-$k\sigma$, m+$k\sigma$], where m is an average value of a given parameter and σ is its standard deviation. It is well known that for the normal law approximately 95% of parameter values are in the interval [m-2σ, m+2σ], and for the any arbitrary law approximately 99% of parameter values are in the interval [m-10σ, m+10σ]. The latter is the consequence of the famous Chebyshev’s inequality. A user sets the critical value k and then the corresponding interval is calculated. If a given parameter has no values outside the mentioned interval then this parameter is marked by ‘1’ otherwise it is marked by ‘0’.

- **Testing variability**

The variability is measured by means of variation coefficient $\theta = \sigma/ m$, where σ and m are defined above. A user sets the threshold for the minimum value of θ. If θ exceeds this threshold then this parameter is marked by ‘1’, otherwise by ‘0’. The latter means that this parameter is a constant or almost constant and therefore it is not interesting for analysis.

- **Normalization**

In this procedure parameter values are transformed according one of the rules: a) the minimum and maximum values of a given parameter are calculated and then all values reduce to the interval [0,1] or [-1,1]; b) a user himself/herself assigns two boundary values a and b and then all values reduce to the interval [a,b]; c) the average value m and standard deviation σ of a given parameter are calculated and then all values p are transformed as (p-m)/σ.

- **Final operations**

User should implement the procedures described above according the following order: completeness=>outliers=>variability=>normalization. Then the user can resolve the problem of iincompleteness if it exists. The typical values of the threshold for completeness are Tc=10%-20%, the typical intervals of acceptable values for testing outliers are To=[m-5σ, m+5σ] or [m-10σ, m+10σ], the typical values of the threshold for variability are Tv=0,1-0,2. The indicators ‘1’ and ‘0’ can help to automate the stage of preprocessing

It is well-known that that preprocessing is the most important stage of calculations, which defines the success of all following calculations. That is why we give much consideration to this question.

Module ‘Success’

Parameters selected on the stage of preprocessing are the basis for formation of the criterion Success. It should say that user is free in formation of this criterion. Namely, he/she can:

- use all parameters or any part of them;
- build compositions of parameters and use them in Success.

An example of the composition is the well-known profitability of sales. It is a ratio of net profit to sales volume. Both net profit and sales volume are included to the list of 51 primary parameters mentioned above.

In one of our experiments we build Success as a weighted sum of 3 primary parameters: *Capital and reserves, Cash*, and *Current assets*. It can be written in the form Success = $\Sigma w_i\, p_i$, i=1,2,3. Here: w_i are weights, and p_i are parameters mentioned above. User assigns the weights w_i according his/her preferences. In particularly, in our first experiment we used: w_1=0.5, w_2=0.3, and w_3=0.2. The program allows to change these weights and immediately recalculate Success for all objects. New values of Success are reflected on the diagram Success-Support.

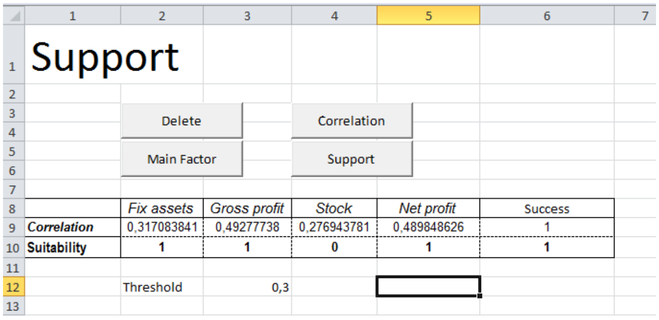
We would like to say once more that the lineal combination of parameters is not the alone possibility for the formation of Success. The criterion Success can be equal to one of the primary parameters, or to any secondary parameter as the profitability of sales, or to lineal combination of the secondary parameters, etc.

Module 'Support'

All primary parameters selected on the stage of preprocessing are the candidates to be elements of the criterion Support. The principal Muchnik's idea consists in substitution for all this parameters by one generalized parameter that could support the criterion Success. The following steps realize this idea:

1) Coefficients of correlation between each parameter and Success are calculated. Parameters having the significant positive correlation with Success are and selected. These parameters are named 'supporting parameters'. Note: we include to the list of 'supporting parameters' the Success itself as an additional parameter if Success contains more then one primary parameter.
2) Correlation matrix of the supporting parameters is built. The first principal component (the main factor) of this matrix is calculated. Therefore we have a variable reflecting a generalized relation between Success and all supporting parameters. This variable is a vector in the space of successful parameters. It is just the axis of the criterion Support.
3) Projection of the supporting parameters of objects on the vector Support gives the value of Support for this object. Therefore the projections for all objects are calculated.

The threshold for selection of parameters with the significant positive correlation is assigned by a user. The typical values of this threshold are *Tr*=0,3-0,5. Obviously, when we change this threshold we change the list of supporting parameters. With this we change the criterion Support. The library of VBA does not contain the procedure for calculation of the first principal component. So, we have to program this procedure using the well-known iteration method [Hoffman, 2001]. Figure 2 shows the part of interface related to the module Support of the program.



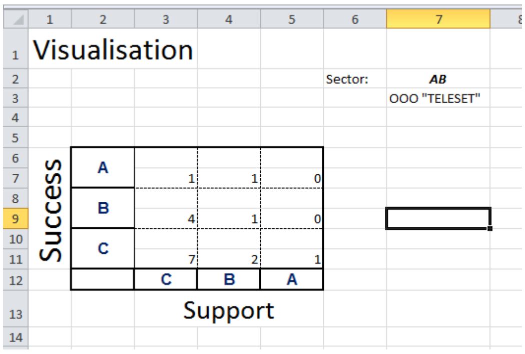
	Fix assets	*Gross profit*	*Stock*	*Net profit*	Success
Correlation	0,317083841	0,49277738	0,276943781	0,489848626	1
Suitability	**1**	**1**	**0**	**1**	**1**

Threshold 0,3

Figure 2. Interface of module Support

Module 'Diagram'

As a result of the previous stages we have the presentation of all objects in 2D space Success-Support. The visualization of these objects is the Muchnik's diagram. Usually axis X is associated with Support, and axis Y is associated with Success. To facilitate the diagram exploration we divide data of each criterion on 5 intervals marked by (A, B, C, D, E). 'A' denotes the best interval and 'E' denotes the worst interval. Therefore the best objects are collected on the right-upper corner. The part of the diagram is presented on Figure 3.



Visualisation

Sector: *AB*

OOO "TELESET"

Success			
A	1	1	0
B	4	1	0
C	7	2	1
	C	B	A
		Support	

Figure 3. Diagram Success-Support, the part of interface

Interface of the module Diagram allows to show objects related to various cells of the diagram. For example, a user can point the cell 'AB' and obtain the list of objects reflected in this cell. Here: A is the value of Success and B is the value of Support. One can see just this case in Fig.3: the cell AB contains only one object and this object is presented at cell (3,7).

Besides the diagram Success-Support the module 'Diagram' builds two additional charts: the histogram of object distribution on axis Success and the histogram of object distribution on axis Support.

Module 'Postprocessing'

Here, a user studies two questions:

- Whether the best or the worth objects on the diagram Success-Support are also the best or worth for the selected parameters.
- What parameters do not affect the selection of companies on Success-Support diagram.

For this the user indicates: a) an area on the diagram, for example, {AA}, or {AA, AB, AC}, or {AA,AB,BA,BB}, etc.; b) any primary parameter from the list of parameters selected on the stage of preprocessing. Then the program builds two histograms: 1) the histogram of all objects distributed on the axis of a given parameter; 2) the histogram of the objects from the marked area distributed on the same axis with the same scale.

The example is presented on Figure 4. Here a user considers the best objects from the area {AA,AB,BA,BB} and the parameter Capital and reserves. One can see that a) the best objects on the diagram are located in the middle part of the values of parameter Capital and reserves; b) these objects are located as one compact group.

Experiments

Plan of experiments

The object of consideration is a set of 784 Russian companies of mobile communication. Each company is presented by its 51 parameters. The following 9 parameters were selected for our experiments: *Fixed assets, Gross profit, Capital and reserves, Cash, Stock, Current assets, Non-current assets, Nominal capital, Net profit*. All these parameters passed all procedures of preprocessing without any correction. In the experiment description we use the term 'results'. Here it means the list of companies belonging to a given area from the diagram Success-Support.

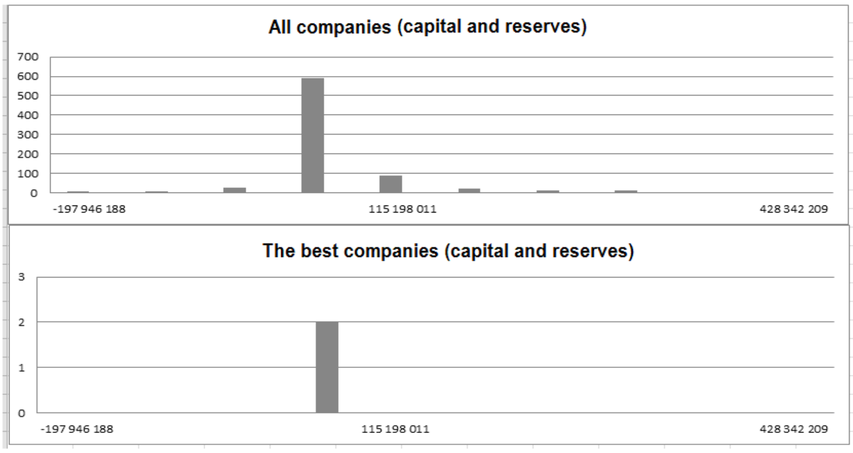


Figure 4. Distribution of all objects and the best objects with respect to parameter capital&reserves

In the first series of experiments we study the sensibility of results to the contents of the criterion Success. For this we form 3 groups of parameters with 3 parameters in each group. The weighted sum of 3 parameters forms the criterion Success.

In the second series of experiment we study the sensibility of results to the number of supporting parameters. For this we implement calculations with the different number of these parameters. We eliminate parameters having the lowest correlation with the Success.

We suppose that the most financially attractive companies should demonstrate stable rankings with different contents of both Success and Support criteria. Just for this reason we study the sensibility to the mentioned criteria.

Variation of the criterion Success

Table 1 summarizes the results of the first series of experiments with the different contents of the criterion Success. The contents of this criterion are reflected in the first column. The criterion Support is based on remaining parameters for each of the four experiments. Here each cell includes the names of companies from the diagram Success-Support.

Table 1. The best companies for different contents of the criterion Success

	AA	***AB***	***BA***	***BB***
Non-current assets, Gross profit, Capital and reserves		OAO "NKS"		ZAO "Samara Telecom", OOO “Prestizh-internet”
Gross profit, Cash, Stocks			OAO “Dagsviaz”, OAO "NKS", OOO "Teleset"	ZAO "Samara Telecom", OOO “Prestizh-internet”
Cash, Stocks, Fixed assets	OOO "Teleset"			OOO "Prestizh-internet", ZAO "IT-center"
Stocks, Current assets, Fixed assets	OAO “Dagsviaz”	OOO “Prestizh-internet”	ZAO "Centel"	ZAO "Bashsel"

Variation of the criterion Support

In this series of experiments the criterion Success is fixed. It contains *Capital and reserves, Current assets*, and *Fixed assets*. To support this criterion we consider parameters Capital and reserves, Gross profit, Fix assets, and Cash. Table 2 shows the correlation between the criterion Success and these parameters. Table 3 summaries the results of experiments with different elements of the criterion Support. These elements are reflected in the first column. Here each cell includes the names of companies from the diagram Success-Support. The letters refer to the criterion Success.

Table 2. Correlation between the criterion Success and elements of the criterion Support

	Capital & reserves	***Gross profit***	***Fix assets***	***Cash***
Correlation	0,34	0,26	0,57	0,27

Table 3. The best companies for different elements of the criterion Support

	A	***B***	***C***
Gross profit, Cash, Capital and reserves, Non-current assets	OAO “Dagsvyaz”	OOO "Prestizh-Internet"	OOO "Infolada"
Cash, Capital and reserves, Non-current assets	OAO “Dagsvyaz”	OOO "Prestizh-Internet"	ZAO "Novgorod telecom"
Capital and reserves, Non-current assets	OAO “Dagsvyaz”	OOO "Prestizh-Internet"	
Non-current assets			OOO "Prestizh-Internet" OAO «Dagsvyaz»

The Table 1 and Table 3 taking together show that the companies OOO "Prestizh-Internet" and OAO “Dagsvyaz” keep their rankings with different contents of the criteria Success ans Support. So, these companies can be considered as the most preferable ones for financial investments. This conclusion completely corresponds to the opinion of experts, which selected the most attractive companies using the other man-machine methods. Speaking more exactly these companies were in their list of preferences.

Conclusions

This article describes the original method of object selection based on visual presentations. The method allows to reconcile a subjective opinion of experts concerning the most attractive objects and the formal object description in multidimensional space of parameters. The key element of the consideration is the technique of principal components.

The method was demonstrated on the real example related to selection of mobile communication companies in Russia. The results were evaluated by experts as the very promising ones.

In future we suppose to develop the new version of the program in the environment SciLab or MatLab.

Acknowledgement

The authors are grateful to Mr. Ilya Mafter for his interest to our work and valuable advice.

Bibliography

[EU, 2008] D. Hubner (Ed.) Guide to cost-benefit analysis of investment projects // European commission, july 2008; http://ec.europa.eu/regional_policy/sources/docgener/guides/cost/guide2008_en.pdf

[Gotze, 2008] U. Gotze, D. Northcott, P. Schuster, Investment Appraisal Methods and Models // Springer, Berlin, 2008

[Hoffman, 2001] J. Hoffman, Numerical Methods for Engineers and Scientists // Taylor & Francis Publ., 2001

[Pashkovskaya, 2013] A. Pashkovskaya. Muchnik's diagram 'Success-Support' and its application to ranking modern mobile communication companies. // B.Sc. thesis, M., RANEPA, 2013

[Walkenbach, 2010] J. Walkenbach, Power Programming with VBA // Wiley Pubishing, inc, 2010

Authors' Information

Anna Pashkovskaya – *M.Sc student, Lomonosov Moscow State University; GSP-1, Leninskie Gory, Moscow, 119991, Russian Federation*

e-mail: APashkovskaja@ gmail.com

Major Fields of Scientific Research: decision-making support, world economy

Mikhail Alexandrov - *Professor of the 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

Yuriy Pervishin – *Professor of the Russian Presidential Academy of national economy and public administration; Prosp. Vernadskogo 82, bld. 1, Moscow, 119571, Russia;*

e-mail: pervishin@ ranepa.ru

Major Fields of Scientific Research: macroeconomy, mathematical modelling

TRAINING OF MANAGERS WITH DSS SVIR (SHORT REVIEW OF M.SC. STUDENT)

Kamila Yangirova

Abstract: *Decision support systems (DSS) are well-known tools for multicriteria object ranking and selection. Such systems are important not only in practical applications but also for training managers to make decisions under complex conditions. This paper shortly describes a) possibilities of DSS SVIR developped at St. Petersburg State Transport University b) personal experience in laboratory practicum for studying SVIR. This review will be useful for those who teach courses on multicriteria selection and who want to train for making rational decisions.*

Keywords*: decision support system, multicriteria selection, object ranking.*

ACM Classification Keywords*: H.4.2 [Information Systems Applications]: Types of Systems – Decision support.*

Introduction

A Decision Support System (DSS) is a computer-based information system developed to solve problems of selection and ranking. When we have only one specified scalar criterion for object evaluation then a usual optimization system can cope with this problem. But in case of complex and informal situations related to decision-making one has to use DSSs. For example, we meet with such problems when we want to find objects for investment or to select a strategy to expand market of given goods. Here complexity is a consequence of variety of attributes and variety of criteria based on these attributes. And informality is a consequence of subjectivity in integral assessment of objects. We can mention here such well-known DSSs as iThink [iThink, http] and Analitica [Analitica, http]. On the Russian market there are two known and accessible DSSs: Pilot [Pilot, http; Semenov, 2004] and SVIR [SVIR, http; Mikoni, 2009a].

Many managers, first of all the managers from large and middle companies, need tools for decision making support. However DSSs are not still as popular among practicians as various optimization systems or systems including optimization procedures. One of the principal reasons (apart from DSS price) is a certain complexity in studying such systems. To get rid of this obstacle two approaches are used. The first one consists in creating a simplified version of corresponding industrial DSS with a set of examples. These examples serve for studying DSS. Such a way is fit for independent study. The second one consists in development of laboratory practicum with corresponding manuals. This laboratory practicum should be studied step-by step independently but under the regular control of a teacher (skilled expert from a company, or a university professor).

The second way is successfully used in the St. Petersburg State Transport University where by this moment many students and Ph.D. students have defended their theses using DSS SVIR applications. They study the theory of multicriteria selection in the course of decision making support and study applications of this theory in the framework of laboratory practicum based on DSS SVIR [Mikoni, 2009b]. In the Russian Presidential Academy of national economy and public administration we study the theory of decision making as a part of the course Data Mining. Some students use in their B.Sc. theses the simplified version of the system Pilot mentioned above. This way proves to be easy for students of economical specialties. The students from the program “Two diplomas” (applied mathematics and economy) select the second way: they implement laboratory works from the laboratory practicum using DSS SVIR. Then they use this experience in their B.Sc. theses. The examples of both types of B.Sc. theses are presented in [Alexandrov, 2013; Mogilyev, 2013].

The purpose of the paper is to pay attention of managers to possibilities of DSS SVIR and to meet them with the laboratory practicum based on this system. I suppose that now DSS SVIR is one of the strongest helpers for those who teach or train to make rational decisions.

The paper is structured by the following way. Section 2 presents the methods of DSS SVIR. Section 3 shortly describes the laboratory practicum based on SVIR. Section 4 includes conclusions.

DSS SVIR

Essential characteristics of the system

The system for selection and ranking SVIR is an instrumental system designed to solve the problems of multicriteria selection. It satisfies the following requirements [SVIR, http]:

- Universality. A universal model of data representation for solving different multicriteria problems is used. Some criteria can be structured as an hierarchy with several levels;
- Ability to use together objective and subjective estimations;
- Autonomy in solving real problems of high dimensionality. The maximum number of objects and attributes is several hundreds;
- Interconnecting with other systems to share data and processing. Ability to import data from a relational database and MS Excel;
- Ergonomics. Ability to manage easily the process of decision making and analysis of results.

Decision making is complicated area. There are no samples of correct decisions for the majority of cases. The main attention should be paid to formulating a given problem, to building tree of goals and, as a result, to creating a model for object selection. Model itself should be confirmed during experiments. SVIR allows one to correct quickly the model.

To solve problems of multicriteria selection one should use several methods and compare their results. SVIR offers a basic set of methods. Cognitive graphics and post-analysis (analysis of attribute contributions, new weights for criteria, etc.) allows one to improve the final selection.

There are two basic types of problems related to multicriteria selection: optimization and classification. They are described below.

Optimization problem

Optimization methods can be divided into two classes: *vector optimization* and *scalar optimization*.

1) Vector optimization on a finite set of objects means revealing objects (alternatives), which have no other objects being better or equal than this one for all its parameters.

The system SVIR includes the following methods of vector optimization:

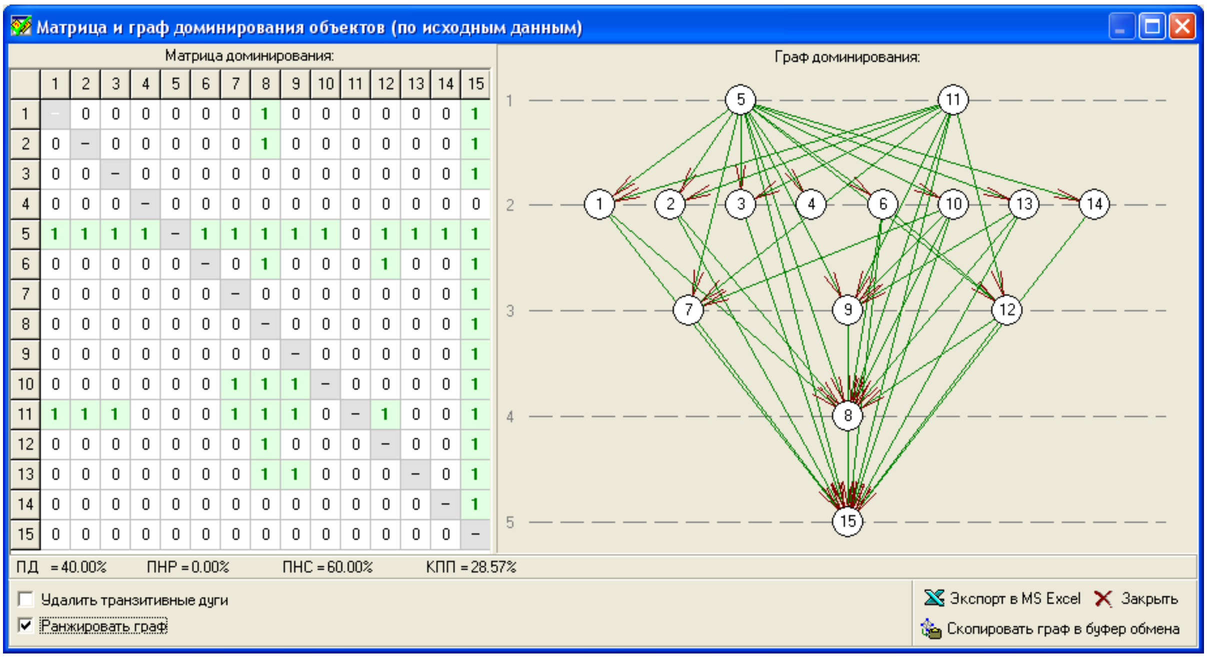
- Pareto optimization;
- Leximin optimization;
- Optimization by priority criteria (lexicographic optimization).

Pareto Optimization uses the Pareto-dominance relation. It gives preference to one object over the other one only if a) the first object is not worse than the second one for all criteria; b) at least one criterion of the first object is better. When this condition is true then the first object is considered to be dominant, and the second one-dominated. The result of Pareto optimization is a ranked graph showing the ratio of Pareto dominance on a set of objects.

Figure 1 illustrates the example of such dominance graph. This graph can be characterized by 4 parameters: dominance, indistinguishability, incomparability and order completeness coefficient. The upper level of the graph forms a Pareto set. Ranking objects by means of the number of links can be used to increase the completeness of the order. Preference is given to the elements of ranged graph that have the minimum number of entering arcs and the maximum number of outgoing arcs. Note. In this moment there are russian and english versions of interface. In this paper the russian version is used.

Leximin optimization is based on the principle of criteria indifference. The order of criteria can be changed for different objects in various ways. Estimates of each object are ordered by quality. The necessary condition for reordering is the equality of scales of all attributes, so here the normalized scale [0, 1] can be used.

Optimization by priority of criteria (lexicographic optimization) is used when the information about importance of criteria is available and estimated in a rank scale. Unlike leximin optimization now the estimates are reordered by criteria priority for all objects together (in leximin optimization the objects were ordered on descending sort and they were considered separately). Further multidimensional sorting guarantees full linear order of the objects.



Матрица и граф доминирования объектов (по исходным данным)

Матрица доминирования:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	–	0	0	0	0	0	0	1	0	0	0	0	0	0	1
2	0	–	0	0	0	0	0	1	0	0	0	0	0	0	1
3	0	0	–	0	0	0	0	0	0	0	0	0	0	0	1
4	0	0	0	–	0	0	0	0	0	0	0	0	0	0	0
5	1	1	1	1	–	1	1	1	1	1	0	1	1	1	1
6	0	0	0	0	0	–	0	1	0	0	0	1	0	0	1
7	0	0	0	0	0	0	–	0	0	0	0	0	0	0	1
8	0	0	0	0	0	0	0	–	0	0	0	0	0	0	1
9	0	0	0	0	0	0	0	0	–	0	0	0	0	0	1
10	0	0	0	0	0	0	1	1	1	–	0	0	0	0	1
11	1	1	1	0	0	0	1	1	1	0	–	1	0	0	1
12	0	0	0	0	0	0	0	1	0	0	0	–	0	0	1
13	0	0	0	0	0	0	0	1	1	0	0	0	–	0	1
14	0	0	0	0	0	0	0	0	0	0	0	0	0	–	1
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	–

ПД =40.00% ПНР =0.00% ПНС =60.00% КПП =28.57%

□ Удалить транзитивные дуги

☑ Ранжировать граф

Figure 1. DSS SVIR interface. Dominance graph

2) *Scalar optimization* is based on the transformation of vector of estimates to a number (scalar). Having these numbers one can set a linear order of objects.

For this purpose the following aggregate functions are used:

- Weighted sum of normalized criteria;
- Average geometric criteria evaluation;
- Multiplicative function with additional cofactors;
- Multiplicative function with factors in the degree of criteria importance;
- Maximin function of criteria.

Scalar optimization can be used either on values of criteria or on deviation of these values from target values. Not only the type of aggregate function is important, but also criteria significance and scale. Because this information should be determined by experts, scalar optimization in multidimensional space cannot give the only one correct result.

Classification problem

The system SVIR aids to solve the following classification problems:

1) Selection of objects that satisfy specified requirements: all objects are divided regarding specified requirements into two classes: admissible and inadmissible objects. There are two types of such selection:

- Selection of non-dominated objects (based on the ratio of Pareto dominance);
- Selection with constraints.

If the feasible set of objects is empty then it is necessary step-by-step to reduce requirements for attributes. This problem can be solved by choosing softer constraints by trial and error. Solution of the problem can be simplified by using the scalar optimization method on deviations from the target values of attributes. It allows finding the object that responds to the most to specified goal, even if it is unattainable.

2) Multicriteria fuzzy classification: objects are grouped into classes that are ordered by their quality. The classes are assigned to each attribute using intervals of their values on given scales. The condition of fuzzy membership in a class is a non-empty intersection of adjacent intervals. The measure of a class membership is estimated by the membership function, formed by an expert. Membership functions can be interpreted as fragments of the utility function. So the utility function can be calculated by membership functions. Additionally it allows to rank the objects.

Application of the system

The system SVIR can be used to solve very different problems. For example, at St. Petersburg State Transport University the following researches were conducted [Mikoni, 2009]: estimation of activity of university departments; estimation of efficiency of the railways in Russian Federation; identification of weak points in railway operations; bank lending; selecting a strategy to expand market share; distribution of cadets from the Military Medical Academy on specializations; assessing the suitability of students to study at the military department; calculation of priorities of football teams using results of the Russian Football Championship, etc.

New features of the system

The latest version of the system SVIR has the following new features [SVIR, http]: ability to use both target and constraint types of criteria; use of constraint criteria in forming the Pareto dominance ratio; the set of basic aggregate functions; subsystem for determining the boundaries of attributes scales with graphical illustration; aggregation of selection problems by preferences (Pareto set) and by constraints (feasible set); unification of utility functions and membership in classes; different ways to build classes of membership functions from the initial data.

Laboratory Practicum

Laboratory works

The basic laboratory practicum consists of five laboratory works [Mikoni, 2009b]:

1) *Designing selection model.* This work is the basis for all consequent works. Designing a selection model is the first step for solving every multicriteria selection problem.
2) *Multicriteria selection with use of vector optimization methods.* For example, this work can be used to select candidates for the job positions.
3) *Multicriteria selection with use of scalar optimization methods*. This work can be used to estimate activity of the university departments.
4) *Multicriteria classification of objects*. This work can be used to distribute students among specializations.
5) *Identification of entities priorities based on pairwise comparisons*. This work can be used to calculate priorities of football teams on the basis of championship results.

The authors of the system now plan to add new laboratory works using new features of the system listed in the paragraph 2.4. Russian and English versions of the interface are available.

Example of laboratory work

Multicriteria selection with use of scalar optimization methods.

Purpose of the work is to study the properties of scalarization methods. Here the selected groups of criteria and all given criteria are considered.

Problem setting:

1) Determining the effect from using scalar estimates in contrast with vector estimates;
2) Finding borders of scales, which provide a stable rating of objects;
3) Comparison of preferences implemented by different aggregate functions;
4) Comparison of results for the lexicographic method and scalar optimization method with additive aggregate function.

The work was completed on the example of choosing the bank for mortgage lending. Selection model consisted of 15 different objects (banks) and 15 attributes. The criteria were structured in the following hierarchy (fig.2):

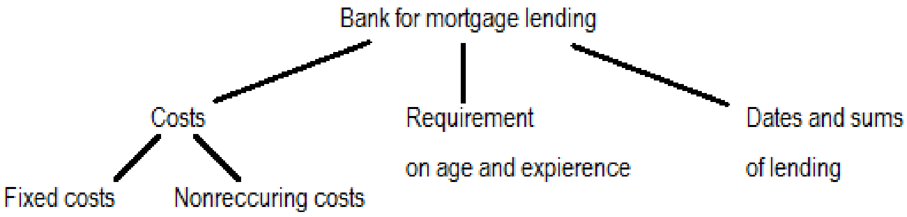
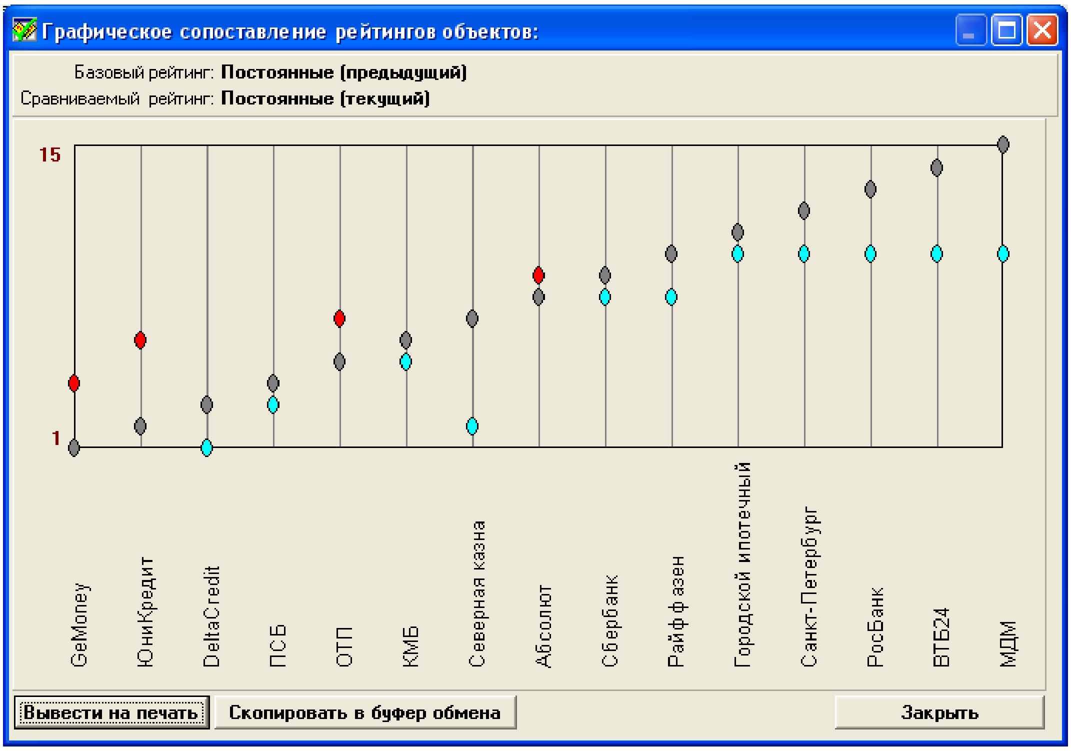


Figure 2. Criteria hierarchy

The main operations of the laboratory work were: tuning SVIR on different methods, work with the borders of attributes scales and basic aggregate functions, compare and analyse the results. On figure 3 the illustration for the third part of the laboratory work is presented.

Conclusions related to laboratory work:

1) Using scalar estimations gives the linear order because scalar estimates provide the comparability of any pair of objects.
2) If one of attribute values lies on the border of the range, the multiplicative aggregate function equals to zero. This problem can be solved by expanding the range for all the primary attributes.
3) The additive and multiplicative aggregate functions give the closest order of objects because they have the compensation property. Minimax function gives the noticeable different order in comparison with averaging functions because it reflects the extreme estimates of objects.
4) Result of lexicographic method differs from scalar optimization with additive aggregate function for two reasons: the different importance of attributes is used; lexicographic method uses not all values of attributes. The second method gives more trustworthy result.



Графическое сопоставление рейтингов объектов:

Базовый рейтинг: **Постоянные (предыдущий)**

Сравниваемый рейтинг: **Постоянные (текущий)**

Figure 3. DSS SVIR interface. Comparison of results of ranking with additive (grey color) and minimax (blue and red colors) aggregate functions

Personal recommendations

To complete the laboratory practicum one needs about 3 months:

- Literature. Reading the main chapters of textbook takes about 1 months;
- Basis for course study. It is discrete mathematics, graph theory, informatics, microeconomics, decision theory;
- Mode of implementation. Each laboratory work takes 1-2 weeks, consultations between laboratory works are required;
- Difficulties. Setting membership functions in the 4-th work requires practice and attention.

Conclusions

The paper includes:

- General description of methods from DSS SVIR;
- General description of laboratory practicum and example of one laboratory work;
- Personal experience of the author.

Official website contains all necessary information to start work with SVIR [SVIR, http]. The free trial version is available. Universities can purchase SVIR for a special price.

Acknowledgement

The author is very appreciated to Dr. Stanislav Mikoni, professor of St. Petersburg State Transport University and the author of SVIR for his numerous consultations during the practicum.

Bibliography

[Alexandrov, 2013] G.M. Alexandrov. Improvement of the company “Gefest" work using DSS Pilot. Bachelor diploma. RPANEPA, Moscow, 2013 (rus).

[Analitica, http] DSS Analitica: http:// www.lumina.com/why-analytica/

[iThink, http] DSS iThink: http:/ /www.iseesystems.com/Softwares/Business/ithinkSoftware.aspx

[Mikoni, 2009a] S.V.Mikoni. Multicriteria selection on the finite set of alternatives. Lan’, Saint-Petersburg, 2009 (rus).

[Mikoni, 2009b] S.V.Mikoni, M.I.Garina. Decision theory. Laboratory practicum. St. Petersburg State Transport University, Saint-Petersburg, 2009 (rus).

[Mogilyov, 2012] P. Mogilyov., M. Alexandrov, S. Mikoni. Multicriteria selection of perspective objects for investments: DSS SVIR vs. Muchnik method. ITHEA Publ, vol.27, 2012., pp. 61-68

[Mogilyov, 2013] P. Mogilyov. Multicriteria selection of perspective objects for investments: DSS SVIR vs. Muchnik method. Bachelor diploma. RPANEPA, Moscow, 2013 (rus).

[Pilot, http] DSS Pilot: http:// www.decisionsupporter.com/

[Semenov, 2004] S.S. Semenov, V.I. Harchev, A.I. Ioffin, Assessment of technical level of weapons and military equipment. Radio, Moscow, 2004 (rus).

[SVIR, http] DSS SVIR: http:// www.mcd-svir.ru/index.html

Author’s Information

Kamila Yangirova – *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, Moscow Region, 141700, Russia*

e-mail: kamila68@ mail.ru

Major Fields of Scientific Research: multicriteria selection, system analysis, investment climate

BAGGING ON SUBSPACES WITH BRAVERMAN 'S CLASSIFIERS[1]

Alexandra Kononova, Mikhail Alexandrov, Dmitry Stefanovskiy, Javier Tejada

***Abstract**: Non-compact and non-uniform object distributions in classes are the well-known reasons, which decrease quality of classification. To improve it we propose to use bagging on subspaces of object parameters with modified method of potential functions (Braverman's classifier). In the paper we shortly describe the proposed technology and end-user software. As an example of application we consider classification of Russian regions related to their investment attractivity.*

***Keywords**: bagging, classification, Braverman's method.*

***ACM Classification Keywords**: I.2 Artificial Intelligence.*

Introduction

Nowadays there are dozens of classification methods tested on many real examples [Bishop, 2006]. But modern approach in classification consists in combining methods instead of their individual use. Such an approach allows to perform successful classification in case of very complex data structures in parameter space. Two technologies realizing this approach are well-known: boosting and bagging.

Boosting is formation of a sequence of elementary classifiers, where each subsequent classifier corrects errors of the previous ones on learning sample. The result is one combined classifier [Schapire, 1999]. One should note that boosting uses elementary classifiers defined on the whole parameter space. The main advantage of boosting is taking into account non-compactness and non-uniformity of objects distribution in classes.

Bagging uses many classifiers, which independently assigns objects to its classes. The final decision is determined by the rule of consensus or the rule of majority. Bagging was firstly proposed in the monograph [Rastrigin, 1981], where the authors used different classifiers in the whole space of parameters. This technology were named collective recognition. Modern bagging technology was considered in [Breiman, 1996]. Here the author divided parameter space on subspaces with their own classifiers. The main advantage of bagging is the simplicity of its realization.

One should note that popular packages related to Data Mining include classification methods but this packages do not contain combinations of methods [Weka, http; Rapid Miner, http]. Boosting and bagging are included into the special package Adabag [Adabag, http]. Bagging in this package is based on the traditional method of decision tree.

In this paper we propose the technology that takes into account the advantages both boosting and bagging. Namely:

1. We use bagging on subspaces chosen by an expert. Here expert himself/herself groups parameters taking into account their mutual relations. We expect that object distributions in subspaces will prove to be more compact.
2. We use the modified method of potential functions in each subspace. Here classifiers learn individually in its subspaces. This procedure corrects errors and makes object distributions more uniform.

The method of potential functions and its modification were proposed in [Braverman, 1970]. But the latter was described without details and therefore by the moment the working version of modified Braverman's method

[1] The work done under partial support of the company Apoyo Total S.A. (subsidized by FINCYT–PERU, Foundation for innovation, science and technology)

is unknown. Also, by the moment neither Braverman's method nor its modification were used inside any bagging technology.

The paper is structured by the following way. Section 2 contains the short description of bagging technology, our version of modified Braverman's method, and developed software. In section 3 we demonstrate the results of several experiments with classification of Russian regions. Section 4 contains conclusions.

Classification Technology

Collective classification and process of decision-making

As we have already mentioned above non-compact object distribution in the space of all given parameters can lead to errors of classification. To reduce the number of errors the complete set of parameters is divided on groups, where each group forms its subspace. The classification is implemented separately in these subspaces. Figure 1 illustrates bagging in 3 subspaces (p1, p2), (p3, p4), (p5, p6). First classifier assigns object x to the class marked '1', the second one and the third one assign this object to the class ´0´ .

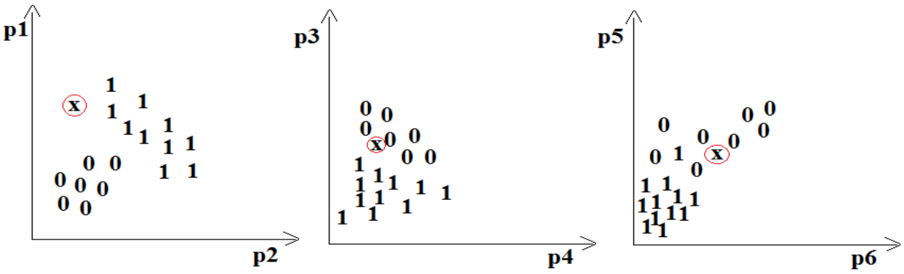


Figure 1. Bagging on subspaces

Subspace definition is implemented by an expert and here he/she tries to join together so-called related parameters. Such an approach is based on the natural hypothesis that compact classes are more probable in spaces with related parameters than in spaces with independent (unrelated) parameters.

To use bagging it is necessary to define a rule of decision-making about the status of object under consideration. Ideally, it is consensus. In this case all classifiers have the same opinion about the object class. If consensus is not reached then the decision should be based on a simple majority. Expert's preferences to certain classes can be taken into account by means of weights to be assigned to classifiers.

Braverman's method of potential functions

The idea of the Braverman's method is the following. There are representatives of each given class in training set. We use here term 'point' to name objects to be classified. Each object creates potential of its class in a test point. Different formulae can be proposed for computing the total potential of a class. Here is one of them:

$$\varphi_i = \frac{1}{n}\sum_{j=1}^{n}\frac{w_j}{(1+ar_j)}$$

where w_j is a weight of object from a training set, r_j is the Euclidean distance between this object and a test point, n is a quantity of objects in a class under consideration and α is a coefficient.

As we have mentioned above one can propose the other formulae. For example, $(\alpha r_j)^2$ can be used instead of αr_j etc. The formula contains the parameter α=1/R, where R is some typical size related to subspace. For example, it can be equal 50% of distance between the farthest objects in training set.

Figure 2 is the illustration of the method of potential functions. Point distribution is shown on the flat and potentials generated by the objects are volumetric figures.

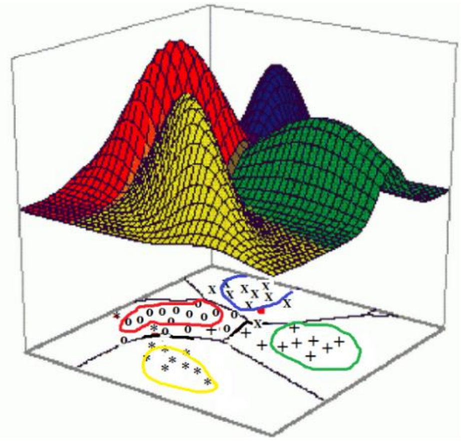


Figure 2. Potentials of classes

In real cases the distribution of points within classes is often essentially non-uniform that needs modification of the method. To do this one corrects weights using procedure of cross validation. The procedure includes 2 steps:

Step 1. Here all representatives of classes are reclassified once more. If an object is classified incorrectly then it is marked as 'incorrect object'. The other ones on this iteration are considered as 'correct objects'.

Step 2. Here all incorrect objects are considered. The nearest correct object to incorrect object increases its weight. Naturally both objects are to belong to the same class.

Step 3. If the quantity of incorrect objects does not decrease then the process stops. Otherwise the new iteration is repeated.

The described procedure looks like augmentation of objects of classes, to which incorrect objects belong. As a result we have more uniform object distribution. Graphic illustration of the method is presented below on Figure 3:

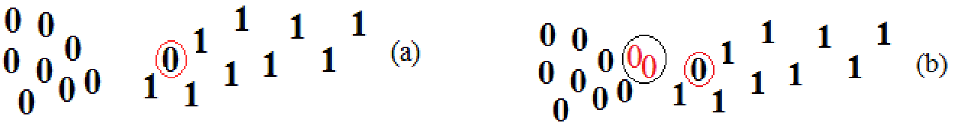


Figure 3. Interpretation of modified method

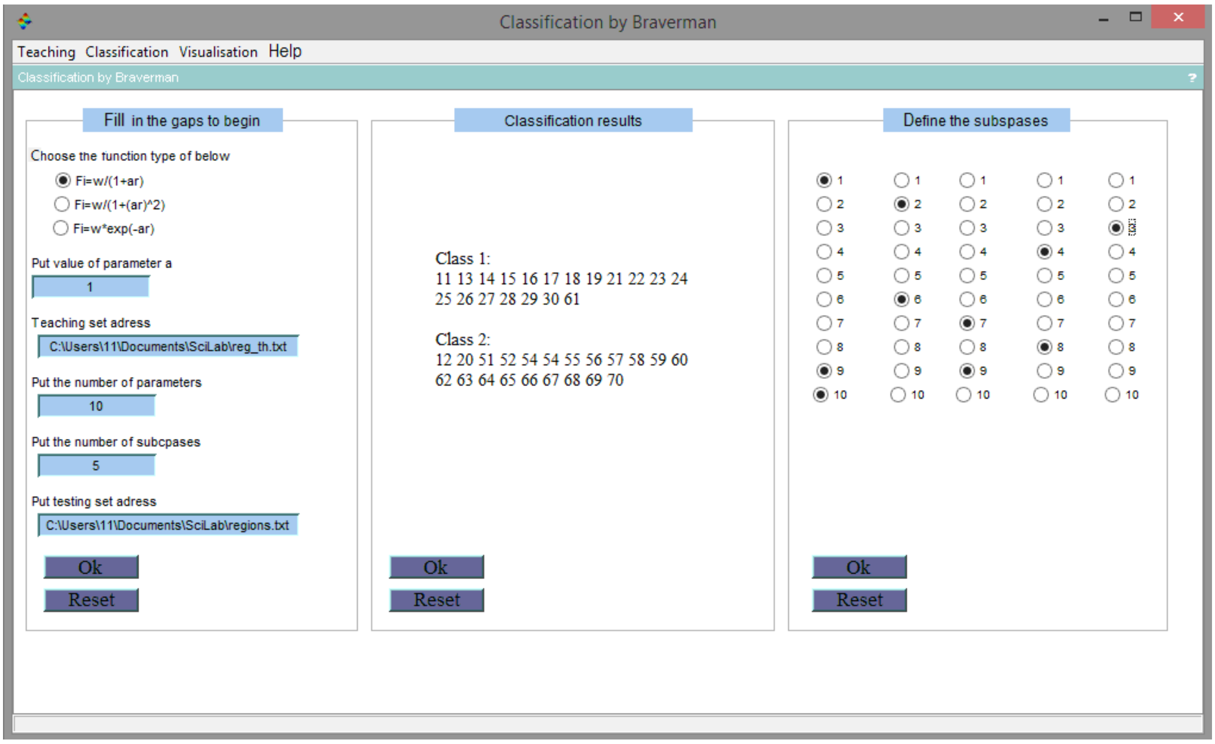
(a) two classes with the incorrectly classified object from the class '0'

(b) the nearest object from the class '0' doubles its weight

Software

The program was developped on the free-share platform SciLab [SciLab, http]. It contains all computational procedures and friendly interface. Preprocessing (normalization and outliers determination) is implemented in a separate module. User defines: number of parameters, contents of subspaces, sources of initial data with representatives of classes and objects to be classified. The program allows to use 2-10 parameters, 1-5 subspaces, 2-3 classes. The quantity of objects is limited by 10000 units. It is sufficient for many applications. Interface for subspace definition is presented on Figure 4 to the right.

Structures of input files for training and testing samples are almost similar. The only difference is: each object of training sample has the class label. The results are reflected in window of the interface. The results are also saved in output file. Figure 4 shows the program interface. It is used to manage the process of classification and to control the results


Classification by Braverman
Teaching Classification Visualisation Help
Classification by Braverman
Ok
Reset
Ok
Reset
Ok
Reset

Fill in the gaps to begin

Choose the function type of below

● Fi=w/(1+ar)
○ Fi=w/(1+(ar)^2)
○ Fi=w*exp(-ar)

Put value of parameter a: 1

Teaching set adress: C:\Users\11\Documents\SciLab\reg_th.txt

Put the number of parameters: 10

Put the number of subcpases: 5

Put testing set adress: C:\Users\11\Documents\SciLab\regions.txt

Classification results

Class 1:
11 13 14 15 16 17 18 19 21 22 23 24
25 26 27 28 29 30 61

Class 2:
12 20 51 52 54 54 55 56 57 58 59 60
62 63 64 65 66 67 68 69 70

Define the subspases

● 1	○ 1	○ 1	○ 1	○ 1
○ 2	● 2	○ 2	○ 2	○ 2
○ 3	○ 3	○ 3	○ 3	● 3
○ 4	○ 4	○ 4	● 4	○ 4
○ 5	○ 5	○ 5	○ 5	○ 5
○ 6	● 6	○ 6	○ 6	○ 6
○ 7	○ 7	● 7	○ 7	○ 7
○ 8	○ 8	○ 8	● 8	○ 8
● 9	○ 9	● 9	○ 9	○ 9
● 10	○ 10	○ 10	○ 10	○ 10

Figure 4. Program interface

Experiments

Source data

To demonstrate the proposed technology we completed experiments with classification of the Russian regions related to their investment atractivity. The initial data contains 80 regions. Each regions is described by 10 parameters. The part of data is presented in Table 1. Here all regions are ordered according their rating.

Table 1. Parameters for Russian regions

Number	***Region***	***p1***	***p2***	***p3***	***p4***	***p5***	***p6***	***p7***	***p8***	***p9***	***p10***
1	Moscow	95,2	87,4	74,8	69,1	39,7	76,2	91,5	100	84,4	61
2	St.Petersburg	74,4	79,2	74,4	57,7	57,4	56,9	81,3	93,1	72,1	58,8
3	Moscow Region	57,2	77,7	48,6	54,8	48,8	57,1	66,1	80,6	71,6	66,7
4	Tatarstan	56,9	75,6	53,1	61	60	50,7	60	45,7	53,4	53,3
5	Krasnodar region	42,9	76,1	47,4	79,7	48,3	52,1	60,2	49,8	52,6	68,5
6	Belgorod region	51,6	68,2	44,1	52,3	63	39,8	61,2	52,6	53	38,4
7	Yugra	64	74	49,5	36,8	59,5	62,1	60,9	19,7	71,9	42,4
8	…										

Here: p1 is income level, p2 is living conditions, p3 is social infrastructure, p4 is ecology, p5 is safety, p6 is demography, p7 is education and health, p8 is transport infrastructure, p9 is economy development level, p10 is entrepreneurship development level.

We deal with 2 classes, the positive and the negative ones. To form these clases and to select the training and testing sets we prepare data according to Table 2.

Table 2. Data preparation

Rating	***%%***	***Category***
1-8	10	Training set (good)
9-24	20	Testing set (good)
25-56	40	To be excluded
57-72	20	Testing set (bad)
73-80	10	Training set (good)

Classification of regions

We completed 4 experiments under different conditions to study the quality of classification. The results are presented in Table 3. In all experiments we used decision-making based on majority rule.

Table 3. Results of experiments

Experiment conditions	***Accuracy***
Classification in the complete 10D space Here we have no subspaces	62,5%
Bagging on 5 subspaces All 10 parameters are included	85%
Bagging on 3 subspaces Parameters related to economy, health and education are excluded	75%
Bagging on 3 subspaces Parameters related to safety and ecology are excluded	90%

The table shows that:

- Bagging essentially improves the results of classification in all cases. It means that our technology leads to higher compactness and uniformity of object distribution in classes.
- Compactness and uniformity are related rather to parameters reflecting economy, health and education than safety and ecology.

Conclusions

In the paper:

- Interactive bagging is developed in the form of end-user software.
- Braverman's method with the nearest neighbor correction is realized and tested.
- Proposed technology demonstrates its advantages on the real example.

In the future we plan to include preprocessing and procedures of visualization to the program.

Bibliography

[Adabag, http] Inside-R Adabag electronic resours http:// www.inside-r.org/packages/cran/adabag

[Bishop, 2006] Bishop. C. Pattern Recognition and Machine Learning, Springer, 2006

[Braverman, 1970] Arcadev A., Braverman E. Learning machine for classification of objects. Science, 1971 (rus)

[Breiman, 1996] Breiman L. Bagging predictors. Machine Learning , 1996.

[RapidMiner, http] electronic resource: http:// rapid-i.com.

[Rastrigin, 1981] Rastrigin L., Erenstein R. The method of collective detection. Energoisdat, 1981.(rus)

[Sci Lab, http] electronic resource: http:// www.scilab.org/

[Schapire, 1999] Freund Y., Schapire R, A Short Introduction to Boosting, Shannon Lab.USA,1999.,pp.771-780

[Weka, http] electronic resource: http:// www.cs.waikato.ac.nz/ml/weka/

Authors' information

Alexandra Kononova – *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, Moscow Region, 141700, Russia;*

e-mail: kononova@ phystech.edu

Major Fields of Scientific Research: data mining, classification

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

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

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

BUILDING NOISE IMMUNITY MODELS FOR GDP FORECAST BASED ON ELECTRICAL POWER CONSUMPTION[2]

Ksenia Terekhina, Mikhail Alexandrov, Oleksiy Koshulko

Abstract: *Forecasting GDP is one of the most popular applications of forecast methods in macro-economy. In this paper we build medium-term predictive GDP-models for 19 countries using Group Method of Data Handling (GHDH). The models are the hybrid ones: they include data of GDP itself and data of electrical power consumption. GMDH is realized by means of neuro- similar algorithm from the software package GMDH Shell. We studied properties of the models related to volume of teaching sample, ratio of teaching/control data, and noise immunity. The built models can be a useful addition to traditional models of GDP forecast.*

Keywords: *GDP forecast, inductive modeling, GMDH Shell.*

ACM Classification Keywords: *I.2 Artificial Intelligence.*

Introduction

Problem settings

Gross Domestic Product (GDP) is one of the most important macroeconomic indicators, which determines a place of country in the world. Dynamics of GDP is used in economic analysis, sociological prognosis, and in political battles. To build the forecast model one needs to define model parameters and to select method for building this model:

1) Predictive models usually use the values of GDP itself or other macroeconomic parameters, such as a price of exported gas or oil, volume of industrial and agricultural products related to export, etc. [Angelini, 2010]. Productions of goods and services, mineral extraction, as well as other economic activities require energy. So, it is very naturally to build GDP models using electrical power consumption as an integral variable.
2) By the moment there are many well-studied auto-regression and regression models for forecasting GDP [Kraay, 1999; Angelini, 2010]. These models are built under strong limitations concerning model structure and noise characteristics. Thus, applying GMDH technique, which allows to build models without mentioned limitations, is very promising. The models with higher noise immunity would be result of GMDH use.

These circumstances taken together define actuality of completed research. In this paper the models for GDP forecast in 19 countries with comprehensive amount of data are built and studied.

Related works

This moment there are many models for GDP forecast using indirect parameters. One of the most interesting indirect parameters is night lights. The model with this parameter was considered by Weil in 2009, the main conclusion of his work was direct correlation night lights' intensity and GDP [Weil, 2009]. Unfortunately, Weil's models proved to be invalid for Russia and some other northern countries due to territorial compression of economy and switching on business activity. Basing on the physics laws Stern claims that energy market has an impact on GDP, because the energy is required for production [Stern, 2010]. The researcher built the model of GDP using labor, capital and energy as input parameters. As for the GMDH application in economic forecast

[2] Work done under partial support of the British Petroleum grant (BP-RPANEPA 2013).

then one should mention here the large experience of Ukrainian researchers. In particularly they used GMDH for forecasting GDP and inflation of Ukraine [Ivakhnenko, 1997; Stepashko, 2010].

In section 2 of this paper the short description of GMDH and models to be built is given. Section 3 contains experimental study of model properties. Finally, section 4 is devoted to conclusions.

Tools and Models

Group Method of Data Handling

Although this method has long history it still is not well-known to researches dealing with modeling [Ivakhnenko, 1968; Ivakhnenko, 1971; Madala, 1994; Stepashko, 2013]. Particularly, the method demonstrates its advantages over the others, when:

- there is no or almost no a priori information about the structure of model and its parameter distribution,
- there are very limited quantity of initial data (data of observations) reflecting model behavior.

However, when a model or set of models are fixed, meaning have a known structure and unknown values of its parameters, and large volume of measured data is accessible, one can successfully use other approaches for model identification.

The GMDH generally follows 5-steps scheme:

1. Model class is given by an expert. Models are ordered with increasing complexity.
2. Learning data set is divided into training set and control set. Models are built on one set and tested on the other one.
3. Model parameters are determined using any internal criterion (the least squares, for example) on teaching set.
4. Model quality is determined using any external criterion (criterion of regularity, for example) on control set.
5. Model complexity is increased until the external criterion will reach its extreme.

The final model has an optimal complexity having in view a) reflection of object behavior in data of observation b) stability to unknown factors, which is titled noise.

GMDH Shell

For modeling we used software package GMDH Shell here-in-after named GS [GS, http].It is a well-known tool for time series forecast, function approximation and object classification including extended possibilities for visualization of results. GS employs GMDH technique in all its algorithms. Actual GS version includes two algorithms:

- Combinatorial GMDH;
- GMDH-type neural networks;

The unique possibility of GS is its automatic adjustment to a given data set. Namely GS itself tests each of the mentioned algorithms on a part of initial data and then selects the algorithm giving the best results. Of course, such a selection is completed with the agreement of user. In our case GS suggested to use neuro-similar algorithm which was used in our research. The comparison of algorithms from GS is presented in [Koshulko, 2011]

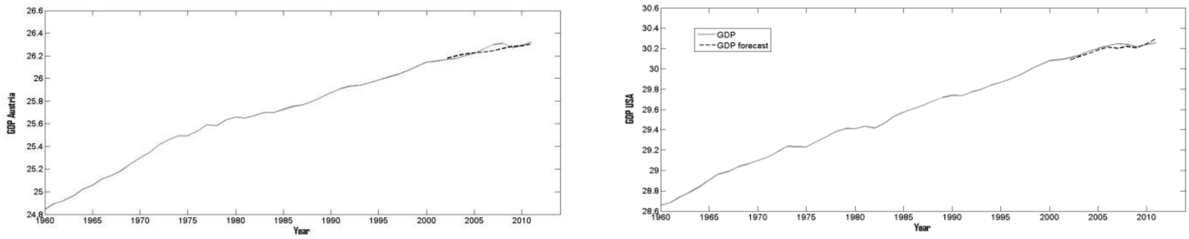
Building Models

Models to be built are the hybrid ones including data of GDP itself and data of electricity consumption. Class of models is polynomials including lags. GDP were measured in local currency and electricity consumption was measured in kW-hour. In our paper we built models for 10 year forecast in 19 countries, for which enough volume of initial data was acquired. The complete learning period was 1960-2011.

Special attention was paid on preprocessing procedures. We used:

- Scaling-1. All data were scaled using logarithmic transformation ln(1+x). After that all the data fall within the range of 20-40.
- Scaling-2. All variables were transformed using cubic root. It allowed to avoid extreme values of the variables under consideration.

The examples of models for Austria and USA are presented below in Table 1. The results of modeling are presented on Figure 1. The data about the best and the worst MAPE are presented in Table 2.



a) Austria b) USA

Figure 1. GDP and GDP forecast, solid line is raw data, dashed line is forecast.

Table 1. Model of GDP

Y1 = -3,657e^{-07}+ N4*0,489 + N3*0,510
N3 = -0,168*10^3 +[Elec(t-14)$_{\text{Austria}}$]$^{1/3}$*0,326*10^3+ N4
N4=-8,9+[Elec(t-14)$_{\text{Austria}}$]$^{1/3}$*8,776+ [GDP(t-14)$_{\text{Austria}}$]$^{1/3}$ *3,282
 	
MAPE = 0,1239
	Y1 = 4,836e^{-07} +[GDP(t-9)$_{\text{USA}}$]$^{1/3}$ *(-1,64e^{-07}) + N2
	N2 = 4,759e^{-05} +[GDP(t-9)$_{\text{USA}}$]$^{1/3}$ *(-1,624e-05) + N3
	N3 = 20,9229 + [GDP(t-9)$_{\text{USA}}$]$^{1/3}$*(-7,155) + N4*1.039
	N4 = -39,578 + [GDP(t-17)$_{\text{USA}}$]$^{1/3}$*20,529 + N5*0,213
	N5 = -61,426 + [Elec(t-5)$_{\text{USA}}$]$^{1/3}$*29,943
	
	MAPE = 0,0699

MAPE stated for Mean Absolute Percent Error. It can be calculated by the formula

$$\mathbf{MAPE} = \frac{1}{h}\sum_{t=T+1}^{T+h}\left|\frac{\hat{y}_t - y_t}{y_t}\right| * 100\%$$

where y'$_t$, y_t are data of modeling and experimental data, h is number of data for forecasting

In the represented formulas, Yi are predicted values, Ni are neurons. Each neuron is an elementary transformator of entrance data

Table 2.

	Value	Country
The best MAPE	0,039%	Australia
Mean MAPE	0,186%	
The worst value	0,499%	Turkey

All values of MAPE mentioned above refer to transformed data that is to natural logarithm of GDP data. The values of MAPE related to real data are approximately 10-20 times greater (it depends on data). In spite of this note one can see that the accuracy keeps high level having in view 10 year forecast.

Studying Models

Sensibility to sample size

We studied dependence of the forecast models quality on the length of samples. On each sample (1960-2011, 1970-2011, 1980-2011) for each country optimal prediction model for the horizon of 5 years was built. The MAPE indicator was used to compare the qualitative characteristics of the models. After the experiment with 19 countries, only 3 of 19 models were found to have the best quality characteristics in the time interval 1960-2011.

Table 3. Sensibility to sample size.

	1990-2011	1980-2011	1970-2011	1960-2011
Country	**MAPE %**			
Australia	0,0493	**0,0430**	0,1142	0,0768
Austria	0,1280	**0,0652**	0,2690	0,1038
Belgium	**0,0256**	0,1035	0,1145	0,0732
Canada	**0,0254**	0,1085	0,0602	0,1595
Denmark	0,2158	0,2327	**0,2146**	0,3361
Finland	**0,1526**	0,2901	0,1675	0,2728
France	**0,0604**	0,1167	0,1207	0,1505
Greece	0,5412	0,7309	0,5843	**0,5312**
Italy	**0,1880**	0,2401	0,2083	0,2173
Japan	0,0892	**0,0732**	0,0978	0,1262
Luxemburg	0,2185	0,9944	0,1884	**0,1824**
Netherlands	**0,0734**	1,8250	0,1486	0,2527
Norway	0,1487	**0,0961**	0,2230	0,1795
Portugal	0,1466	**0,0924**	0,1441	0,5068
Spain	0,2447	**0,2241**	0,2274	0,3312
Sweden	1,2510	0,1239	**0,1238**	0,1284
Turkey	0,5046	0,3597	0,3220	**0,2069**
United Kingdom	0,2506	**0,1385**	0,2385	0,1955
USA	**0,1013**	0,1204	0,2457	0,1461

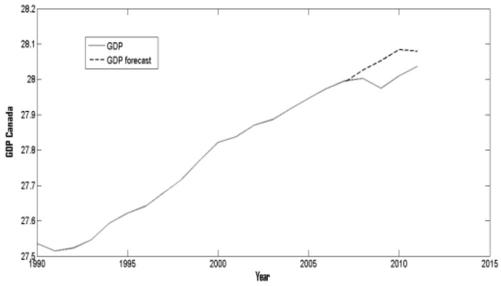


Figure 2. Model for Canadian GDP built on 1960-2011

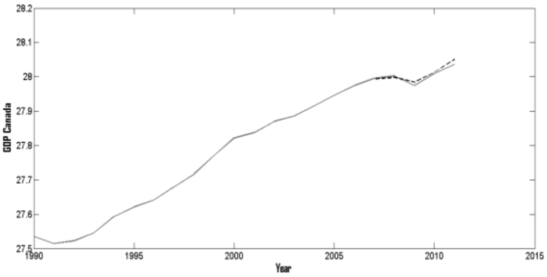


Figure 3. Model of Canadian GDP built on 1990-2011

The left graph represents model for Canadian GDP built on a sample 1960 - 2011, the right one represents model for Canadian GDP built on the sample 1990 – 2011.

Ratio training/control

According to GMDH model, parameters are determined on a training set and model quality is tested on a control set. Obviously:

- when training set is too small then model proves to be too simple,
- when training set is too large then model proves to be too complex.

In both cases the quality of the models are low. The different ratios of volumes for training set and control set in order to find the best ratio were tested. In the experiment we built models for 5 year forecast using learning set 1960-2011. The results of this testing are presented in Table 4.

Table 4. Ratio training/control

Country	40/60	60/40	90/10
Australia	0,0495	**0,0266**	0,0266
Austria	0,1901	**0,0466**	0,1247
Canada	0,1236	**0,0546**	0,1822
Denmark	0,3131	**0,098**	0,1617
France	0,1191	0,1465	**0,1091**
Norway	**0,1099**	0,1928	0,1603
Portugal	0,3463	**0,2973**	0,41
Turkey	0,6703	**0,3141**	0,4541
United Kingdom	0,191	**0,1749**	0,1791
USA	0,2507	**0,1716**	0,2541

According to the testing, the ratio 60/40 is the best one for the majority of the cases. Thus, this ratio was used in our experiments described in section 2.3. One should say that this ratio is recommended by GMDH developers [Koshulko, 2011].

Noise immunity

Noise immunity is one of the principal characteristics of models built with GMDH. Such an effect is reached when model is simplified under high level of noise. Experimental study of this effect is presented in many publications, for example [Ponomareva, 2008]. Theoretical justification is done in [Stepashko, 2008]. Our experiment consisted in the following:

Data (electricity and GDP) were taken for both developed and developing countries. These data were supplemented with gauss noise. The levels of noise were 10%, 20% and 50%. We built models for 5 year forecast on the sample 1990-2011. To compare results the values of MAPE were used.

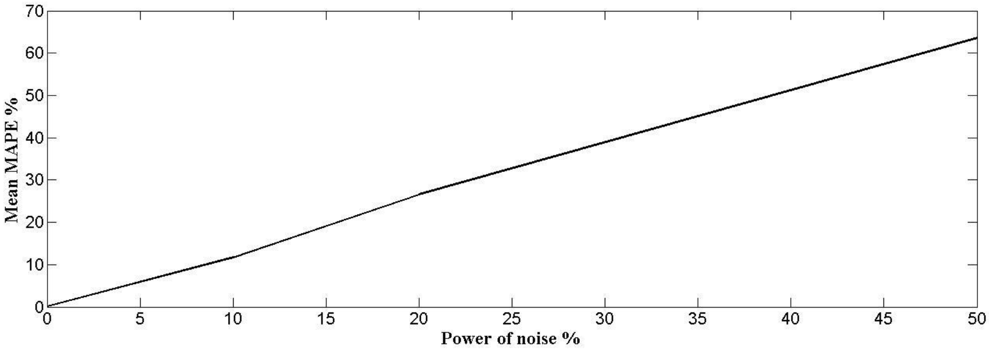


Figure 4. Dependence of the average MAPE values for 10 countries on the noise level

The Figure 4 shows the dependence of averaged MAPE on the noise level. The averaged MAPE was calculated using data of 10 countries. The linear dependence is evidence for the noise immunity of the built hybrid models.

Conclusions

In this paper we built models for 10 year forecast for 19 countries The neuro-similar algorithm was used to build models, which provided high quality of forecast. Namely, MAPE varied from 0,039% to 0,499% with a middle value 0,186 % (all values refer to the transformed data).

We studied characteristics of built models and obtained the following results:

- Long time series do not guarantee better model quality. It proves that better models refer to learning period 20-30 years for the majority of countries.
- The best ratio teaching/testing is equal 3:2. It corresponds to the well-known recommendations concerning GMDH techniques.
- All built models have high level of noise immunity that is one of the main advantages of GMDH technique.

In the nearest future we plan to build models with GMDH technique for other macroeconomic parameters.

Bibliography

[Angelini, 2010] E. Angelini, G. Camba-Mendez , D. Giannone, L. Reichlin and G. Runstler. Short-term forecasts of euroarea GDP growth. Econometrics Journal (2011), volume 14, pp. C25–C44. http://dx.doi.org/10.1111/j.1368-423X.2010.00328.x

[Kraay, 1999] A. Kraay, G. Monokroussos. Growth Forecasts Using Time Series and Growth Models. Policy Research Working Papes, 1999, p.38 http://dx.doi.org/10.1596/1813-9450-2224

[Weil, 2009] D. Weil, H. Vernon, A. Storeygard. Measuring Economic Growth from Outer Space. NBER Working Paper Series,2009, Vol. w15199.

[Stern, 2010] Stern D. I. The Role of Energy in Economic Growth. USAEE-IAEE, 2010, w10-055 http://dx.doi.org/10.2139/ssrn.1715855.

[Ivakhnenko, 1997] A.G. Ivakhnenko, G.A. Ivakhnenko, N.M. Andrienko. Inductive Computer Advisor for Current Forecasting of Ukraine Macroeconomy. http://www.gmdh.net/articles/papers/advisor.pdf

[Stepashko, 2010] V.S. Stepashko, Y.V. Koppa. Comparison of predictive properties of regression models and GMDH. Kiev, 2010. (rus)

[Ivakhnenko, 1968] A. Ivakhnenko. Group Method of Data Handling as a Rival of Stochastic Approximation Method. Journal “Soviet Automatic Control”, No. 3 , 1968, pp. 58-72

[Ivakhnenko, 1971] A. Ivakhnenko. Polynomial theory of complex systems. IEEE Trans. Sys., Man and Cyb., No 4, 1971, pp. 364-378.

[Madala, 1994] H.R. Madala, A.G. Ivakhnenko. Inductive learning algorithms for complex systems modeling. CRC Press, 1994, 368 pp.

[Koshulko, 2011] O., Koshulko. Validation Strategy Selection in Combinatorial and Multilayered Iterative GMDH Algorithms. In: Proc. Intern. Workshop on Inductive Modeling (IWIM-2011), NAS of Ukraine and Prague Tech. Univ, 2011, pp. 51-54

[Ponomareva, 2008] N. Ponomareva, M. Alexandrov, A. Gelbukh. Performance of Inductive Method of Model Self-Organization with Incomplete Model and Noisy Data. Published in Gelbukh.A., Morales, E. (Eds.): Journal of IEEE Computer Society, N_3441, pp.101-108

[Stepashko, 2008] V. Stepashko. Method of critical variances as analytical tool of theory of inductive modeling. Journal of Information and Automation Sciences, Publ.: Begell House Inc, 2008, Vol. 40, N_3, pp. 4-22

[GS, http] Program GMDH Shell, http:// gmdhshell.com

[Stepashko, 2013] V. Stepashko, Ideas of academician O. Ivakhnenko in Inductive Modeling field from historical perspective. In: Proc. of 4th Intern. Conf. on Inductive Modeling (ICIM-2013), 31-37. NAS of Ukraine, Prague Tech. University, Kyev, 2013.

Authors' Information

Ksenia Terekhina – *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: kseniya.terekhina@phystech.edu

Major Fields of Scientific Research: mathematical modeling, world economy

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

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

e-mail: koshulko@ gmail.com

Major Fields of Scientific Research: data science, parallel processing

NOISE IMMUNITY OF INDUCTIVE MODELING WITH J.FORRESTER'S VARIABLES[3]

Olga Proncheva

Abstract: *Traditional model of Forrester´s world dynamics contains 5 variables: population, main funds, capital investment in agricultural fraction, pollution and natural resources. In the paper we consider model with the same variables, which is built using inductive modeling technique. We study influence of additive and multiplicative Gaussian noise on the model and test the theoretical results concerning training on noised data.*

Keywords*: inductive modeling, world dynamics, noise immunity.*

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

Introduction

In 1971 J. Forrester was asked to develop a model of world dynamics. Speaking world dynamics we mean the dynamic interactivity of the main macroeconomical variables. Such models can predict crises and sometimes help to avoid it. J. Forrester in his work [Forrester, 1979] selected five main problems, which could provoke the World Crises. It is overpopulation of our planet, lack of basis resources, critical level of pollution, food shortages and industrialization and the related industrial growth. He tied a single variable with each of these issues. So, we have a five-level system including: population (P), pollution (Z), natural resources (R), fixed capital (K), capital investment agriculture fraction (X). This system is built on the principals of system dynamics and it is presented in the form of five differential equations named the classical Forrester´s model:

Previously we have experimentally studied noise immunity of this [Proncheva, 2014a]. We found out that multiplicative noise, which represents internal system shocks, affects a system less than additive noise that represents external shocks. Also the most sensitive variable was pollution and the most influential variable was natural resources.

In this work we study noise immunity of models built in inductive modeling technique. Here-in-after we will call such models IM-models. Inductive modeling has a long history and many applications [Ivakhnenko, 1968; Madala 1994; Stepashko, 2013]. In the work [Proncheva, 2014b] we shortly describe our experience in building IM-models with J.Forrester's variables but we did not consider noise immunity of these models.

The paper is built by the following way. In section 2 we consider the tools for testing noise immunity of IM-models. Section 3 is devoted to study noise immunity. Section 4 contains conclusions

Tools for Testing Noise Immunity of IM-Models

The classes of predictive models

The models under consideration are supposed to belong to the class of nonlinear difference equations. All data were scaled. So all the values of the variables prove to be in the interval (0, 1), i.e. have the same scale. To do this, the numerical values of population, investment capital and investment capital in agriculture fraction were divided in each year on 10^{10}, the number of remaining natural resources - on 10^{12}.

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

The final models have the following form (1 - 5):

$$P_{t+1} = g_1(P_t, K_t, X_t, Z_t, R_t, P_{t-1}, K_{t-1}, X_{t-1}, Z_{t-1}, R_{t-1}, \dots) \quad (1)$$

$$K_{t+1} = g_2(P_t, K_t, X_t, Z_t, R_t, P_{t-1}, K_{t-1}, X_{t-1}, Z_{t-1}, R_{t-1}, \dots) \quad (2)$$

$$X_{t+1} = g_3(P_t, K_t, X_t, Z_t, R_t, P_{t-1}, K_{t-1}, X_{t-1}, Z_{t-1}, R_{t-1}, \dots) \quad (3)$$

$$Z_{t+1} = g_4(P_t, K_t, X_t, Z_t, R_t, P_{t-1}, K_{t-1}, X_{t-1}, Z_{t-1}, R_{t-1}, \dots) \quad (4)$$

$$R_{t+1} = g_5(P_t, K_t, X_t, Z_t, R_t, P_{t-1}, K_{t-1}, X_{t-1}, Z_{t-1}, R_{t-1}, \dots) \quad (5)$$

In our research we studied two models. The first one contains only given variables with different powers, the second one contains additional pairwise multiplications to consider a combined influence of variables. The variables in the models can use integer, fractional, positive and negative powers.

Checking IM-model

Noise immunity can be checked in two different regimes:

- noise affects on model on the stage of forecast;
- noise can be included to initial data.

In this work we consider the first case. We check noise immunity to additive (external shock of the system) and multiplicative (internal shocks) noises. Noise affects only since 2013 year. Before this year the system dynamics is defines by (1-5).

Additive noise that affected on population was simulated by the following way:

$$P_{t+1} = g_1(P_t, K_t, X_t, Z_t, R_t, P_{t-1}, K_{t-1}, X_{t-1}, Z_{t-1}, R_{t-1}, \dots) + level \cdot \xi \cdot \widehat{P} \quad (6)$$

here:
ξ - white Gaussian noise (0;1);
$level$ - a level of noise (some fractions);

$\widehat{P}$ - an average power of population.

Dynamics of other variables is the same.

Multiplicative noise was simulated by the following way:

$$P_{t+1} = g_1(P_t, K_t, X_t, Z_t, R_t, P_{t-1}, K_{t-1}, X_{t-1}, Z_{t-1}, R_{t-1}, \dots) \cdot (1 + level \cdot \xi) \quad (7)$$

here:

ξ - white Gaussian noise (0;1);

$level$ - the level of a noise (in fractions).

Also we calculate quantitative characteristic of noise immunity. We use the next measure:

$$\sigma_i^{rat} = \frac{1}{15} * \sum_{t=2013}^{2028} \frac{1}{f_i(t)} * \sqrt{\sum_{k=1}^{1000} [\![(f]\!]_i^k(t) - f_i^{mean}(t))^2 * \frac{1}{1000}} \quad (13)$$

here

σ_i^{rat} - mean-root deviation of variable i;

$f_i(t)$ - the value of variable *i* in moment *t* in un-noised function;

$f_i^k(t)$ - the value of variable *i* in realization *k* in moment *t*;

$f_i^{mean}(t)$ - mathematical expectation of variable *i* in moment *t.*

So, the less σ_i^{rat} is, the more stable a function is.

Software

To build IM-models we used the program package GMDH Shell (GS). GS were developed by GEOS company and it covers problems of extrapolation, approximation and classification [GS, http://www.gmdhshell.com/]. Speaking extrapolation we mean time series prognosis. GS is based on Group Method of Data Handling (GMDH). It is realized in 3 algorithms: combinatorial GMDH, GMDH-type neural networks, GMDH-type decision forest. GS is very fast because of parallel processing and deep optimization of core algorithms. In our previous work we used GMDH-type neural networks [Proncheva, 2014b].

For analysis of noise immunity we use the program "Model-IM" developed in MatLab. This program includes a convenient interface to make this program accessible for end-users.

Experimental Study of Noise Immunity of IM-Models

The simple model

The forecast was made on 15 years. The best model in the class of "simple" model is:

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

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

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

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

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

Below (fig. 1) are the results of influence of 20% additive noise. There are 3 lines on the figure: thin uninterrupted line is the initial function, thick line is the forecast, and thin dotted line is the worst function. The mean-root deviation, calculated with (13), is presented in Table 1.

The most sensitive and the most influential variable were also diagnosed. The most sensitive variable reacts the most on shock of other variables. Shock of the most influential variable affects the most the other variable. The most sensitive variable is pollution, as in Forrester's model. The most influential variable is resources. It means that one should pay the main attention on these variables.

Table 1.

Variable	Mean-root deviation, %
Population	1,4%
Capital investment	1,2%
Capital investment agriculture fraction	1,0%
Pollution	1,3%
Resources	1,2%

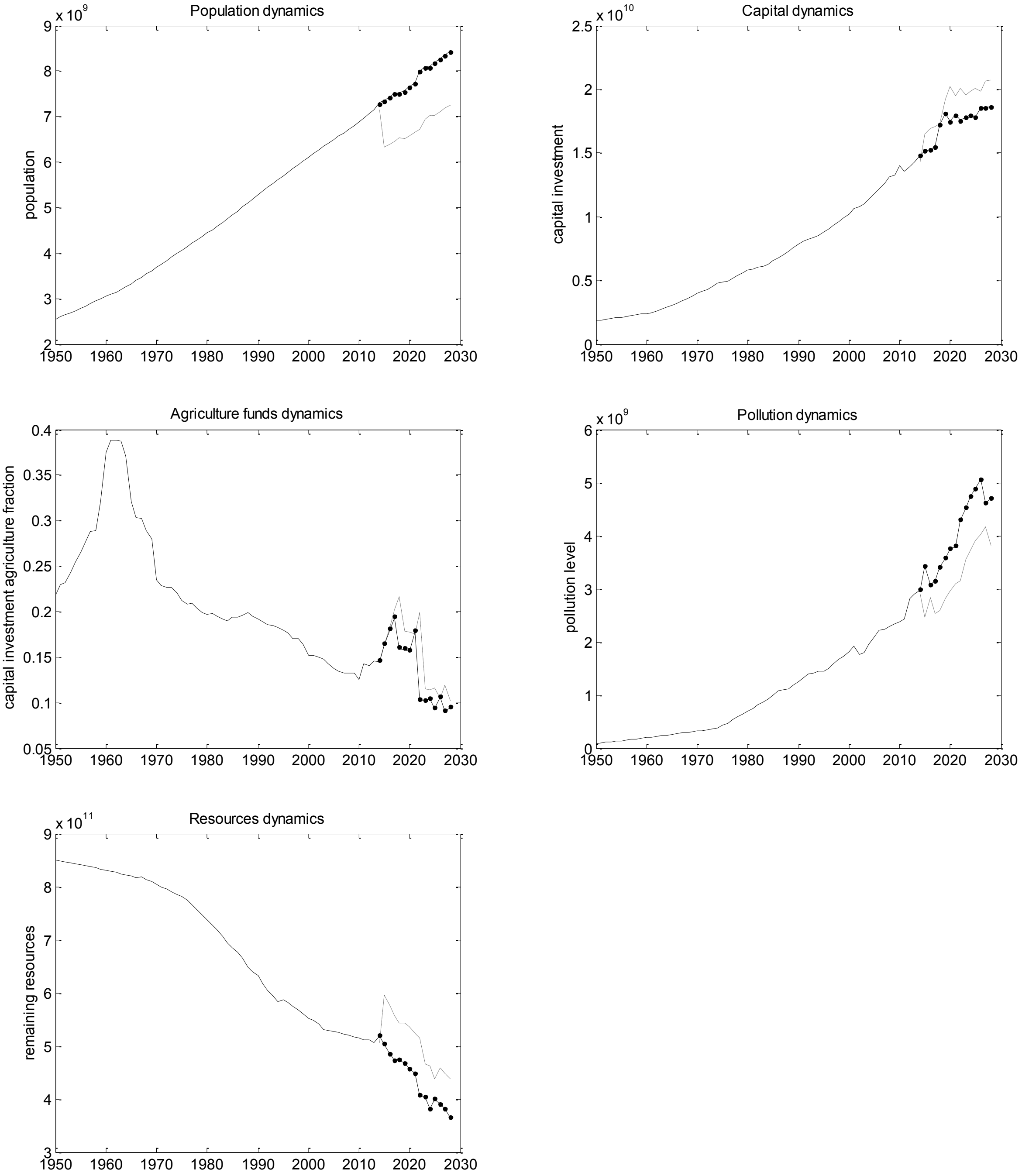


Figure 1. Influence of additive noise on the simple model

The influence of multiplicative noise was also researched. Below (fig. 2) there are the results of influence of 50% additive noise. There are 3 lines on the figure: thin uninterrupted line is the initial function, thick line is the forecast, and thin dotted line is the worst function. The mean-root deviations are presented in table 2. The most sensitive variable is pollution, as in Forrester's model. The most influential variable is resources.

So, we got the analogy with Forrester's model. Multiplicative noise affects the model much less than the additive one, and in both cases the most sensitive variable is pollution and the most influential one is resources.

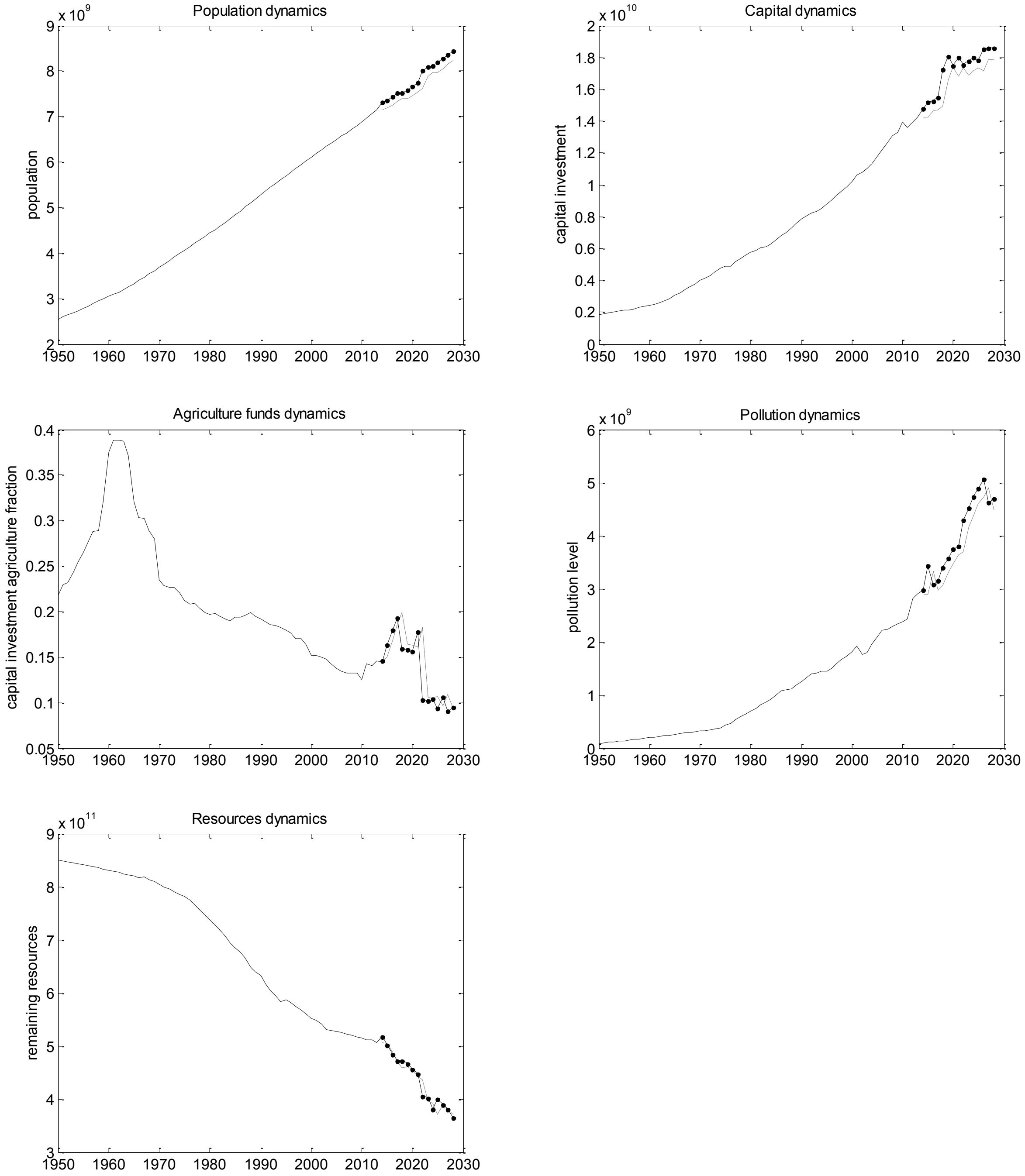


Figure 2. Influence of multiplicative noise on the simple model

Table 2

Variable	Mean-root deviation, %
Population	0,4%
Capital investment	0,5%
Capital investment agriculture fraction	0,3%
Pollution	0,4%
Resources	0,1%

The model with pairwise multiplication

The best 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-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}$

Below (fig.3) there are the results of influence of 20% additive noise on the model with pairwise multiplications. The mean-root deviations, calculated with (29), are presented in table 3. In this case the most sensitive and influential variables are also pollution and resources respectively.

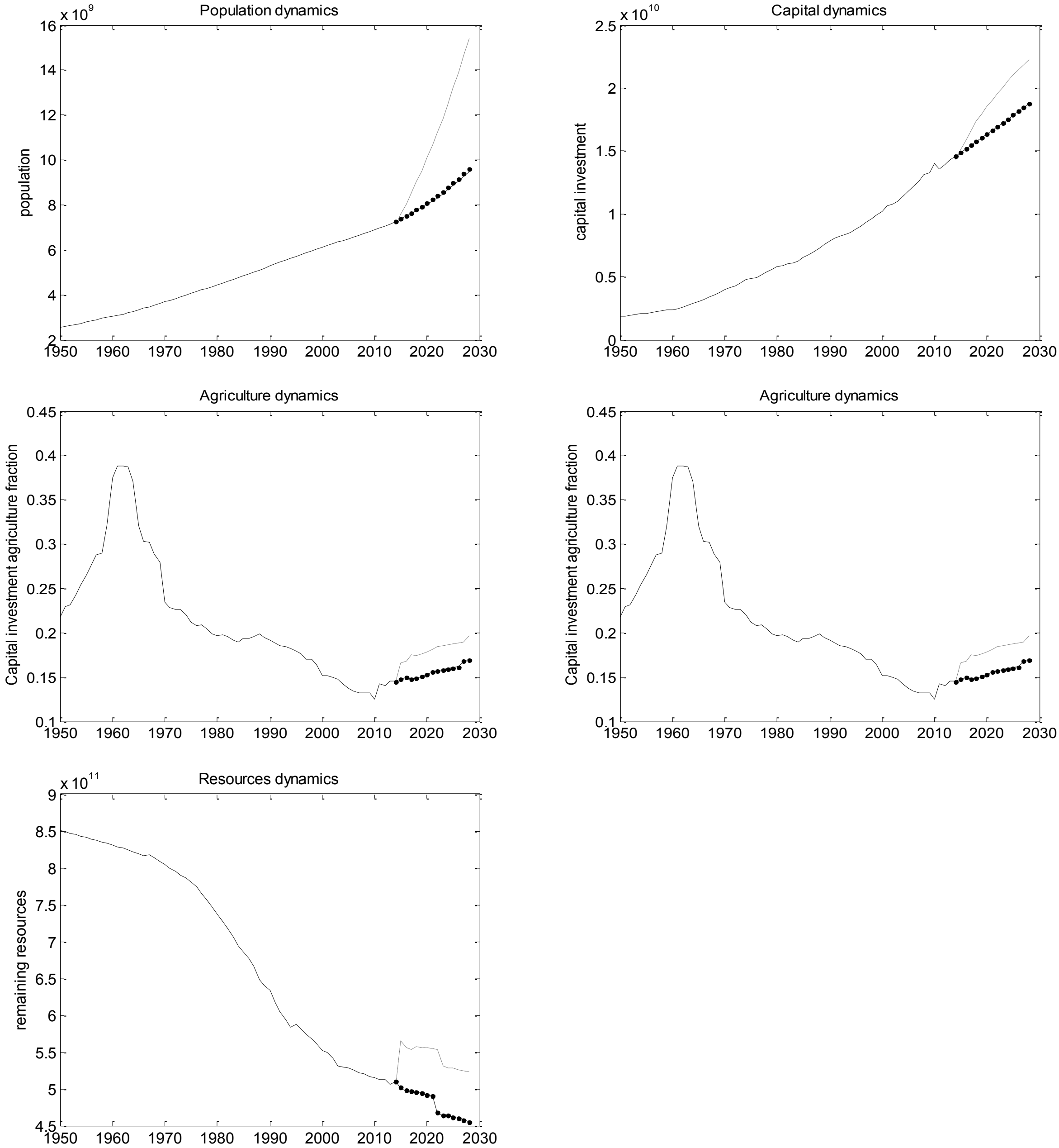


Figure 3. Influence of additive noise on the model with pairwise multiplications

Table 3.

Variable	Mean-root deviation, %
Population	2,4%
Capital investment	2,2%
Capital investment agriculture fraction	2,1%
Pollution	2,1%
Resources	2,4%

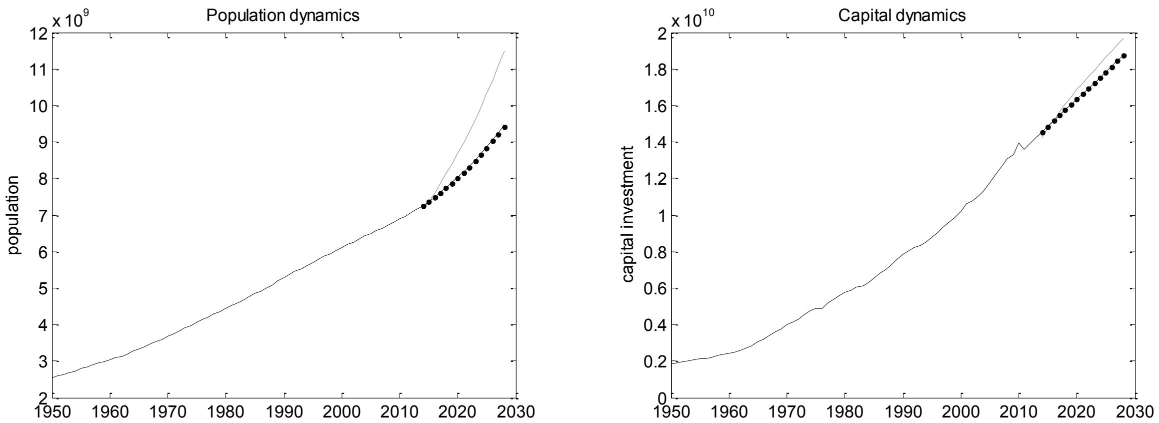
The final experiment was completed with 50% multiplicative noise. Its results are presented on figure 4. Table 4 contains the mean-root deviations.

Experiments showed that the most sensitive variable is pollution, and the most influential one is resources. In case of the model with pairwise multiplications it was also detected that multiplicative noise affects the model less than the additive one.

Table 4.

Variable	Mean-root deviation, %
Population	1,0%
Capital investment	0,5%
Capital investment agriculture fraction	0,4%
Pollution	0,4%
Resources	0,5%

In addition one can say that simple model better adapts to noise than the model with pairwise multiplications. This result confirms the well-known theoretical fact that simple model is more stable to noise but worse approximates real data [Stepashko, 2008]. By the way other our experiments show that the model with pairwise multiplications gives forecast, which almost coincides with real data.



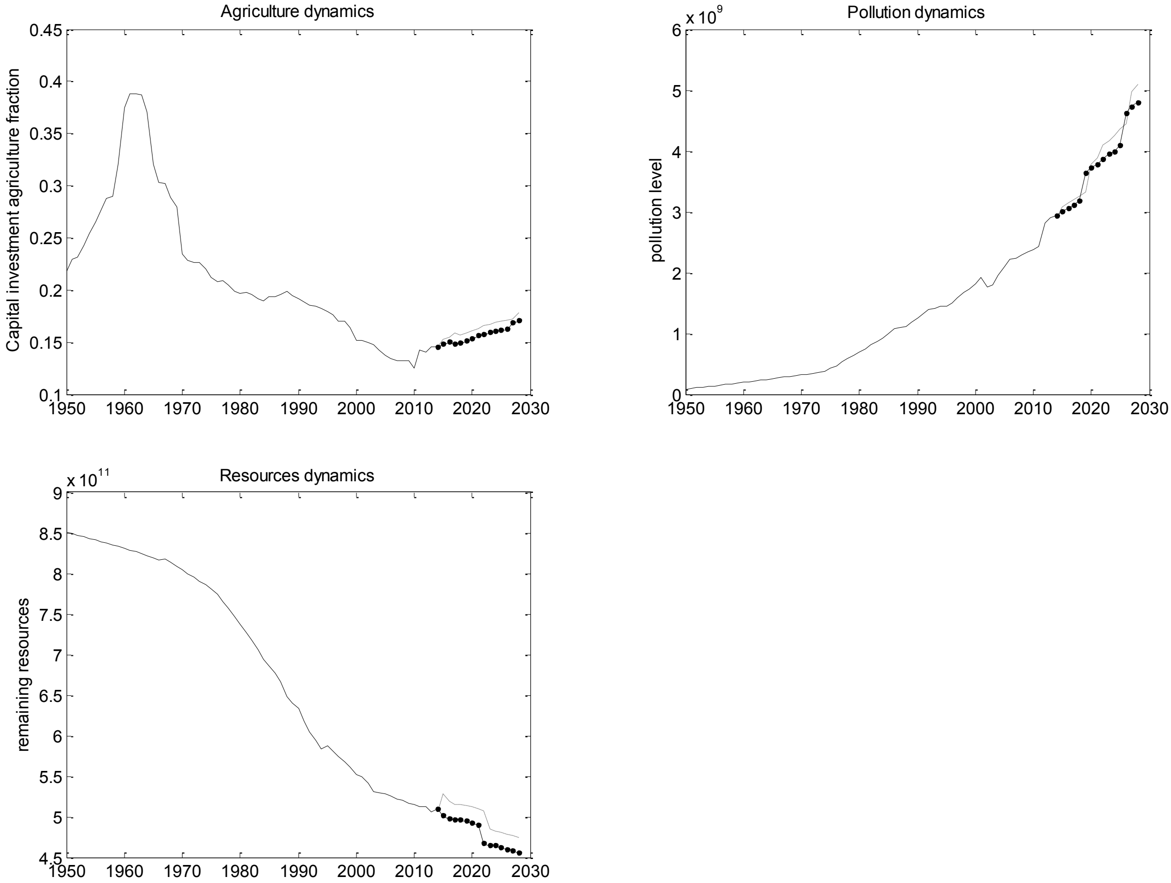


Figure 4. Influence of multiplicative noise on the model with pairwise multiplications

Conclusion

In the paper we studied noise immunity of models built in GMDH technique. The simple model with individual variables and the model with pairwise multiplications were considered. Our results are the following:

- The most influential variable in the model is resources and the most sensitive is pollution. This result coincides with that for Forrester's model [Proncheva, 2014a];
- Additive noise affects on both models more than the multiplicative one. This result coincides with that for Forrester's model [Proncheva, 2014a];
- The simple model better adapts to noise independently whether it is the addivive one or the multiplicative one.

Acknowlegements

The author is grateful to Dr. Alexandrov, Dr. Koshulko, Dr. Makhov and Dr. Stepashko for their interest to my work and advices.

Bibliography

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

[Ivakhnenko. 1968] A. Ivakhnenko. Group Method of Data Handling as a Rival of Stochastic Approximation Method, Journal "Soviet Automatic Control", No. 3 , 1968, pp. 58-72

[Malada, 1994] H. Madala H.R., A. Ivakhnenko A.G: Inductive learning algorithms for complex systems modeling. CRC Press, 1994.

[Proncheva, 2014a] O. Proncheva. WorldDyn as a tool for study of world dynamics with Forrester model / Intern. Journal "Information Theories and Applications" Vol. 21.,.No.2, 2014, pp. 126-141

[Proncheva, 2014b] O.Proncheva, M. Alexandrov, A. Koshulko, V. Stepashko. Optimal prognosis of J. Forrester's variables using gmdh techniques

[Stepashko, 2008] V. Stepashko. Method of critical variances as an analytical apparatus of the theory of inductive modeling. "Problems of Management and Informatics", N2, 2008, pp. 8-26; (rus).

[Stepashko, 2013] V. Stepashko. Ideas of academician O. Ivakhnenko in Inductive Modeling field from historical perspective. In: Proc. of 4th Intern. Conf. on Inductive Modeling (ICIM-2013), 31-37. NAS of Ukraine, Prague Tech. University, Kyev, 2013.

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

GLOBAL ADMINISTRATIVE LAW REALIZATION IN LITHUANIAN E-JUSTICE SYSTEM

Tatjana Bilevičienė, Eglė Bilevičiūtė

***Abstract**: Globalization is an inevitable phenomenon of these days. Economic, cultural, political and social globalization creates the need and of the legal globalization. Rules, standards and templates, which shall apply on a global scale, become the legitimate global functioning of the administration and the guarantor of the decisions binding. At the same time, it becomes the basis to talk about the global administrative law as an autonomous right. The purpose of global administrative law becomes to create universal - international standards that would be laid down by the general public management principles and form the transnational governance institutions, which are attributed to the implementation of administrative functions worldwide. After Lithuania's accession to the European Union and its commitment to take all of the acquits communautaire, the entire Lithuanian legal system, along with administrative law, received a change. European Union legislation in the Lithuanian law has become a priority. Throughout Europe, it is increasing demand of justice for the increasing workload of judicial systems and often difficult budgetary terms, it is imperative to continuously adapt its working methods. Development of e-Justice is one of the most important aspects of the modernization of the judicial system. The main purpose of e-Justice is to do justice in Europe more efficient and better to serve citizens. E-justice issues are not confined to certain areas of law. They arise in many areas of civil, criminal and administrative areas. This paper examines the evolution of global administrative law and its models. The paper analyses as an example of Lithuania for global administrative law application and realization of e-Justice assistance.*

***Keywords**: e-Justice, global administrative law, globalization, Prüm Decision, EURODAC, ECRIS.*

***ACM Classification Keywords**: J.1 Administrative Data Processing – Law.*

Introduction

The concept of global governance can be directly linked to the process of globalization. This process involves the different government regulation areas of every state: international investment, trade, financial and banking activities, environmental protection, health care, security, transportation, communications, and so on. t. Particular attention should be paid at the integrated global management, because the negative consequences for the most part are caused not only by the process of globalization, but both the disability to manage it.

“Globalization and the rise of global governance are transforming the structure of international law, though much of this transformation takes place beneath the surface of the international legal order and often goes unnoticed. From the perspective of classical, inter-state, consent-based international law, global governance may still appear merely as a quantitative increase in international legal instruments, sometimes coupled with stronger enforcement mechanisms and accompanied by some changes in procedures of treaty-making” [Krisch, Kingsbury, 2006].

Administrative law is perhaps one of the most important modern branches of law, cause it is regulated a very wide spectrum of the social fields. Practical aspect of administrative law science is pointed at public administration relationships through public authorities’ services for human [Urmonas, 2009]. Global administrative procedure means a wider, more flexible, neologism of modern legal standards, which fits well to the law, politics, economy, culture and other spheres of social relations [Krisch, Kingsbury, 2006].

Global administrative law doctrine argues that global governance is the administrative actions. Scientists argue that global administrative law conceptualization recognizes that there is a global or international administration.

International regulation is more directly exposed individuals within states. On the other hand, more often mechanisms of participation and the security of individuals are established in international criteria that are directly related to the decisions taken by international organizations. In this way, the legal principles of administrative law are taken from domestic administrative law and are transferred and adapted to the international context [Battini, 2005].

Global Administrative Law is a public area, which poses a challenge to classical international law and argues that the latter is unable to meet the modern needs of the global community, and it is published in the verdict and it is suggested a gradual transition to the global law. Purpose of global administrative law becomes to create a universal - international rates (rates - conglomerates), which lays down the general principles of state management and to form the transnational governance institutions, which are attributed to the implementation of administrative functions worldwide. This phenomenon reflects the legal integration processes and the establishment of global legal space, functioning in accordance with the principles of administrative law.

One of the effects of globalization - that is development of communication and computer networks. Various management and administrative problems are remotely solving. In whole Europe the demand justice is increasing, because of it the workload of judicial systems is also increasing and often in difficult budgetary terms, it is imperative to continuously adapt its working methods. E-Justice can be defined as implementation of information and communication technologies in order to improve citizens' access to justice and to increase effectiveness of legal actions, vol. y. any settlement of disputes or criminal penalties for certain offenses related activities.

Administrative law and European Union law relationship’s analysis has become relevant after Lithuania's accession to the European Union. In order to belong to one of the biggest Europe's economic and political communities - the European Union, Lithuania has undertaken to take all the legal base so-called *acquis communautaire*, in line with its national legislation. So basically the whole legal system, without distinction and of administrative law, received a change.

National e-justice systems are an important power for national economies—they represent the space where the most advanced information and communications technology (ICT) makes the administration of justice better, faster, and less expensive for taxpayers [Integrated Justice, 2010]. National e-justice systems are also an important driver for national economies.

This paper analyzes the application of information technologies in Lithuanian criminal justice system at the point of the realization of global administrative law decisions.

Global Administrative Law

Administrative law as a branch of public law regulates the social relations that arise in public administration. Administrative law discipline is closely related to the administrative regulatory legal practice (in particular in the state’s public management) [Bakaveckas et al., 2005].

Globalization is an inevitable phenomenon of these days, leading both positive and negative consequences for countries. Economic, cultural, political and social globalization creates and the need of legal globalization, because the "legal" global governance must be based on juridical - legal grounds. The era of globalization increasingly emphasize international actors (organizations) role not only in the global context, but also in the point of national law of the countries. The relationship between national law and international law is changing by growing role of administrative law.

“Global governance can be understood as regulation and administration, and that we are witnessing the emergence of a ‘global administrative space’: a space in which the strict dichotomy between domestic and international has largely broken down, in which administrative functions are performed in often complex interplays between officials and institutions on different levels, and in which regulation may be highly effective despite its predominantly non-binding forms. In practice, the increasing exercise of public power in these structures has given rise to serious concerns about legitimacy and accountability, prompting patterns of responses to those concerns in many areas of global governance. Global administrative law proposes drawing together these dispersed practices and understand them as part of a common, growing trend towards administrative - law type mechanisms for holding global regulatory governance accountable, and to inquire into the challenges this set of issues poses to both domestic administrative law and international law" [Krisch, Kingsbury, 2006].

Global Administrative Law includes legal mechanisms, principles and practices that encourage or otherwise seek to influence the global status of the administrative authorities, in particular by ensuring that these bodies meet adequate standards of transparency, performance, rationality and legitimacy of care standards, and could make effective decisions and rules [Kingsbury et al., 2005]. Actors on the international scene could successfully implement the administrative-regulatory functions, and if such functions would be carried out by public institutions - their nature are clearly administrative.

European Union law is a large part of law that is operating by principles of administrative law and for its implementation Member States as well use the national administrative measures or other methods. In other words, the European Union impact on national administrative law is called the Europeanization of administrative law and European Union regulators also seek smooth and fruitful global governance [Valančius Kavalnė, 2009]. Globalization of European Union law is limited by the area of Europe, however, we consider, that the concept of a global administrative law seeks to unify the administrative regulation to a greater extent, and perhaps even global.

Mostly global administrative law is analyzed on the basis of certain models, which are based on U.S. administrative law theory. The researchers of this country assert that global administrative law must be developed and analyzed on the basis of two reflections – "*top down*" and "*bottom up*" [Stewart, 2005].

Global administrative law model of "bottom-up" reveals the ways in which the global regulatory decisions may be taken by state national administrative legal measures. In this way, the development of the global modes of domestic administrative law is being adapted and the use of national instruments is expanded [Stewart, 2005].

The latter model is functioning by ensuring global standards' implementation of national courts, by the country's institutions (officials)' participation in negotiations or decisions on the global level, by the national public administration bodies' legislation or decisions necessary for the implementation of international standards and so on. [Tripathi, 2011].

Meanwhile, other global administrative law model of "top-down" offers to create a new administrative mechanisms to direct global regulatory regimes, thus setting global standards for the implementation of national administrative procedures. For the establishment of this model it will be needed to set up a new global institutional structure in which the institutions of legislative, administrative and independent review would be function [Stewart, 2005]. This system will provide powers of global regulatory entities (countries, organizations, etc.) to participate in the global administrative procedures, decision review would be conducted by independent international bodies, and this would include the review of national decisions related to the global administration.

Current global administrative law problems are usually associated with taxes, fines, illegal migration, and so on issues.

E-Justice

By offering standard tools, techniques, and data structures, information sharing becomes easier, quicker, and less expensive for the justice sector. This is all the more important in the current economic climate when most governments are seeking budget savings in the public sector. Public agencies require software that is intentionally designed to facilitate and accommodate new thinking and reform. When professionals in law enforcement—whether judges, prosecutors, defense attorneys, prison officers, or the police—can connect with each other and securely share information, everything changes. Economy, efficiency, and effectiveness are the principal drivers for all justice and citizen safety organizations when making decisions about ICT solutions [Integrated Justice, 2010].

E-Justice is related to the broader concept of the e-Government and there is a separate part of the phenomenon. European Communities Commission Communication *Towards a European e-Justice Strategy* [Commission of the European Communities, 2008] described e-Justice Strategy, which aims to increase citizens' confidence in the justice of Europe. Main aim of e. Justice is to improve justice in Europe to be more efficient and better serve citizens.

Algimantas Urmonas [2007] argues that the reticence of administrative law, search for solutions in only the legal environment in terms of social technology restricts its ability to enrich it to rely more on other social science information. Development of an optimal institutional framework of administrative justice and the governing legislative framework necessary should implant the advantage of the latest social technologies. The task of technology is not randomly influence the natural and social processes, but achieve the state aims by directing them to human society.

Both the purpose of legal and social technologies is to influence the social environment.

Administrative justice institutions should more widely use the public relations and information technology. Allow access to any person interested in information about their work freely available information resources (without prejudice to the rights and freedoms) must be provided for each person [Kurpuvesas, 2007].

Effective integrated justice solutions collect, analyze, and circulate information across multiple agencies and organizations in the justice community. Using the latest applications that are designed to work together, electronic data collection is more reliable and cost efficient, freeing up officials to detect and investigate crime, instead of managing paper work. These Web-based solutions—designed to comply with international standards— ensure easier horizontal and vertical integration and allow interoperability with data management software [Integrated Justice, 2010].

E-Justice issues are not confined to certain areas of law. They arise in many areas of civil, criminal and administrative law areas. Therefore, e-Justice is the horizontal case of cross-border matters in Europe. The European e-Justice is designed to create a European judicial area by the use of information and communication technologies [Council of the European Unijon, 2008].

Since 2003 Commission of the European Communities develops portal of judicial cooperation in civil and commercial matters web: http://ec.europa.eu/civiljustice/. In Europe, the several professional organizations are developing useful electronic exchange of information or interconnection projects, such as the Association of State Boards website: http://www.juradmin.eu/, the general practice of the Supreme Courts Portal: http://www.network-presidents.eu/ or the register of wills: www.cnue.be. E-Justice activity of European Union should enable citizens, particularly victims of criminal acts, to access to information and to cross the multiplicity of systems to overcome the linguistic, cultural and legal barriers [Commission of the European Communities, 2008].

The e-Justice portal: https://e-justice.europa.eu/home.do?action=home&plang=en&init=true purpose is to facilitate citizens and businesses' right to justice's implementation in Europe. Eventually, the portal should become the symbol of European justice area and the common Internet communication policy. The site provides access to the case law (civil, criminal, and administrative cases) of different countries of the European Union.

Lithuania in the European Union

After Lithuania's accession to the European Union and its commitment to take all of the *acquis communautaire*, the whole Lithuanian legal system, along with administrative law, received a change. European Union legislation in the Lithuanian law has become a priority. In Lithuanian administrative law the whole legal regulation of public administration should be implemented. Since its accession to the European Union, Lithuanian public administration institutions in the field of administrative law should apply European Union law, and administrative courts hearing cases and making decisions – should take in account the rich jurisprudence of judicial institutions of the European Union law in the interpretation and validity issues. European Union law influences almost all areas of administrative law.

Traditionally, the greatest impact of it the so-called economic administrative law is experiencing. This is evident in the competition law, state aid, and grants law's areas. However, the classical administrative law areas such as agriculture law, environmental administrative law, police law, foreigners, refugees' affairs and state responsibility are increasingly pervaded by European Union law. Particularly the strong direct influence of European Community law is on such areas as state services [Valančius, Kavalnė, 2009].

However, there are some major obstacles that still face when it is trying to create a common European administrative law. First of all, it's too small legal mechanisms for the direct implementation of European Union law. If the European Union, as the new legal regime, seeks to ensure compliance with the rule of law and the principles of direct effect, it should have its own arrangements for implementing such policy. The vast majority of European Union law today is implemented via national administrative authorities. Over the one year Lithuanian authorities transposes over 100 directives, implemented over 2000 regulations and decisions. Over the one year, there is acceptance of 50 laws, 30 government decrees and 200ministerial orders [Europos Sąjungos teisės įgyvendinimas].

In order to ensure the Lithuanian tax authorities' opportunity to cooperate with third countries (non-EU Member States) in various tax collection related fields, including assistance in calculating and collecting taxes, as well as the fight against tax fraud, Lithuania ratified the Convention on mutual administrative assistance in tax matters. Since 2012, Lithuania became the member global transparency and tax information exchange forum of Economic Cooperation and Development Organization (OECD), which the main goal - an effective international cooperation in tax field. In 2013 the evaluation of adequacy of Lithuanian legislation (the first stage) with the Economic Cooperation and Development standards for with regard to taxation was carried out, concerning the availability of relevant information, its use and return. At this stage, the highest rating was given to Lithuania [Lietuvos Respublikos Vyriausybės, 2014].

In 2013 Lithuania became a non-permanent member of the UNSC for 2014-2015 terms. UNSC Membership not only provides a unique opportunity to participate in global policy processes in the world seeking international peace and stability, but is also an important lever for the most important contribution to the principles, values and relations between countries of the rules for dissemination and consolidation. This is especially important for enhancing national security aspects, in order to protect its interests and in the global world, and in the neighborhood [Lietuvos Respublikos Vyriausybės, 2014].

Implementation of Information Technologies in Lithuania

Lithuanian information society development program for 2011-2019 [Lietuvos Respublikos Vyriausybės..., 2011a] was prepared taking into consideration the fact that the development of information society is a dynamic, rapidly changing process of a number of public and state activities and the success of public participation in promoting the positive and reducing the negative consequences of this process would contribute to the sustainable development of the information society.

The purpose of the program is to identify priorities, the goals and objectives for the development of the information society, in order to make better use of information and communication technologies providing the social and economic opportunities, and firstly online - it is very important economic and social tool that enables users to provide and receive services, work, entertain, interact and express freely their opinions [Lietuvos Respublikos Vyriausybės..., 2011a].

The program complies with the 2010' May 19 European Commission Communication to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions *A Digital Agenda for Europe* [COM (2010) 245 final] and the objectives set out in line with the European Commission in 2010' March 3 Communication *Europe 2020 A strategy for smart, sustainable and inclusive growth* [COM (2010) 2020 final]. The program aim is to promote Lithuanian population to gain knowledge and skills that they successfully use ICT to engage in a knowledge society, improve their quality of life, and reduce social exclusion, in favorable conditions [Lietuvos Respublikos Vyriausybės..., 2011a].

Administrative E-Justice Solutions in Lithuania

Public confidence in law and democratic values, independent judiciary is a necessary condition for the survival of the state. Principles such as transparency, publicity, openness help to update the administrative court system, to restore the trust of justice, in the belief that justice is implementing transparently, in open and professional courts, that for citizens the constitutional right to a fair trial is guaranteed [Piličiauskas, 2011].

Fighting against corruption in the courts and in order that all judicial decisions would be publicly available on the Internet, and the case related information would be available in cyberspace for legitimate interest holders, according to the National Courts Administration implemented project “E-services for the administration of justice in the process” since 2013 actors in civil and administrative cases are provided by access to the Lithuanian Legal System LITEKO data – to store in the system of procedural documentation. On Lithuanian courts' electronic services portal “e.teismas.lt” persons may review cases in which they are actors, listen to audio recordings of court hearings, access to the proceedings, provide to the court process documents, to prepare documents by the forms, be notified about the acceptance, found errors in the proceedings, to pay the stamp duty, court fines or costs awarded against the State [Lietuvos Respublikos Vyriausybės, 2014].

The court order the electronic booking system T[EUS (http://liteko.teismai.lt/tieus/) allows natural and legal persons to submit an application for a court order for electronic (online). This system facilitates lenders' access to justice; the court is to examine the possible cases of this type. The system is available only to legal entities and natural persons who are qualified electronic signature certificate. The certificate is necessary because without it cannot connect to the system and sent documents to sign.

Information dissemination process development through modern electronic means is certainly significant for judicial system. In 2011 by the taken in the summer change of the Civil Procedure Code has been established that since 2013 by electronic means of communication the pleadings can be given to the court. The use of electronic communication means is legal by proceedings for enforcement and bailiffs, the possibility of the trial use the electronic information and communication technologies (video conferencing, teleconferencing, and so on) is arranged. Court could present the documents by electronic means and to others, if they so wished and ordered the necessary contact data [GLIMSTEDT 2012].

Currently, an electronic service information system - e-Fine, for administration of the non-contentious manner of fines imposed on natural and legal persons is developed. Project goal is to encourage voluntary payment of fines, improving non-adversarial manner fines imposed by administrative processes and providing tools to conveniently pay. E-Fine project will allow the different administration and collection procedures to promote the efficiency of the search for fines and penalties paid or voluntary. According to the implementation of the project, the payment for the penalty will be carried out remotely [GLIMSTEDT 2012].

Currently, it is to strengthen the control of illegal migration mechanism, as well as provide more favorable conditions for entry to foreigners who are highly professional, or invest in Lithuania or come to study, teach, and conduct research or experimental development works in Lithuanian science and study institutions [Lietuvos Respublikos Vyriausybės, 2014].

In 2008 the European Union adopted the so-called *Prüm Decision* [Council Decision, 2008], for a cross-border law enforcement cooperation's improvement. This EU legislation act allows for an automated way to check the information in the other EU Member States in accordance with existing databases of DNA, fingerprint and vehicle registration data. Legislative implementation of all EU Member States requires the implementation of a number of technical solutions.

The emergence of more advanced technologies, as well as the implementation of the *Council Decision 2008/615/JHA of 23 June 2008 on the stepping up of cross-border cooperation, particularly in combating terrorism and cross-border crime*, the Government of Lithuania in 2009 adopted a resolution on the Council Decision 2008/615/JHA on the cross-border cooperation, particularly in combating terrorism and cross-border crime action plan approval [Lietuvos Respublikos Vyriausybės nutarimas, 2009]. Based on this decision, Lithuanian Police Forensic Science Centre in 2009 implemented the United States and the Austrian company's *Cogent Systems GmbH* (now *3M Cogent, Inc.*) product (CAFIS), which is used in most of the world, including the European Union. This is a high quality standardized foreign practice tested and fully functioning product that meets the requirements of Council Decision 2008/615/JHA on the exchange of data. "CAFIS provides the solution for agencies that: require fingerprint/palm print matching systems with databases of up to tens of millions of records; need rapid response times; must support a few users to thousands of users; want to include Live Scans and wireless biometric input devices; require integration of existing information systems; need to provide secure Web-based identification services" [CAFIS]. This system provides the function of *Prüm*.

In 2013 the Habitoscopical data register (HDR) started to work. Stored in the register the personal identification tags help quickly to identify persons', who illegally crossed the border, identity, and find wanted persons in Lithuania and other EU countries, to better investigate crimes [Lietuvos Respublikos Vyriausybės, 2014].

"Information about citizens of EU Member States and about third country nationals is available in many forms and systems in the Member States and at EU level. National and European instruments lay down the rules and conditions under which law enforcement authorities can have access to this information in order to carry out their lawful tasks. Fingerprint data is especially useful information for law enforcement purposes, as it constitutes an important element in establishing the exact identity of a person. The usefulness of fingerprint databases in fighting crime is a fact that has been repeatedly acknowledged" [Commission of the European Communities, 2009].

The growth and the introduction of new technology open up the new opportunities for researchers to take advantage of the pre-trial investigation fingerprint data systems not only at national but also at international level. Execution of Prüm, ECRIS and EURODAC functions is set at the level of international commitment.

Prüm is the direct exchange of fingerprint data [Lietuvos Respublikos Vyriausybės nutarimas, 2009] between the states that have implemented Council Decision 2008/615/JHA of 23 June 2008 on the stepping up of cross-border cooperation, particularly in combating terrorism and cross-border crime and Council Decision 2008/616/JHA of 23 June 2008 on the implementation of Decision 2008/615/JHA on the stepping up of cross-border cooperation, particularly in combating terrorism and cross-border crime. Prüm include automatic exchange of DNA fingerprint data and national vehicle registration data. Prüm data exchange is strictly regulated [Lietuvos Respublikos Vyriausybės, 2011b].

In 2011 adopted by the European Union Council Decision of 13 December 2011 on the launch of automated data exchange with regard to fingerprint data in Lithuania (2011/888/EU) [Council Decision, 2011] by the automated fingerprint database search in the field of Lithuania has fully complied with the general Decision 2008 / 615/TVR chapter 6, the provisions on data protection and on the implementation of this Decision shall be entitled to receive and supply personal data pursuant to the provisions of Article 9.

“In the Hague Programme for strengthening freedom, security and justice in the European Union of November 2004, the European Council set forth its conviction that for that purpose an innovative approach to the cross-border exchange of law enforcement information was needed” [Council Decision, 2008]. “The aim of this Decision is to lay down the necessary administrative and technical provisions for the implementation of Decision 2008/615/JHA, in particular as regards the automated exchange of DNA data, dactyloscopic data and vehicle registration data” [Council Decision, 2008].

EURODAC is the European Union countries’ electronic control system for controlling the European Union border crossing for illegal migrants and asylum seekers on the basis of fingerprint identification technology (System of EU member states for processing flows of asylum seekers and illegal migrants by means of dactyloscopy)

EURODAC system was established in 2000 and operates according to by European Union adopted the Dublin Convention. Each EU country must take part in the work of the EURODAC system and to establish a national EURODAC unit. In 2004 and Lithuania joined the EURODAC system. EURODAC tasks in Lithuania are implemented by the Lithuanian Police Forensic Science Center, the State Border Guard Service under the Ministry of the Interior and the Migration Department under the Ministry of the Interior.

“A system known as “Eurodac” is hereby established, the purpose of which shall be to assist in determining which Member State is to be responsible pursuant to Regulation (EU) No 604/2013 for examining an application for international protection lodged in a Member State by a third- country national or a stateless person, and otherwise to facilitate the application of Regulation (EU) No 604/2013 under the conditions set out in this Regulation. This Regulation also lays down the conditions under which Member States' designated authorities and the European Police Office (Europol) may request the comparison of fingerprint data with those stored in the Central System for law enforcement purposes [Regulation…, 2013]. ”

The main purpose of EURODAC is to prevent the same illegal migrant or asylum seeker to seek asylum in several European countries identifying them by hand fingers.

ECRIS is the European Criminal Records Information System. In 2009 Lithuania implemented by Council Framework Decision 2009/315/JHA of 26 February 2009 on the organisation and content of the exchange of information extracted from the criminal record between Member State [Council Framework Decision, 2009] and Council Decision 2009/316/JHA of 6 April 2009 on the establishment of the European Criminal Records Information System (ECRIS) in application of Article 11 of Framework Decision 2009/315/JHA [Council Decision, 2009] aim - since 2012 Lithuania started exchanging information on convictions through the development of the criminal information system ECRIS.

Conclusion

Formed transnational legal relationships are experiencing the natural changes that directly affect the process of globalization. This process was exempted and for management field, - it develops a new legal phenomenon, which is proposed to call global administrative law. This law is a new approach to international relations, ongoing management of the processes that take place in different countries administrative ideas for adapting and moving beyond national boundaries. In this way, the legal neologism created global norms governing international administrative bodies of the external and internal administrative actions and the implementation of functions, as well as support national management entity within the country to implement international norms.

Global Administrative Law provides quite a colorful range of transnational control. Global assignment management features not only are for the classical international organizations, but also private (non-profit), or "mixed" status with operators shows that on the one hand international administration becomes characteristic of such structures, which had originally been the regulatory function, on the other hand it illustrate the diversity of global administration and flexibility.

The criteria of technologies affecting social and legal practice may be considered the simplicity (the technology does not have to be too complicated), agility, changing (adaptation to the changing social legal environment), reliability (endurance reliance on technology stocks), economy (technology can be affected, but not cost-effective), the use of convenience (well-designed technology is useless if it is inconvenient for people who have to work with it). Social technology expression in law is associated with social and legal status of scientific knowledge and social effectiveness of legal activity, as dictated by both the social and legal conditions and the objectives pursued by the decision-society ways.

E-Justice issues are not confined to certain areas of law. They arise in many areas of civil, criminal and administrative law areas. Therefore, e-Justice is the horizontal case of cross-border matters in Europe. The European e-Justice is designed to create a European judicial area by the use of information and communication technologies.

Lithuania is implementing the means of European e-Justice strategy. Lithuanian administrative courts are implementing the means of public awareness about the court actions on the court websites. Lithuanian Police Forensic Science Center installed a new Automated Fingerprint Identification System is one of the most effective CAFIS identification tools. It meets the highest modern standards and guaranteed by the European Council Decision 2008/615/JHA on the stepping up of cross-border communication. Lithuania realized Prüm function - opened to researchers the access to fingerprint data systems not only at national but also at international level. EURODAC and the ECRIS system open up the new possibilities for the identification of individuals in the international arena.

Bibliography

[Bakaveckas et al., 2005] A. Bakaveckas et al. Lietuvos administracinė teisė. Bendroji dalis. Vilnius: Mykolo Romerio universiteto Leidybos centras, 2005.

[Battini, 2005] S. Battini. International Organizations and Private Subjects: A Move Toward a Global Administrative Law? International Law and Justice Working Papers. 2005(3), p. 3.

[CAFIS] CAFIS - CogentAutomated Fingerprint Identification System. http://solutions.3m.com.sg/3MContentRetrievalAPI/BlobServlet?lmd=1318828957000&locale=en_SG&assetType=MMM_Image&assetId=1273693021373&blobAttribute=ImageFile [visited 2014 03 19].

[Commission of the European Communities, 2008] Commission of the European Communities. Communication from the Commission to the Council, the European Parliament and the European Economic and Social Committee. Towards a European e-Justice Strategy. COM(2008)329 final. http://ec.europa.eu/civiljustice/docs/com_2008_329_en.pdf> [visited 2014 03 19].

[Commission of the European Communities, 2009] Commission of the European Communities. Proposal for a Council Decision on requesting comparisons with EURODAC data by Member States' law enforcement authorities and Europol for law enforcement purposes. COM(2009)344 final.

[Council Decision, 2008] Council Decision 2008/615/JHA of 23 June 2008 on the stepping up of cross-border cooperation, particularly in combating terrorism and cross-border crime. Official Journal of the European Union L 210/1, 6.8.2008.

[Council Decision, 2009] Council Decision 2009/316/JHA of 6 April 2009 on the establishment of the European Criminal Records Information System (ECRIS) in application of Article 11 of Framework Decision 2009/315/JHA. Official Journal of the European Union L 93/33, 7.4.2009.

[Council Decision, 2011] Council Decision of 13 December 2011 on the launch of automated data exchange with regard to dactyloscopic data in Lithuania (2011/888/EU). Official Journal of the European Unijon L 344/36, 28.12.2011.

[Council Framework Decision, 2009] Council Framework Decision 2009/315/JHA of 26 February 2009 on the organisation and content of the exchange of information extracted from the criminal record between Member State. Official Journal of the European Union L 93/23, 7.4.2009.

[Council of the European Unijon, 2008] Council of the European Union. (2008). European e-Justice action plan. 15315/08, Brussels, 7 November 2008. http://register.consilium.europa.eu/pdf/en/08/st15/st15315.en08.pdf>.

[Europos Sąjungos teisės įgyvendinimas] Europos Sąjungos teisės įgyvendinimas. Lietuva Europos Sąjungoje portalas. http://www.euro.lt/lt/apie-lietuvos-naryste-europos-sajungoje/lietuva-europos-sajungoje/es-reikalu-koordinavimas-lietuvoje/es-teises-igyvendinimas [visited 2014 03 19].

[GLIMSTEDT, 2012] GLIMSTEDT. Teisės žinios 2012 Nr. 5(5). http://www.glimstedt.lt/e-laikrastis/e-laikrastis-glimstedt-teises-zinios-5-e-paslaugos/2952>.

[Integrated Justice, 2010] Integrated Justice. A Microsoft White Paper. June 2010. http://www.microsoft.com/government/ww/safety-defense/solutions/Pages/integrated-justice.aspx> [visited 2014 03 19].

[Kingsbury at al., 2005] B. Kingsbury, N. Krisch, R. B. Stewart, J. B. Wiener. Foreword: Global Governance as Administration - National and Transnational Approaches to Global Administrative Law. Law& contemporary problems, Volume 68, Summer/Autumn. 2005. Numbers 3 & 4, p. 2.

[Krisch, Kingsbury, 2006] N. Krisch, B. Kingsbury. Introduction: Global Governance and Global Administrative Law in the International Legal Order. The European Journal of International Law Vol. 17(1), 2006, p. 1-13.

[Kurpuvesas, 2007] V. Kurpuvesas. Socialinės technologijos administracinėje justicijoje. Jurisprudencija. Mokslo darbai 6(96), 2007, p. 72–77.

[Lietuvos Respublikos Vyriausybės nutarimas, 2009] Lietuvos Respublikos Vyriausybės nutarimas Nr. 310 „Dėl Europos Tarybos sprendimo 2008/615/TVR dėl tarpvalstybinio bendradarbiavimo gerinimo, visų pirma kovos su terorizmu ir tarpvalstybiniu nusikalstamumu srityje, įgyvendinimo veiksmų plano patvirtinimo", 2009 m. balandžio 15 d., Žin., 2009, Nr. 49-1957.

[Lietuvos Respublikos Vyriausybės..., 2011a] Lietuvos Respublikos Vyriausybės nutarimas. Dėl elektroninės informacijos saugos (kibernetinio saugumo) plėtros 2011–2019 metais programos patvirtinimo, 2011 m. birželio 29 d. Nr. 796.

[Lietuvos Respublikos Vyriausybės, 2011b] Lietuvos Respublikos Vyriausybės 2011 m. lapkričio 9 d. nutarimas Nr.1324 „Dėl tarpvalstybinio keitimosi DNR duomenimis, daktiloskopiniais duomenimis, transporto priemonių registracijos, jų savininkų ir valdytojų duomenimis ir informacija, susijusia su didelio masto tarpvalstybinio pobūdžio renginiais ar teroristinių nusikaltimų prevencija, tvarkos aprašo patvirtinimo" (Žin., 2011, Nr.137-6494).

[Lietuvos Respublikos Vyriausybės, 2014] Lietuvos Respublikos Vyriausybės 2014 m. kovo 26 d. nutarimas Nr. 257. Lietuvos Respublikos Vyriausybės 2013 metų veiklos ataskaita.

[Piličiauskas, 2011] R. Piličiauskas. Atvirai visuomenei - atviri teismai. 2011. <http://www.lvat.lt/atvirai-visuomenei-atviri-teismai.aspx> [visited 2013 10 19].

[Regulation…, 2013] Regulation (EU) No 603/2013 of the European Parliament and of the Council of 26 June 2013 on the establishment of Eurodac' for the comparison of fingerprints for the effective application of Regulation (EU) No 604/2013 establishing the criteria and mechanisms for determining the Member State responsible for examining an application for international protection lodged in one of the Member States by a third-country national or a stateless person and on requests for the comparison with Eurodac data by Member States' law enforcement authorities and Europol for law enforcement purposes, and amending Regulation (EU) No 1077/2011 establishing a European Agency for the operational management of large-scale IT systems in the area of freedom, security and justice (recast). Official Journal of the European Union L 180/1, 29.6.2013.

[Stewart, 2005] R. B. Stewart. U.S. Administrative Law: A Model for Global Administrative Law? Law& contemporary problems. Volume 68. Summer/Autumn. Numbers 3 & 4.2005.

[Tripathi, 2011] R. Tripathi. Concept of Global Administrative Law: An Overview. India Quarterly: A Journal of International Affairs. December, 67, 2011, p. 366.

[Urmonas, 2007] A. Urmonas. Socialinių technologijų konceptualių modelių pritaikymo administracinėje teisėje paieška. Jurisprudencija. Mokslo darbai 6(96), 2007, p. 9–15.

[Urmonas, 2009] A. Urmonas. Administrative Law as a Macro-System Phenomenom. Social Sciences Studies, 2009, 3(3), p. 273–287.

[Valančius, Kavalnė, 2009] V. Valančius, S. Kavalnė. Europos Sąjungos teisės įgyvendinimas Lietuvos administracinėje teisėje. Vilnius: Registrų centras, 2009, p. 31.

Authors' Information

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

Major Fields of Scientific Research: Mathematics (statistics), IT, management and administration, knowledge management

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

Major Fields of Scientific Research: research management and law, environmental law, administrative law and procedure, law informatics and IT law, criminalistics, criminal procedure

LINGUISTIC TECHNOLOGIES

APPLIED LEXICOGRAPHY AND SCIENTIFIC TEXT CORPORA

Larisa Beliaeva

***Abstract**: Nowadays applied lexicography is a special domain of applied linguistics and language engineering in the framework of problem–oriented automated and automatic dictionaries and databases. Modern approach to dictionary creation assumes preliminary work with parallel or comparable text corpora to be considered as reference database for solving both research and practical lexicographic problems. Parallel text corpora are not always available. One of the options is to create a source lexicographical material as a text corpus with parallel presentation of initial texts, their machine translations and post-editing results. Analysis of comparable text corpus permits to reveal the set of terminological collocations (mostly noun phrases) on the translation level. The paper considers this process on the example of creating a dictionary on the Bologna process domain. The procedure permits to specify translations of lexical units in large text collections and to reveal the domain structure and its terminological system. This idea is shown on the examples of analysis for the collocations with component "higher education", being the most frequent in the Bologna process text corpora.*

***Keywords**: applied lexicography, automatic dictionary, parallel text corpora, noun phrases, terminological system, machine translation, postediting.*

***ACM Classification Keywords**: A.0 General Literature - Conference proceedings, H.2.5 Heterogeneous Databases, data translation, I.2.7 Natural Language Processing.*

Introduction

Dictionary creation is based on preliminary work with parallel or comparable text corpora, which can be considered as reference databases for solving both research and practical lexicographic problems. Parallel text corpora are the perfect source of lexicographical materials as these corpora are constructed on the basis of problem-oriented texts (articles, monographs, conference materials and their translations into other languages). Such corpora are to be sentence–by–sentence aligned, that permits to reveal and analyse terms and their translations, evaluate their level of standardization and translation conformity as well as prevalence of special variants. However, creation of such a corpus is not always possible. One of the options is to create a source lexicographical material as text corpora with parallel presentation of initial texts, their machine translations and post-editing results. These edited machine translations are to be agreed with experts in the proper knowledge domains. It is important, that the quality and potential of such corpus depends on cooperation with experts when selecting the source material and editing the machine translations. We ask authors to follow some simple guidelines.

Applied Lexicography and Automatic Dictionaries

Range of modern information technologies (IT), knowledge of which is now a significant part of any professional competence, makes it possible and really necessary to develop and implement different types of computerized tools for philological research work. These tools are used for developing new types of computer adaptive testing and tutorial systems, various systems of natural language processing (NLP), including technologies for computational lexicography. Nowadays in the framework of open and multilingual communication a series of research and practical lexicographic problems are to be solved quickly and adequately. In general the main problems solving of which involve construction of modern lexicographic systems are as follows:

- information retrieval, information and knowledge mining when using various multimedia and multilingual sources;
- retrieval, presentation and dissemination of multilingual information;
- automatic mining of new facts from multimedia resources;
- using special knowledge sources for knowledge tagging and access (knowledge sources for various types of lexicons, thesauri, encyclopedia databases, etc);
- supporting human-computer natural language and interpersonal computer-based interaction;
- supporting distance learning in the open learning systems, including adaptive knowledge testing, electronic textbooks and computer-assistant tutorial systems development;
- creating intelligent tools for automatic bibliography, texts analysis and understanding;
- modeling and predicting of user needs and intentions on the basis of possible quests to different information systems;
- supporting human-computer oral interaction and speech analysis and generation.

These problem solutions define the necessity for creation and use of specialized systems for multilingual information processing in different domains. To solve these problems we need special lexicographic bases, thus all these problems relate to computational lexicography as creation of appropriate dictionaries determines translation and communication quality.

Correspondence with crucial research problems, relevancy and adequacy of lexicographic system spectrum define the level and relevance of knowledge and data mining from the texts of different nature, composition and function. Unfortunately modern translation dictionaries, both paper and automatic, do not correspond with the science and technology levels. The case is better for the pairs of the foreign language – Russian language, but is absolutely incredible with the pairs: Russian language – foreign language. This situation is not only the result of natural lagging of the lexicographical results but the result of traditional approach for creation of new dictionaries on the basis of dictionaries published. If we compile a new dictionary on the base of old lexicons and different glossaries with small part of the terms found by a lexicographer in course of his/her translation work and do it with the help of any IT means, the situation doesn't change. In this case any information technologies used have nothing to do with modern approach. In this context the information technologies only make the lexicographer work not so difficult and tiresome when comparing and compiling different dictionary sources and when editing the final word list. Thus finding a new way using computer tools for effective creation of the dictionaries that reflect the real terminology and domain structure is a special task [TKE, 2010].

Using IT in the applied lexicography gives us an opportunity for

1. Supporting the lexicographer work at dictionary creation and maintenance:
   - solving the problems of lexical units (LU) selection, their lexicographic description, extraction of lexical unit information from the domain–oriented text files [Cerbach, Euzenat, 2001];
   - creation, editing and correction of the dictionary layout;
   - word lists creation and maintenance on the basis of LU selection from lexicographical databases according to the given criterion or a set of criteria;
   - creation and maintenance of terminological databases and ontologies.

2. Supporting the work of a specialist and/or interpreter when using different type of dictionaries in electronic or paper format:

- information extraction from various lexicographical sources (automatic, automated, resident dictionaries);
- research of lexical composition and lexical spectrum dynamics for a certain language/ sublanguage.

The task of effective terminology mining and description is solved when creating various types of automatic dictionaries to be used as the basic part of NLP systems and the quality of the NLP results depends on dictionary completeness and adequacy. Thus a sound approach to automatic dictionary (AD) creation is determined by the necessity to process a large volume of domain-oriented texts in order to access real terminology used. Special problem-domain orientation is one of the most important characteristics of a modern AD as it permits to solve lexical level ambiguity when parsing and translating separate words and collocations (terminological units).

Databases which are designed for various intellectual systems differ about their structure, composition, type of components, set of information and relation systems between the elements. But in spite of their differences all possible databases of NLP systems have common features and common problems which are to be solved when designing a database.

Dependency of the database structure on the knowledge domain and the main task of a natural language processing system and as a consequence, the necessity of the AD to be adjusted to the domain peculiarities are now mutually recognized. The same refers to the volume of a NLP-system database. It is now absolutely clear that creation of a practically usable expert system requires to design a huge database, items of which represent the main concepts and conventional terminology of the domain in question.

Not less than 95% of the source text items are to be distinguished and described with the help of a database if the NLP system is designed as a practical one. Naturally, particular volume of a NLP system database depends on the typology of the source language and the chosen procedure of morphological analysis, the aim of which is high-speed and accurate identification of the source text wordforms with the help of AD. The bottleneck of automatic machine phrases dictionaries lies in the necessity to establish for any database the following:

- the typology of machine phrases;
- the method of their recognition in course of text analysis;
- the method of storing the automatic machine phrases dictionaries.

The problem of automatic machine phrases dictionary corresponds with the fact that new and important notions in all contemporary languages are often expressed by means of phrases. Usage of a special automatic dictionary of phrases is absolutely necessary because of the different focal points of nomination: one and the same object in different languages has special descriptions and special features. In general, automatic dictionaries are created on the base of processing of huge samples of original texts (not less than 1 000 000 wordforms) translations, dictionaries and consultations with experts in the domain in question.

Thus, any linguistic database being a part of a machine translation system or a special entry to a knowledge or terminology database of any expert system shall include:

- source word dictionaries, which are organized both as dictionaries of words and dictionaries of stems,
- source phrase dictionaries and
- machine morphology for source and target languages.

In the most general sense selection of lexical units (words and collocations) for an AD shall be done on the basis of:

- statistical criterion that determines the necessity to include in the AD all the units for recognition of 95% wordforms of a text from the domain under consideration;
- criterion of syntactic independence that determines the necessity to include in the AD the units, structure of which is independent of the sentence structure and the nearest context structure;
- relevance criterion that determines the necessity to include in the AD the terminological units, which enter the terminology system, irrespective of their frequency in the learning text samples and their standardization level.

It is to be specially noted that in case of expanding information systems functions the machine translation and translational memory systems domain– and problem–oriented archives of such system are the optimum source for lexical items selection and description. The fact is that orientation on a specific data domain is very important characteristic of any AD, as it permits to solve the lexical unit ambiguity and to standardize the terminology translation on the lexical analysis level.

Modern approach to a translational dictionary creation assumes preliminary formation and use of parallel or comparable corpora of modern texts, which can be considered as a database for solving not only research tasks, but practical lexicographic tasks as well. Written text corpora, as a rule, include the texts as they are, as well as text tagging results: format boundaries and features, morphological characteristics of lexical units etc. These texts serve for creation of concordances, word and collocation lists in case of monolingual corpora, as well as for creation of multilingual lexicons and concordances if we have parallel corpora.

It is necessary to take into account, that lexicographical work even when using the whole set of IT means remains the work of art and can't be fully automated. At the same time, there is a vast potential for preparation of text files for automatic lexicographical analysis. Parallel text corpora are the perfect source of lexicographical materials [Lefever et al., 2009] as these corpora are to be constructed on the basis of problem-oriented texts (articles, monographs, conference materials and their translations into other languages). Such corpora are to be sentence–by–sentence aligned, that permits to reveal and analyze terms and their translations, to evaluate their level of standardization and translation conformity as well as prevalence of special variants. However, organization of such a parallel corpus is not always possible. One of the options to create a material for subsequent lexicographical analysis is formation of special text corpora that include parallel presentation of initial texts, their machine translations and the same translations after human postediting. These edited machine translations are to be agreed with experts in the proper knowledge domains. It is important, that the quality and potential of such corpus to a great extent depends on cooperation with experts when selecting the source material and editing the machine translations.

In case of lexicographical analysis sentence–by–sentence text alignment permits to compare initial sentence, its machine translation and the final sentence translation, thus we are able to reveal and describe the set terminological expressions (mostly noun phrases) on the translation level. In order to receive machine translation results it is expedient to use the dictionaries of a MT system, which contains the needed or comparable words and expressions. Let's consider this process on the example of creating a dictionary on the Bologna process domain (the examples presented below show in thick print those lexical units and their translations that require special attention and modification). Thus, initial sentence was as follows:

Student-centered learning *produces a focus on the* ***teaching-learning-assessment*** *relationships and the fundamental links between the design,* ***delivery****, assessment and* ***measurement of learning****.*

The sentence was translated using Word^{+} machine translation system with specialized AD for the linguistics domain:

Ориентированное на обучающегося обучение *производит фокус на отношениях* ***teaching-learning-assessment*** *и фундаментальных связях между проектом,* ***поставкой****, контролем знаний и* ***измерением обучения****.*

After human editing we had received:

При ***личностно-ориентированном обучении*** *основное внимание уделяется отношениям типа* ***преподавание-обучение-оценка*** *и фундаментальным связям между проектом,* ***подачей материала****, контролем знаний и* ***измерением качества обучения****.*

Thus, comparison of these three sentences permits to reveal the following units and their translations for dictionary registration:

student-centred learning	*личностно-ориентированное обучение*
teaching-learning-assessment relationships	*отношения типа преподавание-обучение-оценка*
measurement of learning	*измерение качества обучения.*

Besides, comparison of these three sentences permits to specify the translation of the word *delivery* as *подача материала*, this meaning corresponds with the terminology of the domain in question. All these expressions are terms and need specialized translation.

Using the sentence–by–sentence aligned texts gives us opportunity to specify translations of the lexical units in large text collections, domain structure and its terminological system. Let's consider this idea on the examples of analysis for the collocations with component *higher education*, being the most frequent in the Bologna process text corpus, a small research corpus (500 000 wordforms), the aim of this corpus was to verify the composition, structure and translations included in different Russian–English glossaries used for this very domain.

To show the potential of defining terminological system of the *higher education* field we analyze all lexical units in this corpus. There are 126 such elements in the corpus under study, for example:

higher education area, higher education assessments, higher education authority, higher education awarding bodies, national higher education frameworks of qualifications, national higher education qualifications.

The maximum length of noun phrases with the *higher education* component was 8 elements, the only noun group of this length is *New European Quality Assurance Network for Higher Education*, in which the collocation *network for higher education* is its head component. This last collocation has not been registered in this corpus as a separate lexical unit. At the following phase of analysis the whole set of noun phrases with the component *higher education* was used as the sample for receiving a frequency dictionary of lexemes. The function words were excluded from the word list, as a result we had received a key word list, which could be used as the base for terminological system structurization. The most frequent elements were lexical units *qualification, system, institution, program(mes), research, area, minister.* Convertible terms from the whole frequency list were united with the main (most frequent) key words, for example, a group with a key word *institution* consisted of the words *institute, school, university* from the word list.

On the basis of grouping the collocations in accordance with all the key words we had received the following subfields of the terminological system in question:

Programs of Higher Education, Systems of Higher Education, Institutions of Higher Education, Qualifications of Higher Education, Structure of Higher Education, Legislative Base of Higher Education, Types of Higher Education, Management of Higher Education, Degrees of Higher Education, Audit of Higher Education Quality.

These subfields correspond with the subfields for this domain which had been established on the basis of its structure analysis. Besides, the analysis of each of subfields permits to install the nomination peculiarities and variants. For example, the cluster *Management* of *Higher Education* consists of the following lexical units:

Department of Science and Higher Education, European Ministers in Charge of Higher Education, European Ministers of Higher Education, European Ministers Responsible for Higher Education, French Community Ministry *for Higher Education and Research, Minister for Higher Education, State Minister of Higher Education and Science*.

When this corpus is expanded for lexicographic research this list of departments and ministers could be more exhaustive.

Research of the collocations revealed and organised on the basis of this procedure permits both to determine the collocations from the texts and there translations to be included in the dictionaries and glossaries and to define the potential collocations (see, for example, *network for higher education, higher education law, higher education program* which are not in the list but could be constructed from the longer ones).

If we use a full-text parallel corpus as a lexicographic base it is necessary to expand them with a corpora of machine translation results. Analysis and comparison of these text files will make it possible to allocate such lexical units, which should be considered as dictionary entries. The main problem is to establish the boundaries and structures of these lexical units – noun phrases.

Noun Phrases in a Scientific Text

During the translation process the text analysis is based on formal parsing and semantic analysis. Both of these processes are based on our possibility to understand the surface structure of a sentence and semantic relations between its components [Beliaeva, 2009].

Noun phrases are the objects of special research in both theoretical and applied aspects. Such phrases are functionally equivalent to a word, but at the same time they represent a convolution of a sentence, i.e. they are, rather, units of syntax, not lexicon. So we can assume that internal structure of a noun phrase correlate with internal dependencies structure of the sentence. The problem is to find a procedure to recognize this structure in a concise form of a noun phrase. The problem is related to the fact than when translating from English to any inflectional language we should know the relation structure between the noun phrase components. If NP is analyzed in scientific or technical texts its denotative or referential status is important, but the object definiteness is given by situation of a speech act both for the author, and for the recipient. This definiteness of an out-language object is the basis of scientific text understanding by the specialists in a domain, the necessary condition of such understanding. In machine translation such understanding can't be simulated which means necessity of postediting. The same problems arise as to whether some of these phrases should be included in the automatic or some other dictionary.

The main problem of special text postediting is to recognize information on noun phrase component relations in the domain in question. This information can be received on the basis of the whole text analysis. This approach seems expedient for as it is based on the formal indications of the author's intentions which are reflected both in the text structure and in the composition of different NP with the same constituents. Establishing the relation structure is only the first step for translation of word combinations as this translation is to be adjusted to the domain in question. In doing so we are confronted with the problem of context-sensitive terms, translation of which depends not only on the domain relations and meaning but on the nearest context.

In machine translation procedure the structure of each noun phrase and its boundaries are to be determined at the sentence analysis step, thus the task of noun phrase translation is to be performed in the framework of the following operations:

1. Establishing the head element of the English noun phrase;
2. Establishing semantic and syntactic structure of the English noun phrase;
3. Finding semantic, syntactic and lexical structure of the Russian noun phrase;
4. Translation.

Since a NP is a sentence convolution, a compression of this structure, such external simplification of both the structure and the form results in the noun phrase semantic complication. The markers of relations between actual components and types of relations between elements, which sentence shows with the help of different means, are absent in the English noun phrase. Absence of morphological markers of case and gender makes it impossible to establish the "host" for an attribute or a set of attributes in the preposition to a noun or a chain of nouns.

Basic noun phrases in English are two-element combinations with a head noun, frequency of which in scientific text three times exceeds the frequency of three-element combinations. However external simplicity of frequent English noun phrase structures is misleading. The fact is that this simplicity could be the result of initial noun phrase or sentence compression. Such compression, formal simplification of NP structure leads to its semantic complication. Pursuant to these, formation of noun phrases in a real text is based on either merging noun phrases and separate lexical units in a new, more complicated nominative construction, or on condensing multicomponent NPs at the expense of deletion of the units which are implicitly obvious.

Formation of a multi-component noun phrase in a text is realized in any of two ways depending on the type of nomination: either as a process of a step-by-step complication and specification of the nomination object (gradual complication of a noun phrase with addition of its head element characteristics), or as a process of sequential noun phrase convolution. This process is realized successively on several levels:

Level 1: transfer from a complex noun phrase to a simple one due to element inversion.

Level 2: elimination of component duplication in a new noun phrase.

Level 3: coordination of semes and elimination of components with duplicated semes.

Referential status of noun phrases in a scientific text permits us assume that author's attitude on information translation and its understanding requires explication of relations within the text. Analysis of texts in different subject domains had shown, that occurrence in the text a noun phrase with length more than 2 elements is followed by occurrence a 2-compound NP in the nearest context, within the limits of 2-3 sentences or combination of title, key words and abstract. Hence, at human translation we can use this situation as a key for structure diagnostic. At MT we need to create a special text translation memory.

The peculiarities of noun phrase formation in the text are to be analyzed as follows:

Connection of two two-element NPs into a new one which results in occurrence of

- four-component noun phrase, the structure of which depends on the structure of the merging NPs, for example, if two groups of *Adjective* + *Noun* type merger the noun phrase which plays the role of an attribute is embedded in the position of the head element of the first noun phrase attribute:

indirect method + *seismic analysis*⇒ *indirect seismic analysis method*

adult learner + *second language*⇒ *adult second language learner*

- three-component noun phrase in case, when one of the elements in two initial NP coincides, for example

mental processing + *processing operation* ⇒

mental processing operation

with establishing direct relations between (in this case) the adjective *mental* and the noun *processing*,

- three-component noun phrase in case, when semantics of one of the NP elements is supported as a part of a new noun phrase by the semes of other noun phrase components, for example, merging the noun phrases

communicative method + *language learning*

results in occurrence of a new noun phrase

communicative language learning

- three-component noun phrase in case, when semantics of one of the elements is implied as a part of a new noun phrase at the expense of extralinguistic information of the domain, for example, merging the NPs *seismic stability + direct analysis*

results in formation of a noun phrase

seismic stability direct analysis,

which in a text may be convoluted to a three-component noun phrase

seismic direct analysis

The cases of noun phrase transformation considered here under the condition of text coherence and cohesion do not show all possible variants of their development in a text, however, they give the basis for consideration of possible translation of a noun phrase with high degree of structure compression. Besides the research conducted permits to show that exactly two-element noun phrases present special difficulties at their analysis and translation.

To solve the problem of such noun phrase translation we can see only two approaches which can be used both in human and machine translation.

The first approach includes modelling the knowledge base of the domain in question (in the framework of a MT system) or appealing to such factual knowledge of a translator. In case of machine translation this approach is based on vast investigations of the possible relations between both the main concepts of the domain and the items of the linguistic data base. Creation of such a thesaurus or a semantic net is not only extremely laborious but space-consuming. But the most serious disadvantage of this approach is that an unambiguous solution of the problem sometimes can't be achieved. For example, for a noun phrase *constant amplitude deformation cycle* a semantic network would show relations between the nodes *constant* and *amplitude*, *constant* and *deformation*, *constant* and *cycle* and it is impossible to use this information to establish the dependencies structure of the noun phrase both in human and machine translation.

The second approach could be more formal: we can use the information, which can be received on the basis of the whole text analysis. This approach seems more expedient as it is based on the formal indications of the author's intentions which are reflected both in the text structure and in the composition of different noun phrases with the same constituents.

Conclusion

Investigations of text structure in terms of noun phrase composition in different subject domains (medicine, seismic isolation, space systems, power plants construction, language teaching etc) had shown that dependency structure of a noun phrase with three or more constituents can be obtained from the nearest context: a 2-component NPs would show the accurate relations relevant for this special text.

This means that for an English-Russian machine translation system we need a special tool for noun phrase analysis and translation within the text boundaries, something like text translation memory, which could store the history of noun phrase development. The same problem is to be solved with human text translation or dictionary creation.

Solving the problem of dictionary creation requires to distinguish between linguistic automata with which this problem could be solved automatically or semiautomatically and linguistic automata in work of which a specialist – lexicographer could participate both at he level of corpora tagging and alignment and text analysis, and at the level of lexicographic problem solving. In using the parallel corpora the second type is more expedient.

Bibliography

[Beliaeva, 2009] L.N.Beliaeva. Scientific Text Corpora as a Lexicographic Source // SLOVKO 2009. NLP, Corpus Linguistics, Corpus Based Grammar Research, Proc. from the Intern.Conference, November 25 – 27 2009, Smolenice, Slovakia. – pp. 19-25

[Cerbach, Euzenat, 2001] Cerbach F., Euzenat J., Using Terminology Extraction to Improve Tracebility from Formal Models to Textual Representations. // NLDB 2000, LNCS 1959. Berlin Heidelberg: Springer Verlag 2001. – pp. 115-126

[Lefever et al., 2009] Lefever, E., Macken, L. and Hoste, V., Language-independent bilingual terminology extraction from a multilingual parallel corpus. // Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, Athens, 2009. - pp. 496-504.

[TKE, 2010] TKE 2010: Presenting terminology and knowledge engineering resources online: models and challenges. – Dublin: Dublin City University, Ireland, 2010. – 102 p

Authors' Information

Larisa Beliaeva – *Herzen State Pedagogical University of Russia, Professor, Chief of Machine Translation Laboratory; e-mail: lauranbel@gmail.com*

Major Fields of Scientific Research: Human and Machine Translation, Applied Lexicography, Multi-dimensional information systems

MEDICAL TEXTS CLASSIFICATION BASED ON KEYWORDS USING SEMANTIC INFORMATION

Roque López, Javier Tejada, Mikhail Alexandrov

Abstract: *This paper presents a method to classify medical texts based on keywords with the support of additional semantic information. The classification is performed in two phases. In the first phase, keyword sets are extracted for each type of disease presented in the training set. Keywords are ranked according to their semantic relatedness. In the second phase, medical texts are classified basing on the resulting keyword lists. The experimental results proved to be encouraging.*

Keywords: *Text Classification, Semantic Information, Natural Language Processing.*

ACM Classification Keywords: *I.2.7 Natural Language Processing.*

Introduction

Automatic text classification, also known as text categorization, is the task of assigning a text into a set of predefined classes or categories [Sebastiani, 2002]. During the last decades different methods of automatic document classification have been proposed. Text classification is commonly defined as a two-stage process. The first stage deals with learning a classification model on a set of pre-classified documents. At the second stage, the model is used to classify new documents. Most existing algorithms and methods are based on statistical data such as term frequency (TF), term frequency–inverse document frequency (TF-IDF), etc. The classification results based on this information can be enhanced by using some additional information.

Each document from the medical documents set includes data on clinical examinations, diagnoses, treatments, indications, medical monitoring, etc. Medical documents are short texts having lots of keywords in common (e.g.: patients, illness, treatment, etc.). In this context, relying on statistical findings alone does not help to distinguish properly between document categories.

In this work, an alternative solution is proposed, which aims to improve the classification of medical documents taking advantage of the semantic relatedness of keywords. The semantic relatedness data is obtained from the ontology of biomedical concepts UMLS (Unified Medical Language System). To evaluate the performance of the classifier, we used the OHSUMED corpus, a collection of medical documents, where each document is assigned a disease type.

The rest of the paper is organized as follows. In Section 2, the related work is outlined. Section 3 describes the proposed method of automatic medical texts categorization. The experiments and results are presented in Section 4. In Section 5, conclusions are drawn, and the future work items are identified.

Related Work

There is an extensive research on the algorithms of medical text classification. Naïve Bayes [Olszewski, 2003], Neural Networks [Farshchi, 2013], Rocchio Algorithm [Figuerola, 2001], etc., have been widely used in text classification. In most papers, statistical approaches are used [Elberrichi, 2012]. However, recently the interest has increased towards the use of semantic information for the improvement of clinical text classification. In this context, one of the earliest efforts is the work by [Wilcox, 2000]. This paper investigates the application of two knowledge resources (UMLS, a repository of biomedical vocabularies, and NLP, a medical language processor) to improve the classifier performance. The UMLS synonym set is used to enrich the representation of medical records. [Perea, 2008] presents an automatic text categorization system, which uses the UMLS ontology to expand the set of terms in the training and test collections. The obtained results show that increasing the number of terms significantly improves the performance of categorization systems. [Elberrichi, 2012] proposes a method for clinical documents classification based on the information provided by the medical thesaurus MeSH (Medical Subject Headings). Instead of the standard bag-of-words approach, the document representation based on MeSH concepts is applied, which improves the classifier performance. In [Lakiotaki, 2013], a three-stage architecture is proposed: (1) data recovery and terms extraction, (2) representation and data modeling, and (3) documents classification. The main idea is to take advantage of the UMLS semantic network data. The semantic network provides a categorization of UMLS concepts.

The Proposed Approach

The method takes into account both the statistical data and the semantic relatedness between keywords in a medical document. It includes a training phase and a classification phase. In the first phase, keywords for each type of disease in the training set are extracted and ranked according to their semantic relatedness. In the second phase, we calculate the similarity between the medical text to be classified and the keywords of each disease type. Figure 1 shows the architecture of the proposed approach.

Training Phase

The main purpose of this phase is to automatically extract the most relevant keywords for each type of disease that is present in the training set (manually classified documents). This phase has three modules: preprocessing, keyword extraction and keyword ranking.

A. Preprocessing

This module filters out irrelevant passages from the medical documents that cannot contribute to the training process. It includes three steps: tokenization, stopwords removal and part-of-speech tagging.

- **Tokenization**: In this step, the text is divided into simple tokens such as words, numbers, punctuation marks, etc.
- **Stopwords Removal**: In this step, the most frequent words are removed (i.e. pronouns, prepositions, conjunctions, etc.), which do not convey any important semantics. The punctuation is also eliminated.
- **Part-of-Speech Tagging**: In this step, nouns, adjectives and verbs are selected, which carry most of the semantics [Liu, 2009]. For the experiments in this work, the tagger proposed in [Malecha, 2010] was used.

B. Keywords Extraction

The keywords of a document are the words and phrases that can precisely and compactly represent the content of the document [Jiang, 2009]. Some words, such as patient, infection, treatment, etc., appear frequently in all medical documents and do not provide important information about the class (disease) to which they belong. For this reason, in this module we use the method proposed by [Alvarez, 2009], where the weight of a keyword indicates its importance for a class and becomes discriminant for the other classes. Within this method, a word has more weight for a given class when it appears more times in this class and less in the others.

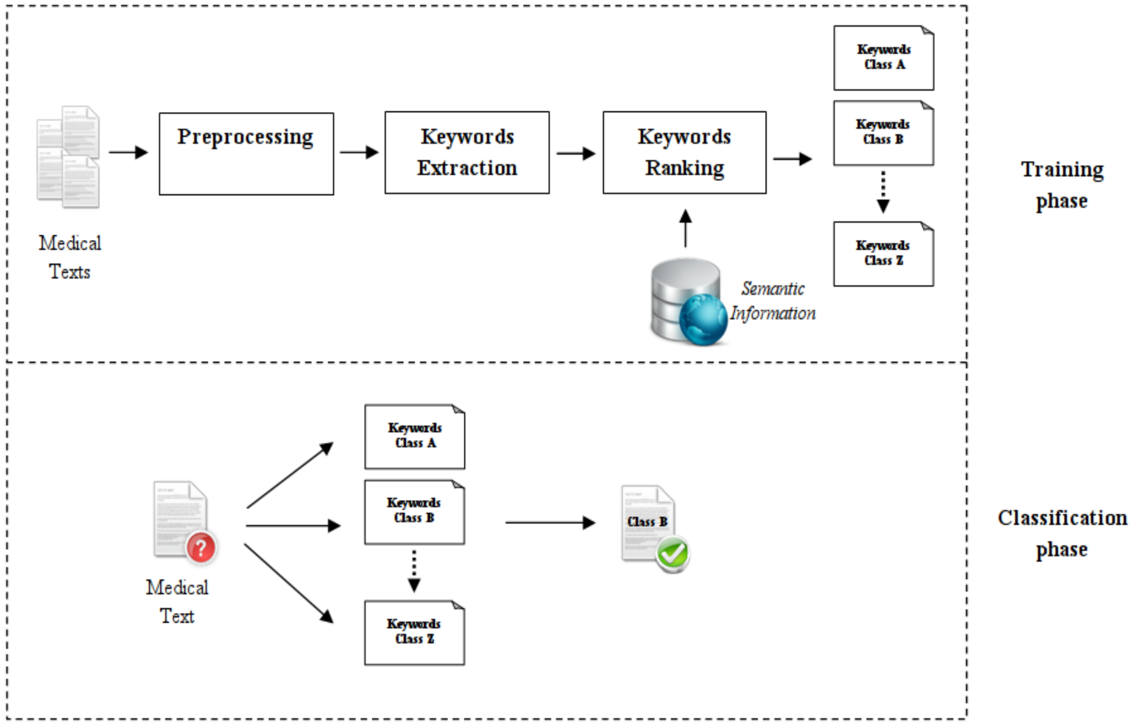


Figure 1: Architecture of the proposed approach

The weight of the word $\boldsymbol{w}_{iclass}$ for $\boldsymbol{c}$ given class is calculated as follows:

$$w_{i_{class}} = tf_i * \log\left(\frac{N_{classes}}{n_{i_{classes}}}\right) \tag{1}$$

where $\boldsymbol{tf}_i$ is the number of medical documents of the class $\boldsymbol{c}$ in which the word $\boldsymbol{w}_{iclass}$ appears. This value is normalized by the total number of documents in the class $\boldsymbol{c}$; $\boldsymbol{N}_{clases}$ is the total number of classes; and $\boldsymbol{n}_{iclases}$ is the number of classes that have medical documents containing the word $\boldsymbol{w}_{iclass}$. Based on this statistical information, for each clinical document in the training set we extracted three words with the highest weights as keywords.

C. Keywords Ranking

At the following stage, the three keywords obtained for each medical document in the training set are ranked according to their semantic relatedness. The semantic relatedness considers the relations of all types between two concepts or terms in a taxonomy (i.e. hyponymic, meronymic and any kind of functional relations including *has-part*, *is-made-of*, *is-an-attribute-of*, etc.) [Strube, 2006]. If two concepts or terms tend to occur together more often than usual, their semantic relatedness level is deemed to be higher. For example, the words *endoscopic* and *epigastric* have more semantic relatedness than *endoscopic* and *brain*. We pretend to use this information to get the most similar keywords for each type of disease. To achieve this, we use the semantic relatedness provided by UMLS. UMLS is a widely used database of biomedical terminologies, it includes over 100 terminologies and contains more than 1.7 million active concepts [Liu, 2012].

To rank the keywords, we propose a modification of the PageRank algorithm [Page, 1998]. The PageRank algorithm is used by Google to determine the website importance level. This algorithm builds a graph with websites as nodes, and the input and output links as edges. The PageRank provides a numeric value that represents the relevance of a website on the Internet. In our case, this value represents the importance of a keyword in the training set. Unlike the original PageRank algorithm, our proposed modification takes into account the weights between nodes, i.e. the semantic relatedness provided by UMLS. In this scenario, the importance of one keyword depends of the keywords that recommend it and the semantic relatedness shared between them. The modified PageRank algorithm is shown in Equation 2:

$$S(V_i) = (1 - d) + d * \sum_{j \in In(V_i)} \frac{w_{ij}}{|Out(V_j)|} S(V_j) \tag{2}$$

where ***S(V$_i$)*** is the PageRank value for the keyword ***V$_i$***; ***d*** is a damping factor that can be set between 0 and 1; ***S(V$_j$)*** are the PageRank values of each keyword that appears in the same medical document with ***V$_i$***; ***In(V$_i$)*** is the total number of the input links of keyword ***V$_i$***; ***Out(V$_j$)*** is the total number of the output links of keyword ***V$_j$***. The weight of the edge that links keywords ***V$_i$*** and ***V$_j$*** is calculated as follows:

$$w_{ij} = tf_{ij} * UMLS_{V_i,V_j} \tag{3}$$

where ***tf$_{ij}$*** is the number of occurrences of keywords ***V$_i$*** and ***V$_j$*** in the same medical document. ***UMLS$_{V_iV_j}$*** is the weight assigned by the UMLS ontology, which corresponds to the semantic relatedness between these keywords.

Classification Phase

We calculated the similarity between a given medical document and keywords of each class (disease). The class with the highest similarity index is assigned to the medical document.

The main reason of calculating the similarity is to have an idea on how many features are shared between a given medical document and the selected keywords, and also on the level of importance of those features. In [Alvarez, 2009], it is denoted as *Heavy Intersection* to this way of comparing documents with classes and is defined as:

$$similarity(d, k) = \sum_{i \in d} w_{i_{doc}} * w_{i_{class}} \tag{4}$$

where ***d*** is the document that we want to classify; ***k*** is the set of keywords of the class ***K***; ***w$_{iclass}$*** is the weight of keyword ***i*** in the class ***K***; ***w$_{idoc}$*** represents the weight of the word ***i*** in the document (frequency of the word ***i***).

Experiments

Dataset

For the experiments, we used the corpus OHSUMED [Hersh, 1994], which includes 50,216 medical documents written in English. Usually, the first 10,000 are used for training and the remaining 10,000 - for evaluation. This corpus contains medical documents describing 23 different cardiovascular diseases included in the MeSH vocabulary.

Results

The experiments are conducted to evaluate the utility of the semantic relatedness in clinical text classification. Additionally, we have made a comparison with Naïve Bayes and Rocchio algorithm.

We performed a comparative evaluation of the proposed method against a variation of the same, which does not use semantic information in keywords ranking. The two types of ranking are denominated Simple Ranking and

Semantic Ranking. The Simple Ranking uses the original PageRank algorithm, while the Semantic Ranking uses the modification proposed in this paper (Equation 2) and considers the semantic relatedness extracted from the UMLS ontology.

Table 1: Rankings comparison (5 classes)

	Simple Ranking			Semantic Ranking		
Class	**Precision**	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**
Cardiovascular (C14)	0.59	0.47	0.53	0.61	0.63	**0.62**
Digestive System (C06)	0.50	0.32	0.39	0.51	0.42	**0.46**
Immunology (C20)	0.63	0.41	0.50	0.64	0.51	**0.57**
Neoplasms (C04)	0.65	0.52	0.58	0.63	0.64	**0.64**
Pathology (C23)	0.45	0.67	**0.54**	0.51	0.55	0.53
Average	0.57	0.48	0.51	**0.58**	**0.55**	**0.56**

The performance evaluation of the proposed classifier is based on Accuracy, Precision, Recall and F-Measure. We have performed experiments using the 5 most frequent diseases in the OHSUMED corpus. Table 1 shows that the Semantic Ranking obtained the best results (0.58, 0.55 and 0.56 for Precision, Recall and F-Measure, respectively). Baseline for these 5 diseases is equal 0.34. With such a baseline these results demonstrate that semantics helps to improve the classifier performance. This improvement has been achieved due to the fact that in the proposed method the terms *gastric*, *esophageal* and *endoscopic* have stronger semantic relatedness and have more relevance to the class *Digestive System* than to the other disease types.

Classifying 23 types of diseases we achieved the accuracies 40.82%, 40.98% and 41.58% for the Rocchio algorithm, Naïve Bayes and the proposed method respectively. Here baseline were equal 16.90%, therefore all the methods showed good results. Some improvement of the results with the proposed method can be explained by more careful selection of keywords as it were described above. As explained in Section 3.1, the proposed method selects only the three most important keywords of each medical document in the training set, while the Rocchio algorithm and Naïve Bayes use all the tokens. Despite using fewer tokens, the proposed classifier ensures better results, which indicates that the tokens selected by the proposed method are representative for the disease classes.

Conclusion and Future Work

In this paper we present a method to classify medical documents, which improves the results of Naive Bayes and Rocchio algorithm. This method, in addition to considering statistical data, takes into account the semantic relatedness between keywords. The development of this classifier has been motivated by the specific features found in the medical texts. The most important points to highlight in this paper are: first, the proposed method ensures acceptable results in automatic classification of medical documents; second, the use of semantic information has proven to enhance the performance of the classifier.

The following is proposed as the future work: (1) use different weights for part-of-speech tags in the keywords ranking; (2) use different similarity measures in the classification phase; (3) run experiments on other datasets.

Bibliography

[Alvarez, 2009] Alvarez J. Clasificación automática de textos usando reducción de clases basada en prototipos. Master's thesis, Instituto Nacional de Astrofísica, Óptica y Electrónica, México, 2009.

[Elberrichi, 2012] Elberrichi Z., Amel B., Malika T. Medical Documents Classification Based on the Domain Ontology MeSH. arXiv preprint arXiv:1207.0446, 2012.

[Farshchi, 2013] Farshchi S., Yaghoobi M. Categorization of Medical Documents Using Hybrid Competitive Neural Network with String Vector, a Novel Approach. In Intelligence Computation and Evolutionary Computation, Volume 180 of Advances in Intelligent Systems and Computing, pp. 1045-1054, 2013.

[Figuerola, 2001] Figuerola C., Rodríguez G., Berrocal J. Automatic vs Manual categorisation of documents in Spanish. Volume 57, pp. 763-773, 2001.

[Hersh, 1994] Hersh W., Buckley C., Leone T., Hickam D. Ohsumed: An interactive retrieval evaluation and new large test collection for research. In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 192-201, 1994.

[Jiang, 2009] Jiang X., Hu Y., Li H. A ranking approach to keyphrase extraction. In Proceedings of the 32nd International ACM SIGIR conference on Research and development in information retrieval, pp. 756-757, 2009.

[Lakiotaki, 2013] Lakiotaki K., Hliaoutakis A., Koutsos S., Petrakis E. Towards personalized medical document classification by leveraging UMLS semantic network. In Health Information Science, Volume 7798 of Lecture Notes in Computer Science, pp. 93-104, 2013.

[Liu, 2012] Liu Y., McInnes B., Pedersen T., Melton-Meaux G., Pakhomov S. Semantic relatedness study using second order co-occurrence vectors computed from biomedical corpora, UMLS and WordNet. In Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, pp. 363-372, 2012.

[Liu, 2009] Liu Z., Li P., Zheng Y., Sun M. Clustering to find exemplar terms for keyphrase extraction. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1, pp. 257-266, 2009.

[Malecha, 2010] Malecha G., Smith I. Maximum Entropy Part-of-Speech Tagging in NLTK. Unpublished course-related report: http://www. people. fas. harvard. edu/gmalecha, 2010.

[Olszewski, 2003] Olszewski, R. Bayesian classification of triage diagnoses for the early detection of epidemics. In Proceedings of the FLAIRS Conference, pp. 412-416, 2003.

[Page, 1998] Page L., Brin S., Motwani R., Winograd T. The Pagerank Citation Ranking: Bringing Order to the Web. In Proceedings of the 7th International World Wide Web Conference, pp. 161-172, 1998.

[Perea, 2008] Perea J., Valdivia M., Ráez A., Díaz M. Categorización de textos biomédicos usando UMLS. Revista Procesamiento del Lenguaje Natural No 40, pp. 121-127, 2008.

[Sebastiani, 2002] Sebastiani, F. Machine learning in automated text categorization. ACM Computing Surveys 34, pp. 1-47, 2002.

[Strube, 2006] Strube M., Ponzetto S. Wikirelate! computing Semantic Relatedness using Wikipedia. In Proceedings of the 21st National Conference on Artificial intelligence – Volume 2, pp. 1419-1424, 2006.

[Wilcox, 2000] Wilcox A., Hripcsak G., Friedman C. Using Knowledge Sources to Improve Classification of. Medical Text Reports. In KDD-2000 Workshop on Text Mining, 2000.

Authors' Information

Roque López – *Master Student in Computer Science, Universidade de São Paulo, Avenida Trabalhador São-carlense, 400 – Centro, São Carlos, São Paulo, Brazil;*

e-mail: rlopezc27@gmail.com

Major Fields of Scientific Research: Natural Language Processing, Sentiment Analysis, Opinion Summarization

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

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 Modeling

EVALUATING EFFECTIVENESS OF LINGUISTIC TECHNOLOGIES OF KNOWLEDGE IDENTIFICATION IN TEXT COLLECTIONS

Nina Khairova, Gennady Shepelyov, Svetlana Petrasova

***Abstract**: The possibility of using integral coefficients of recall and precision to evaluate effectiveness of linguistic technologies of knowledge identification in texts is analyzed in the paper. An approach is based on the method of test collections, which is used for experimental validation of received effectiveness coefficients, and on methods of mathematical statistics. The problem of maximizing the reliability of sample results in their propagation on the general population of the tested text collection is studied. The method for determining the confidence interval for the attribute proportion, which is based on Wilson's formula, and the method for determining the required size of the relevant sample under specified relative error and confidence probability, are considered.*

***Keywords**: recall, precision, relevance, confidence interval, sample size.*

***ACM Classification Keywords:** I.2.7. Natural Language Processing, G.3. Probability and Statistics –Statistical Computing.*

Introduction

Nowadays linguistic technologies have become not only tools for modelling language but also the production factor. Computer linguistics is getting now the most strongly developing direction of information technologies. In fact, every intelligent information system with a user interface, both text and web-content processing systems, uses linguistic technologies.

The effectiveness of such technologies depends on morphological, contextual, and syntactic analysis and synthesis as well as on solving the semantic analysis problems. The number of linguistic and information approaches to solve this problem is constantly growing. Therefore common metrics should be introduced for evaluation of the effectiveness of such technologies and their comparison. But currently, there are no standard benchmarks to measure effectiveness using mentioned technologies in text collections. We propose here some indicators and test their reliability.

Usually the method of test collections is used for estimating the effectiveness of linguistic technologies in different systems of text classification, information retrieval, text mining, opinion mining, web mining etc. [Cormack, 1998]. The essence of this method consists in comparing the results of the tested technology at predetermined texts with expert evaluation for the same texts.

However comparing results of the method with experts' opinions generates the two main problems:

- expert subjectivity;
- the need for the determination of the text collection size to make experimental results reliable.

Notion the reliability means here that experimental results, which were received, will be true under certain conditions also in the framework of a certain wider class of objects.

Integral Effectiveness Coefficients of Knowledge Identification

Let's use the quantitative effectiveness coefficients of retrieval and classification approved by interstate standards for information, library science, and publishing [ISO 12620:2009]. These coefficients are precision, recall. All these coefficients are based on the subjectively determined concept of relevance. The concept of relevance is difficult to define and has a rather psychological nature. We use the definition of relevance [Mizzaro, 1997], in which relevance depends on four concepts of Relevance (IR, IN, C, T) Here IR is an information resource, which is presented by a set of collection texts for processing, IN are information needs, C is context and T is time.

Relevance is defined by experts on the scale of relevant/irrelevant/undefined and shows the correspondence or discrepancy of a text to a certain knowledge domain.

To calculate the coefficients of system recall and system precision for each domain of expert's knowledge, it is necessary to determine the following parameters:

- n_{yy} - a number of elements identified by the system as relevant, which are relevant to the local domain knowledge from an expert's viewpoint too,
- n_{yn} - a number of elements identified by the system as relevant, which are irrelevant to the local domain knowledge from an expert's viewpoint,
- n_{ny} - a number of elements that the system has not identified as relevant, which are relevant to the local domain knowledge from an expert's viewpoint.

Using these parameters coefficients of system precision and system recall are determined by the following formulas:

$$\text{precision} = \frac{n_{yy}}{n_{yy} + n_{yn}}, \tag{1}$$

$$\text{recall} = \frac{n_{yy}}{n_{yy} + n_{ny}}. \tag{2}$$

Sampling Method for Text Collections

As the determination of effectiveness indicators is based on the notion of relevance that expert defines subjectively, the reliability of recall, precision, and other effectiveness indicators of the linguistic technologies requires experimental verification on text collection.

Since used text collections are huge it makes sense to study only a part of the objects from the experimental collection, that is to execute the so-called sampling research of the population and make valid conclusions about the properties of the whole population.

As the general population of any model of natural language texts processing is tending to infinite size, the ratio of the sample size to the general population size is much less than 5 - 10%, therefore the mathematical apparatus of the sampling with replacement theory can be used how it was shown by Chetyrkin [Chetyrkin, 1982]. Furthermore, in some cases, using the necessary correction coefficient, the results for a sample with replacement can be transferred to the corresponding results for a sample without replacement.

Within our issue let's evaluate an attribute proportion in the general population on a basis of the corresponding attribute proportion in the sample. Let's consider the share of relevant texts in the collection R as the attribute proportion that shows the ratio of the number of relevant texts to the total number of collection texts. The sample estimate R_S of the proportion R is $R_S = M / N,$ where N is the size of an experimental return sample, and M is the number of identified relevant texts in a sample using the identification method. It can be shown that the evaluation satisfies all of the requirements to statistical estimates (consistency, unbiasedness, sufficiency and effectiveness) [Chetyrkin, 1982].

Since the sample estimate R_S is a point estimate of the attribute proportion, the interval estimation R_S should be used in order to find the sampling error. Since sampling errors are random variables with the same probability distribution, we can define interval estimate within which the attribute proportion of the population will be found with a certain confidence probability P.

Usually, this approach leads to three issue types:

- determination of the confidence probability for a given confidence interval and sample size;
- determination of the confidence interval for a given confidence probability and sample size;
- determination of the necessary sample size for a given confidence probability and error limit.

The determination of the confidence interval and the necessary size of the sample are the most important in our problem.

Determination of the Confidence Interval for a Given Confidence Probability and Sample Size

The determination of the confidence interval of the attribute proportion is based on the binomial distribution law [Clopper, 1934]. However, starting from samples that are more than 20 in size, the binomial distribution is symmetrized and is well approximated by normal distribution with parameters: average $<R_S> = R$, variance $D(R_S) = R(1 - R)/N$, standard deviation $\sigma(R_S) = [D(R_S)]^{1/2}$. In this case the confidence interval can be calculated using the formula:

$$P(|R - R_S| < E_\alpha) = 2\Phi(Z_\alpha) = 1 - \alpha, \tag{7}$$

where $\Phi(Z_\alpha)$ is the Laplace function. The margin error of the sample is found from the equation:

$$E_\alpha = Z_\alpha \sigma(R_S). \tag{8}$$

Let's choose the value of 0.95 usually used as the value of confidence probability, then the significance level α is 0.05. At that $Z_{0.05} = 1.96$. Then we can get expressions for right and left limits of the confidence interval R from the relation:

$$|R - R_S| < Z_\alpha[R(1 - R)/N]^{1/2}. \tag{9}$$

To do so we should solve the corresponding quadratic equation for R [Wilson, 1927]. The adequacy of using this approach to estimate the confidence intervals of the attribute proportion for small samples was proved by L.D. Brown and M. A. García-Pérez [Brown, 2001; Garcia-Perez, 2005]. Using the results of the paper [Agresti, 1998], we can get values of confidence limits in simpler way:

$$Z_\alpha = |R - R_S|/[R(1 - R)/N]^{1/2}, \tag{10}$$

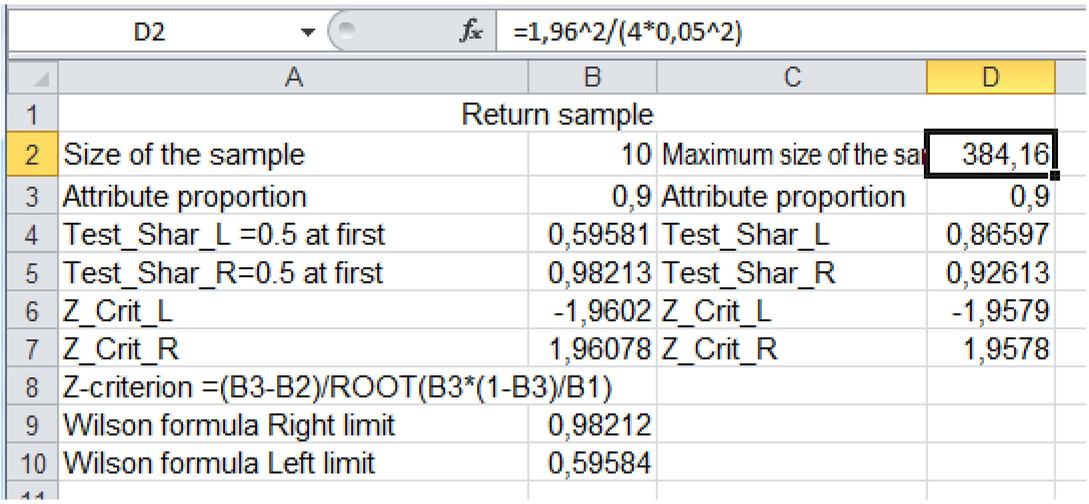
Determination of the Necessary Sample Size

To determine the size of the necessary sample for given confidence probability and the margin of error, we replace $|R - R_S|$ in (10) with E and determine N. We can see that:

$$N = [Z^2 R_S(1 - R_S)]/E^2. \tag{11}$$

The ratio (11) for size of the sample includes yet unknown sample proportion R_S. Since this proportion is unknown, it is reasonable to determine it so that the size of the sample N is maximal. Then it will be acceptable for all feasible R_S. It is easy to see that the maximum N as the function from R_S is reached at $R_S = ½$, that is $N_{MAX} = Z^2/4E^2$. Certainly, if there are a priori assumptions about the value of the attribute proportion during the research (by analogy, from the experience), this value should be used in the ratio (11). The necessary size of the sample is less than maximum one.

Quite often the value of the margin of error E = 0.05 is used for determining the proportions of attributes. Using MS-Excel, let's consider the following illustrative example given in Figure 1.



D2 =1,96^2/(4*0,05^2)

	A	B	C	D
1	Return sample			
2	Size of the sample	10	Maximum size of the sa	384,16
3	Attribute proportion	0,9	Attribute proportion	0,9
4	Test_Shar_L =0.5 at first	0,59581	Test_Shar_L	0,86597
5	Test_Shar_R=0.5 at first	0,98213	Test_Shar_R	0,92613
6	Z_Crit_L	-1,9602	Z_Crit_L	-1,9579
7	Z_Crit_R	1,96078	Z_Crit_R	1,9578
8	Z-criterion =(B3-B2)/ROOT(B3*(1-B3)/B1)			
9	Wilson formula Right limit	0,98212		
10	Wilson formula Left limit	0,59584		

Figure 1. Example of the Evaluation of the Necessary Size of Sample

Let us have a sample with replacement of size N = 10 objects and the attribute proportion of R_S = 0.9. Using *Goal Seek* procedure, we get values of right and left confidence limits for the attribute proportion in the population R for 0.95 confidence probability: 0.59 < R < 0.98. The obtained confidence interval is too wide. Let's find the maximum size of the sample N_{MAX} for the same confidence probability 0.95 (then Z = 1.96) and the margin error E = 0.05 (recall that $N_{MAX} = Z^2/4E^2$). Rounding the computed necessary size of the sample up to an integer, we have: N_{MAX} = 385. Using *Goal Seek* once again, we get a new narrower confidence interval: 0.87 < R < 0.93. It has been achieved at the cost of considerable increase of the necessary size of the sample.

Conclusion

Thus this paper substantiates the usage of recall and precision to evaluate effectiveness of knowledge identification by means of linguistic technologies in text collections. The methods of mathematical statistics are used for determining the evaluation error of chosen coefficients. We considered the problem of the evaluation of the results obtained from the sample and evaluated an attribute proportion in the general population on a basis of the corresponding attribute proportion in the sample. The attribute proportion is considered as a share of relevant texts in the collection. It shows the ratio of the number of relevant texts to the total number of collection texts. The confidence interval for the attribute proportion was computed and the necessary size of the relevant sample was determined in given confidence probability. The experimentally determined values of recall and precision coefficients for the sample size correspond to the values of the same coefficients for the complete text collection for given confidence probability equal to 0.95 and the error limit equal to 0.05.

Bibliography

[Agresti, 1998] Agresti A., B. Coull A. Approximate is better than exact for interval estimation of binomial proportions // American statistician. – 1998. – N 52. – C. 119–126.

[Brown, 2001]. L. D. Brown, T. T. Cai, A. Dasgupta. D. Interval estimation for a binomial proportion // Statistical science. – 2001. – N 2. – P. 101–133.

[Chetyrkin, 1982] Chetyrkin Ye. M., Kalichman I. L. Probability and Statistics. M.: Finance and Statistics, 1982 (in Russian).

[Clopper, 1934] Clopper C. J., E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial // Biometrika. – 1934. – N 26. – P. 404–413.

[Cormack, 1998] Cormack G.V. A Efficient construction of large test collections / G. V. Cormack , C. R. Palmer, C. L. Clarke // Proc. of the SIGIR'98 — P. 282—289.

[Garcia-Perez, 2005] Garcia-Perez M. A. On the confidence interval for the binomial parameter // Quality and quantity. – 2005. – N 39. – P. 467–481.

[ISO 12620:2009] "ISO 12620:2009 - Terminology and other language and content resources -- Specification of data categories and management of a Data Category Registry for language resources". Retrieved 9 November 2011.

[Mizzaro, 1997] Mizzaro S. Relevance: The whole history. Journal of American Society for Information Science. — 1997. — V.48. — № 9 — P. 810-832.

[Wilson, 1927] Wilson E. B. Probable inference, the law of succession, and statistical inference //Journal of American Statistical Association. – 1927. – N 22. – P. 209–212.

Author's Information

Nina Khairova *– Doctor of Computer Linguistics, Professor of Intelligent Computer Systems Department of National Technical University "Kharkiv Polytechnic Institute",*

21, Frunze str., Kharkiv, Ukraine, 61002

E-mail: nina_khajrova@yahoo.com

Major Fields of Scientific Research: artificial intelligence, knowledge identification in texts, text mining, opinion mining, web mining, natural language processing.

Gennady Shepelyov *– Head of Laboratory "Computer systems based on knowledge" of Institute for systems studies of RAS,*

Prospect 60-letiya Oktyabrya, 9 Moscow 117312 Russia

e-mail: gis@isa.ru

Major Fields of Scientific Research: mathematical modelling, probabilistic methods, interval analysis, generalized interval estimations, comparing interval alternatives.

Svetlana Petrasova *– Postgraduate of Intelligent Computer Systems Department of National Technical University "Kharkiv Polytechnic Institute",*

21, Frunze str., Kharkiv, Ukraine, 61002

E-mail: svetapetrasova@gmail.com

Major Fields of Scientific Research: artificial intelligence, knowledge engineering, intelligent systems of knowledge representation, computer linguistics, natural language processing.

THE SEMANTIC MODEL OF A LINGUISTIC KNOWLEDGE BASE

Olga Yeliseyeva, Yury Kim

Abstract: *This paper suggests basic principles for creating the semantic model of a linguistic knowledge base (SMLKB). The approach we're using here allows us to unify all stages of the automated processing of natural language structures. The given model enables us to create applied natural language systems using the same principles as with knowledge-based systems. The new SMKLB also makes possible deep-level language research and the creation of corresponding knowledge bases to store the results of such research. For the developers of applied intelligent systems SMKLB can become a starting point for a more efficient creation of natural language interfaces. Finally, language intelligent tutoring systems could be created within the new approach.*

Keywords*: knowledge base, linguistic knowledge base, semantic model.*

ACM Classification Keywords*: I.2 ARTIFICIAL INTELLIGENCE - I.2.4 Knowledge Representation Formalisms and Methods - Semantic networks, I.2.7 Natural Language Processing - Language models.*

Introduction

When formalizing a language one must distinguish the different levels and sub-levels of its structure (at the most general level, morphology, syntax and semantics ought to be represented) as well as describe the rules of transition between these levels in order to ensure the functioning of the natural language system. Due to structure and system differences existing between the levels developers have to use a special approach and a distinct formal language for each one of them. As a result, the problem of establishing transitions between levels becomes more complex which leads to less efficient solutions.

Formalized natural language knowledge is highly demanded in a whole range of applications, to name but a few: foreign language learning systems; Machine Translation; Semantic Search; Semantic Text Markup; Information and Knowledge retrieval from texts; Natural language interface for applied intelligent systems, etc.

We can assume that creating a unified base of linguistic knowledge (hereafter referred to as LKB, or the Linguistic Knowledge Base) will make the development of such systems a far simpler and faster process. Many developers even research teams are pursuing this goal, some are even fairly close to achieving it. This being said, there is still a large number of issues that haven't yet been solved or only partially solved. Following are several examples of such issues:

- Keeping the LKB up-to date, so that it reflects the current state of the language;
- Saving and adequate processing of phenomena that contradict the current norms (mistakes, dialects, and other individual features that exist in the production of every language speaker;
- Storing and efficiently processing large amounts of complex structured information about all language levels in the unified memory.

In this work we attempt to represent linguistic knowledge using a special homogenous semantic network based on a knowledge representation language called SC (Semantic Code, http://ostis.net). The creators of this language [Golenkov, 2001], [Ivashenko, 2009] believe that it will make possible to solve the above mentioned issues. The representations in the SC language will be hereafter referred to as the '**semantic model of the linguistic knowledge base**' (SMLKB). This model has some particularities including:

- It offers a unified representation of all language levels and stores corresponding information in the unified complex structured knowledge base;

- It represents linguistic knowledge using non-linear graph structures that are best suited for corresponding human cognitive models;
- It enables a formalized description of many conclusions based the discussion of the properties of objects or phenomena belonging to a given subject area (in this work we are using a knowledge-based approach and the subject area is “Natural Language”).

The above statements don’t mean that we assume we are obtaining some significant results. In this paper we only provide an outline and suggest yet another formal approach to the creation of a linguistic knowledge base. Representing the language as a homogenous semantic network we are also staging an experiment that should confirm or show the falsity of our hypothesis which says that such a method of visualization of language knowledge is more efficient and straightforward (“semantic”) than that of linear texts.

We should also mention that this topic is being investigated by many researchers across the world, and experimental models are being created, among others in projects that belong to the “semantic web” category. The most frequently mentioned achievements in this direction are improvements made in the information retrieval systems for Internet. However many issues, unfortunately, remain unsolved. For instance in may 2013 Russian web search giant Yandex announced its new project named “Ostrova” (‘Islands’) (http://beta.yandex.ru/promo), which makes use of a new search technology (http://www.seonews.ru/analytics/yandex-ostrova-tehnologiya-interaktivnogo-poiska) with a main goal to help users to achieve their goals significantly faster. It’s been more than a year now, however the project is still at the beta stage. The most popular internet search engine Google uses an additional markup system to tag its search results which saves time for the by helping him to figure out what a site is about before even proceeding to it. Since 2012 Google uses Knowledge Graph - semantic technology and knowledge base to improve the quality of its search engine with semantic-search information gathered from various sources (http://www.google.com/insidesearch/features/search/knowledge.html).

SC language developers have launched their website http://ims.ostis.net where all the information is presented as a knowledge base that can be navigated by using special search terms understood by the semantic network.

In our opinion, creating and researching various efficient and visual information representation methods in the Internet and in a whole range of specific applications is one of the most important tasks at this stage. This task is particularly important for intelligent tutoring systems for foreign languages.

It is also worth mentioning that the external representations (aimed at the user) of the information are tightly connected with its internal representation inside the computer system’s memory. This is one of the reasons why a lot of efforts are directed nowadays at researching new formats for knowledge storage and processing.

Basic Remarks

In order to keep things simple and logical we will narrow this discussion to the creation of a LKB for an intelligent tutoring system for foreign language teaching (ITS for FLT) [Yeliseyeva, 2012]. We will also use a metaphor when referring to this objective, i.e., when formalizing the knowledge about natural language what we really do is actually teach the foreign language to the computer. We will use this metaphor when talking about different aspects of the structure and contents of the educational LKB building its semantic model.

The results of our reasoning will be presented by means of a special form of homogenous semantic network which is described using a knowledge representation language called SC (SC-code), the base language of the open-source project OSTIS (Open Semantic Technology for Intelligent Systems) [http://www.ostis.net]. Its aim is to create a popular semantic technology for component-based design of intelligent systems with various purposes.

The structures of the language SC are called sc-texts or sc-structures. Sc-structures consist of sc-elements, sc-nodes and sc-arcs being the basic ones. The SC language can represent information using 2 notations: 1) graphical – the knowledge is represented using graphs in the SCg-code (Semantic Computer Graphic Code). It is one of possible flat (2D) visualizations of sc-structures; 2) textual (linear).

SC language semantics are based on the set-theoretical relation of belonging, where an oriented sc-arc from sc-element *X1* to sc-element *X2* means that *X2* belongs to the set *X1*. Sc-elements may have identifiers that we will write down using italic.

Hereafter we will discuss the formalization of Russian and/or Belarusian languages but we shall not claim that the approaches suggested in this work are complete or universal. However, we do hope that many of the ideas and formalisms offered hereafter, with necessary precisions and additions will be useful for creating LKB of many other languages.

Our reasoning is based on the experience we gained when creating online learner dictionaries for Russian and Belarusian hosted on http://rus.lang-study.com and http://by.lang-study.com respectively. These sites are the testing ground where Belarusian State University students enrolled in a bachelor, master and doctoral program upload their content. Since within these projects mainly learner's dictionaries of foreign languages are created, the contents of the dictionaries and thematic word groups are determined by the communicative objectives. To a certain extent these resources are being developed and extended spontaneously which, we have to admit reduces the quality of the results achieved. On the other hand, such projects tend to be constantly modified and improved as the competences of their creators improve. After all, their main objective is to train qualified professionals capable of creating linguistic resources rather than to produce a final product.

We would like to add that the present work is a new step in the development of these projects and should provide a foundation for creating ITS based on the aforementioned sites.

Definitions

First we shall clarify the meaning of the term 'linguistic knowledge base'. For that, a definition of "knowledge" by Gavrilova will be useful: "**Knowledge** is well-structured data" [Gavrilova, 2001].

Thus, we will define linguistic knowledge base as well structured data about the language.

Another useful concept is that of language structure [Kobozeva, 2009], which includes the dictionary (vocabulary) and the grammar. Thus, structured language data should include information about the vocabulary (lexis) and grammar of the language. Thus, we can subdivide the linguistic knowledge base in two components:

1) **lexical knowledge** (structured data); 2) **knowledge** (structured data) **about grammar**.

Hereafter the above elements of the LKB are considered in more detail.

As we've already mentioned above, the results of our reasoning will be presented in the form of sc-texts. In order to create such descriptions, the OSTIS technology requires the identification and comprehension of the notions and relations of the knowledge domain being formalized. This enables the creation of a sc-sublanguage that corresponds to the given knowledge domain and consists of sets of sc-elements (mainly, sc-nodes). The identifiers of such sc-elements are signs of concepts, relations and attributes identified in the knowledge domain. Therefore we will refer to the collection of resulting sc-texts as well as the description of the ontology of identified concepts and relations as **the semantic model of the linguistic knowledge base** (SMLKB).

Lexical Knowledge

It is well known that in order to master the vocabulary of a foreign language one needs not just the words with their translations in his native language, but also additional information about these words. Such additional information includes: 1) lexical meaning (meaning explanation) of the word; 2) grammatical properties and inflection rules; 3) information about word compatibility; 4) semantic and other types of relations; 5) particularities and examples of use: word combinations, sentences, texts.

Note that grammatical knowledge is also included in this list (cf. items 2 and 3). Thus, we don't separate the lexis from grammar. Moreover, in the present work we intentionally avoid using the common approach to the identification of language levels (morphology, syntax, semantics, etc.). As we've already mentioned, we are generalizing the experience of the creation of Russian and Belarusian dictionaries for http://rus.lang-study.com and http://by.lang-study.com. Besides, we also use the approach to creation of explanatory combinatorial dictionary described in the works of Melchuk [Melchuk, 1974]. In both cases we have noticed that solving lexicographic issues leads to results that can be applied to many other areas besides the creation of dictionaries.

Let's elaborate on the remarks we've made concerning the vocabulary of natural language (NL) as well as the possible ways of its formalization using the SC language or the experimental hypertext model of the aforementioned online learner dictionaries.

Lexical Meaning (meaning explanation) of a Word

Works on lexical semantics [Apresyan, 1995], [Kobozeva, 2009] suggest various ways of describing the meaning of a word. One of the approaches involves forming semantic fields. Componential analysis is carried out to define the hierarchy of topic groups as well as to match each word of the language with one or several of these groups. Many dictionaries, including semantic [Shvedova, 1998, 2002], ideographic and thematic dictionaries and thesauri are created in this way. An example of Russian language classification has been suggested in the semantic dictionary under the general editorship of Shvedova which can be accessed online at http://www.slovari.ru/default.aspx?s=0&p=2672. Learner's dictionaries at http://rus.lang-study.com and http://by.lang-study.com can be systematized by determining thematic word groups.

It's worth mentioning that at present there is no universal approach to the problem of defining the contents and the limits of semantic fields. In our opinion, that doesn't even seem to be possible, taking into account the objective nature of the natural language and the subjectivity of human perception. One thing is certain though: structures that are being used to systematize vocabularies of various languages are complex and non-linear. This is why we chose the SC language of semantic networks designed for describing and processing this type of non-linear structures.

Now let's consider a simple example. Fig.1 shows a fragment of a sc.g-text describing a series of words related to "human" – which is one of the topics («Человек») presented on the website http://rus.lang-study.com. Fig.1 A) displays a simplified structure where many nuances of Fig.1 B) are simply omitted. Let's take a closer look at both fragments of the semantic network. Arcs leaving the *части тела* (=parts of the body) sc-node fully reflect the basic semantics of the part-of relation described above. Indeed, the head (sc-node with the *голова* identifier), hand (*рука*), leg (*нога*) – all these are parts of the body, i.e., they belong to a set marked with the *части тела* sc-node. Semantics of all the sc-arcs leaving the sc-node *внешность* (=appearance) is similar. Naturally, the arcs leaving the *человек* (=human) sc-node should have different semantics and define other type of relation. However, it is not reflected in the Fig. 1 A). The arcs present in this figure can be understood so that the *parts of the body*, *appearance* and other sc-nodes belong to the *human* set. This description is obviously not really correct. That's why in the figure 1 B) we use a different style for the sc-arcs leaving the *человек* sc-node. These are so-called oriented couples which depict an oriented binary relation between the given sets. The semantics of the given relation are described by the arcs leaving the sc-node with the *включение множеств** (=inclusion of sets) identifier. This way we show that the *части тела*, *внешность* and some other sets are subsets (and not elements) of the *человек* set.

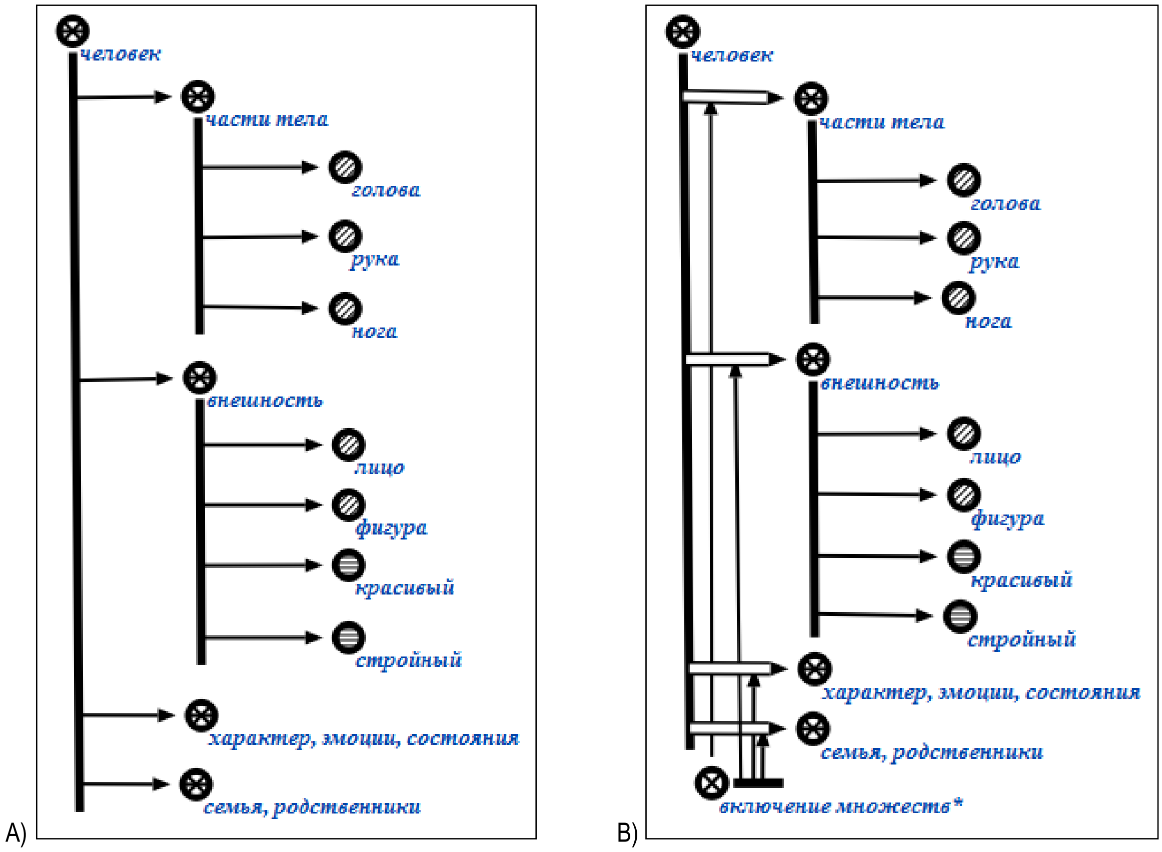


Figure 1. An example of thematic groups description in the dictionary

Fig.1 also uses various images of sc-nodes. These are extended possibilities of the SC language for describing the semantics of corresponding elements. For instance, the sc-node ⊗ defines sets (groups) of elements of the same nature. In Fig.1 we have words that have in their meaning one common semantic marker. Using sc-nodes of type ⊘ and ⊖ subject (constant) elements of a subject domain (concrete, tangible and abstract, intangible respectively). No sc-arcs usually leave such sc-nodes. Finally, the sc-node ⊗ in Fig 1 B) defines a relationship. Besides the relations we can use special sc-nodes to show attributes (⊕) and a bunch of relations (⊙). SC language graphical notation offers one more possibility to place elements in a more a more straightforward way. It's the so-called sc-tire – a thick line leaving the sc-node that becomes longer. This allows for additional arcs to be drawn to and from this sc-node without overloading the picture and making already complex networks even more confusing.

Let's consider the LKB fragment further. In order to make a more complete description of semantics of all sc-nodes shown in the Fig.1 we need to introduce two more set signs – *тематическая группа* (=thematic group) and *лексема* (=lexeme). Let's describe the corresponding statements about element typology using a linear sc-text notation:

тематическая группа -> *части тела*; *внешность*; *характер*, *эмоции*, *состояния*; *семья*; *родственники*;;

лексема -> *часть тела*; *голова*; *рука*; *нога*; *лицо*; *фигура*; *красивый*; *стройный*;;

Obviously some of the lexemes can be included in several thematic groups simultaneously. For instance, the word *лицо* (=face) can belong to *части тела* and to the *внешность* (Fig. 2).

Besides the attribute *характеристика_* (=property) in Fig. 2 indicates an additional semantic marker of the words *красивый* (=handsome) and *стройный* (=slender).

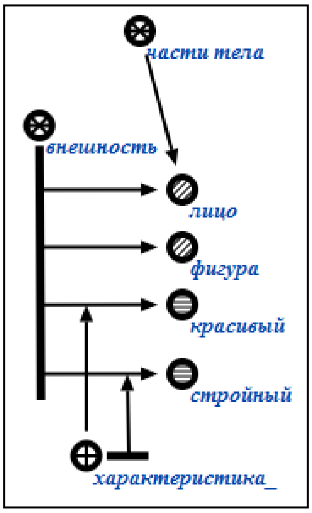


Figure 2. Additional descriptions the dictionary fragment

We will mention once more that we are not claiming that our reasoning is linguistically complete and precise. Our task is to consider in detail formalisms used for describing the semantic model of linguistic knowledgebase. We are well aware that many of our statements may be criticized by linguists and we fully accept this fact.

Of course the above mentioned method of describing word semantics of a natural language by assigning words to various semantic fields (thematic groups) is not a silver bullet. Componential analysis of lexical meaning consists in defining several semantic markers that together form this meaning. Linguists call these markers semes and distinguish at least three types of semes: differential seme, archiseme, and contextual seme. Fig. 3 shows an sc-text describing the lexeme *father* using the relation of *семантическая декомпозиция** (=semantic decomposition). For this description we've used the article «Сема» (Seme) from the linguistic encyclopedia accessible online at (http://tapemark.narod.ru/les/437c.html).

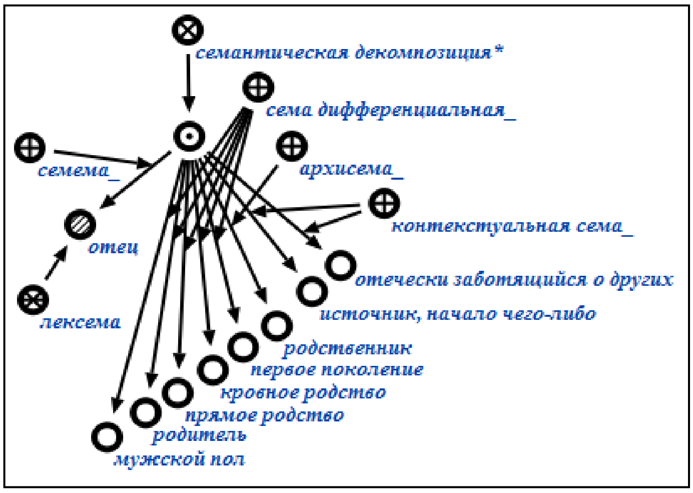


Figure 3. Description of lexeme father through enumeration of various semes

Let's mention here yet another particularity of the SC semantic network language we're using. The sc-nodes used in descriptions having same identifiers undergo the operation of joining when they are loaded into the sc-memory (also called graphodynamic by developers). Thus if we are an sc-node with the same identifier in different fragments of an sc-text, we are basically expanding the description of the same element of the semantic network. When creating a knowledgebase we mostly identify sc-nodes, so every time we mention the same node in our descriptions we basically add sc-arcs leading to or leaving this node.

In accordance with the above mentioned possibilities of the SC language, when similar semantic descriptions of the lexeme *mother* are added to the knowledgebase, sc-nodes like *parent*, *direct relationship*, *real kinship* and others will be completed with new sc-arcs leading to and from them. Further if we make a more complete description of a number of natural language words we will obtain quite a complex structure clearly showing semantic connections between these words. Using the connections between the sc-nodes defining semes finding semantically close words are relatively easy to find by certain markers (semes). Adding detail to the semantics of words that have already been described in such a way is as easy as adding missing connections (arcs) to the sc-node that represents the target lexeme.

Using the above mentioned method one can describe almost any kind of information involving semantics and functions of words in the language. To do that we need to introduce and describe respective relations and their attributes. We will add several examples in what follows.

Grammatical Properties and Inflection Rules

In Fig. 4 we see a fragment of an sc-text describing grammatical properties of several words that have been described above in the context of a hyerarchical structure representing semantic field in the Fig.1. Unlike in Fig. 1, we've used a slightly different depiction of sc-nodes, more specifically, rectangular boxes with words inside. These are the so-called text (string) contents of sc-nodes. To avoid unnecessary complexity we don't display the identifiers of these nodes since in this case they are identical with the contents. Thus, in a general case identifiers are not necessary and may only be needed to ensure more precise merging of sc-nodes when they are being loaded into the graphodynamic memory. We will skip the textual explanation of the sc-structure in Fig. 4, assuming that is quite straightforward.

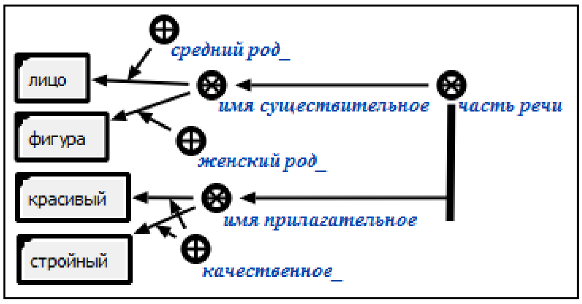


Figure 4. Description of grammatical properties

Up until now we've been describing different LKB at the declarative level. So, for instance, in the Fig. 5 we've described the inflectional paradigm of the word *аудитория* (=lecture hall) using an sc-text at a declarative level (cf. the dictionary article at http://rus.lang-study.com/slovar/100-slov-dlya-1-kursa/auditoriya-3).

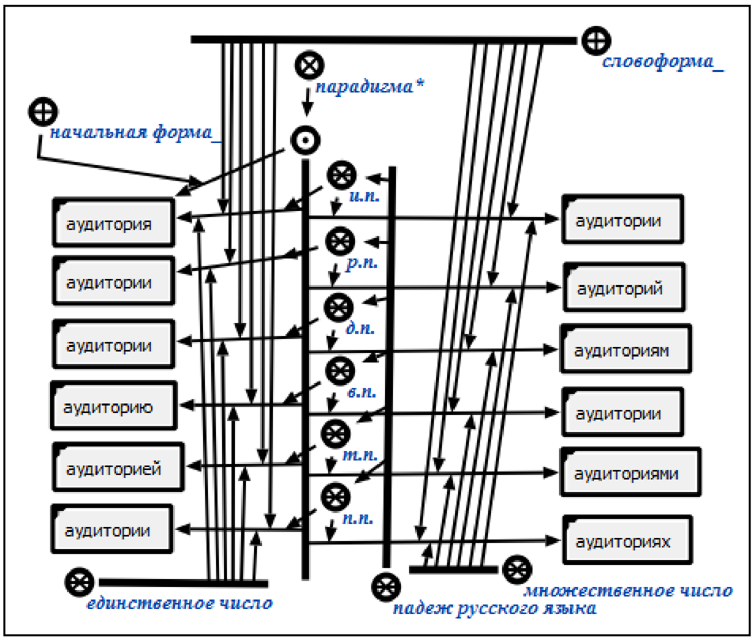


Figure 5. Description of the word inflection paradigm

At the procedural level of knowledge the knowledgebase should describe the corresponding rules of inflection, rules for automatic identification of grammatical properties of the input word forms, etc. In this work we are not attempting to provide such descriptions because it would require a lengthy explanation of the functionality of the SC language, of the processing operations used by the graphodynamic memory [Golenkov, 2001], as well as of the procedures of automatic analysis and synthesis. Taking into account the current state of the research we don't consider such exhaustiveness necessary, since the number of available works on automatic text processing nowadays is impressive. The tools of the OSTIS project used in this research allow for combination of different types of knowledge including external procedures and algorithms in one system.

Here we will only mention some examples of use of the suggested LKB descriptions for problem solving. The Fig. 6 shows an example of an sc-text describing the search for a word form using a given lexeme (*аудитория*) and a set of grammatical properties (*singular*, *Genitive Case*). This description makes use of sc-variables defined with squares instead of circles for sc-nodes or with dashed lines for sc-arcs. Pattern-matching search is a basic operation of information retrieval used by the sc-machine, initialized using respective sc-descriptions and semantic network search procedures – all stored in the sc-memory.

Fig.7 shows an sc-text describing the search of values for the grammatical properties of *case* and *number* of the given word form *аудитории* and its initial form.

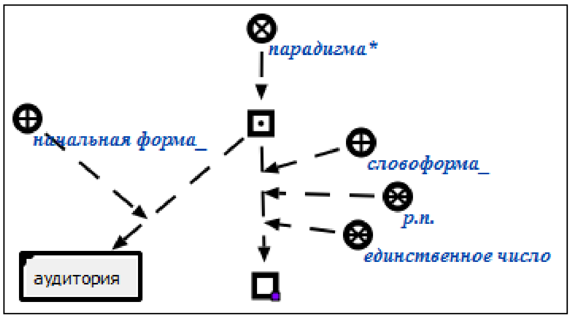


Figure 6. Example of word form search based on the inflectional paradigm

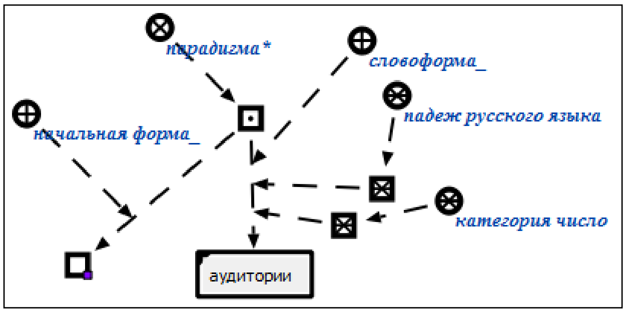


Figure 7 Example of search for grammatical properties of a word form

Information About Word Compatibility

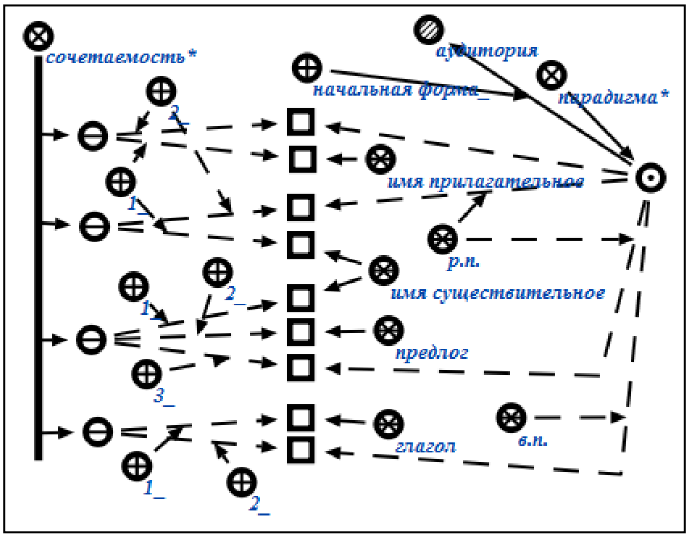


Figure 8. Description of information about the compatibility of a lexeme with other words

To help students develop the ability to use words in combinations and complete sentences and, as a result, to ensure effective communication many foreign language textbooks offer information about the compatibility of new words with other lexical units. For instance, when teaching Russian as a foreign language one can use specific formulas, some of which can be found in the dictionary article «Аудитория» at rus.lang-study.com (http://rus.lang-study.com/slovar/100-slov-dlya-1-kursa/auditoriya-3). The Fig. 8 shows a fragment of description of some of these formulas. For this we have introduced the *сочетаемость** (=compatibility) relation, oriented bundles of which form templates of a certain type (the sc-construction contains variables) which, when supplied with actual word forms that have specified grammatical properties, will provide appropriate word combinations. In this example, containing many sc-variables we have intended to demonstrate the most general form of description and the morpho-syntactic compatibility. If some variables are replaced with actual sc-nodes (lexemes) and/or word forms we will obtain an example of the description of lexical compatibility in the LKB [Popov, 2004]. Finally if similar structures also indicate semes and the thematic groups (semantic fields) or other semantic markers of lexemes, then we will obtain a rough description of semantic compatibility.

The use of such descriptions in the functioning of an application system seems sufficiently obvious. Appropriate fragments of the semantic network along with sc-variables will serve as a model for searching and/or generating necessary information in the sc-memory.

The above suggested description is not the only and perhaps not the most optimal method. Many aspects of lexical compatibility, in particular, can be described using the lexical functions, suggested by Melchuk in the framework of the Meaning ⇔ Text model [Melchuk, 1974]. Some examples of such descriptions will be provided further.

Semantic and Other Types of Relations

Let's consider the possibilities that our SMLKB offers for describing various relations between lexemes.

Fig.9 and Fig.10 present examples of description of the synonymic relation. To indicate such a relation in the SC language we use the abbreviation *Syn**, suggested by Melchuk in his Meaning ⇔ Text model [Melchuk, 1974]. In Fig. 10 the type of synonymic relation between lexemes is specified using the attribute *экспрессивно-стилистическая_* (=expressive-stylistic). Besides, in Fig.10, a symmetrical bundle is defined by a double line to increase the visual clarity. This is supported by the graphical notation of the SC language.

The synonymic relation is used more widely in the SC language. In particular, it is used to describe interlingual synonymy (Fig. 11). In some cases, however (for instance, when there is no exact equivalent), it is more appropriate to use the sc-relation of *трансляция** (=translation) (Fig.12).

Figures 13 and 14 show graphic depiction the relation of hyponymy (based on the example by [Kobozeva, 2009]).

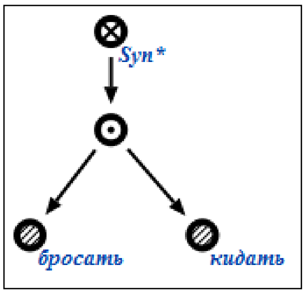


Figure 9. Describing a synonymic relation

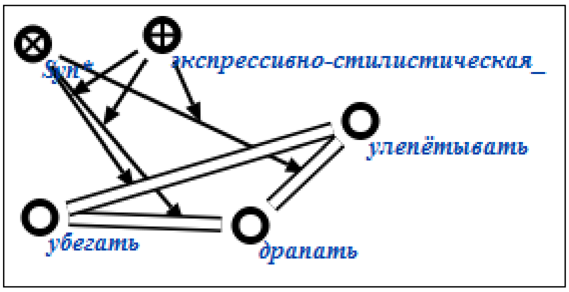


Figure 10. The description of an expressive type of synonymic relation

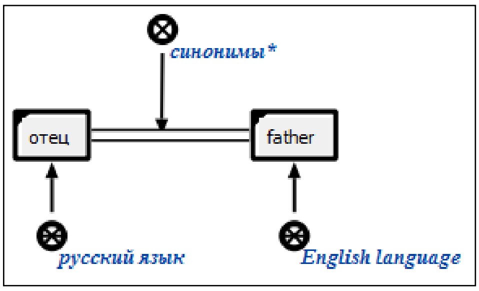


Figure 11.The description of the interlingual synonymy

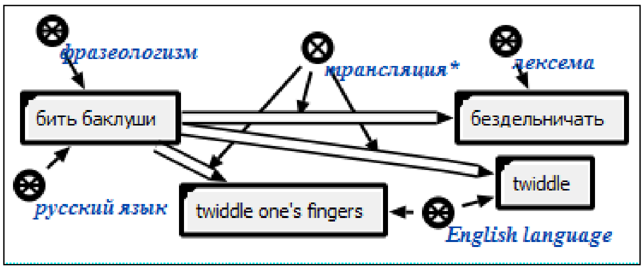


Figure 12. Using the translation relation

In the same manner we could describe various kinds of relations and lexical functions between separate words, word combinations and phrases. No less important is, for instance, the description of associative (http://wordassociations.ru/) and some other relations.

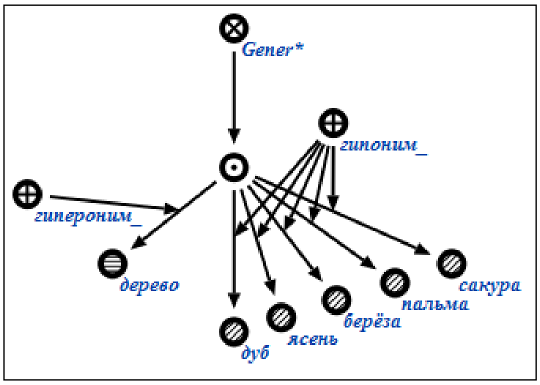


Figure 13. Describing the relation of hyponymy

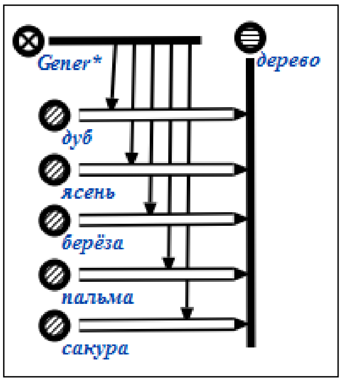


Figure 14. An alternative visualization of the hyponymy

The Figure 15 shows an example of relation “whole – part” tagged using the identifier *Sing** , borrowed from the [Melchuk, 1974]. The contents of this example have been borrowed from [Kobozeva, 2009].

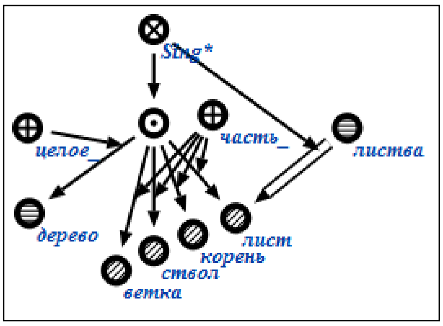


Figure 15. The description of the “whole – part” relation

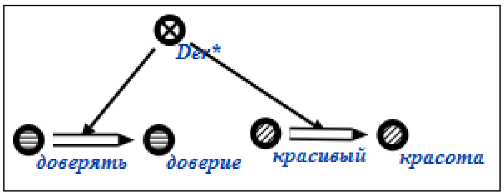


Figure 16. Examples of derivational relations

Finally, in the Fig.16 we provide an example of a derivational relation (*Der**). We can use additional attribution to specify particular derivation methods or appropriate rules to enable the function of automatic generation, but won’t do that due to space limitations.

We have started systemizing some of the relations described above in the dictionary article sat http://by.lang-study.com by determining appropriate fragments with subheadings (cf. http://by.lang-study.com/slounik1/ezha-harchavanne/sadavina1/yablyk).

Particularities and Examples of Word Use in Combinations, Sentences and Texts

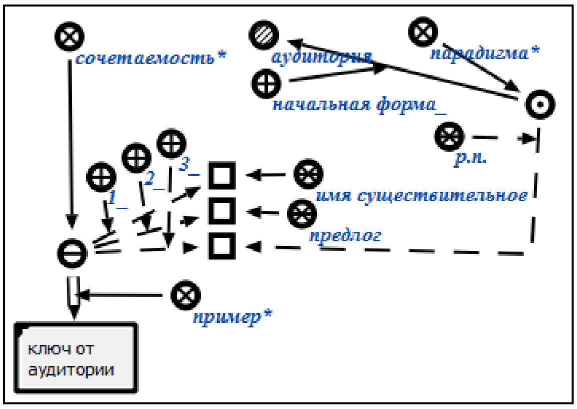


Figure 17. Describing an example of use along with as the compatibilityFig 3. Description of lexeme father through enumeration of various semes

Some particularities of word use in speech have been mentioned by us before during the discussion on the compatibility and other relations and lexical functions. Here we will provide example of fragment of sc-text that contain explicit references to instances of word use (Fig. 17).

Such descriptions are especially useful for an instructional LKB of an ITS. For instance, on our websites such examples represent the essential information on a lexeme. Incorporating such examples along with their most adequate translations in the LKB of machine translation systems can be used to improve the quality of translation.

Knowledge (structured data) About Grammar

As we've mentioned before, while formalizing the vocabulary of a language we also described some aspects of its grammar. Such aspects include grammatical properties (Fig.4), inflectional paradigm (Fig.5), information on compatibility (Fig.8) etc. In this work we discuss the formalization of knowledge of Russian language which has a complex morphology. Professionals that deal with the lexical semantics note that Russian doesn't have a clear distinction between the lexis and the grammar [Kobozeva, 2009]. Therefore one may get an impression that the LKB described in this work pays too much attention to the morphological level. At the syntactical level, the formal description of natural language texts through sc-texts is done in the same way. So, for example, in [Yeliseyeva, 2014] shows a somewhat simplified description of the sentence structure at the surface syntax level. Moreover is shown there the correspondence between the text in a natural language and its translation in the SC language.

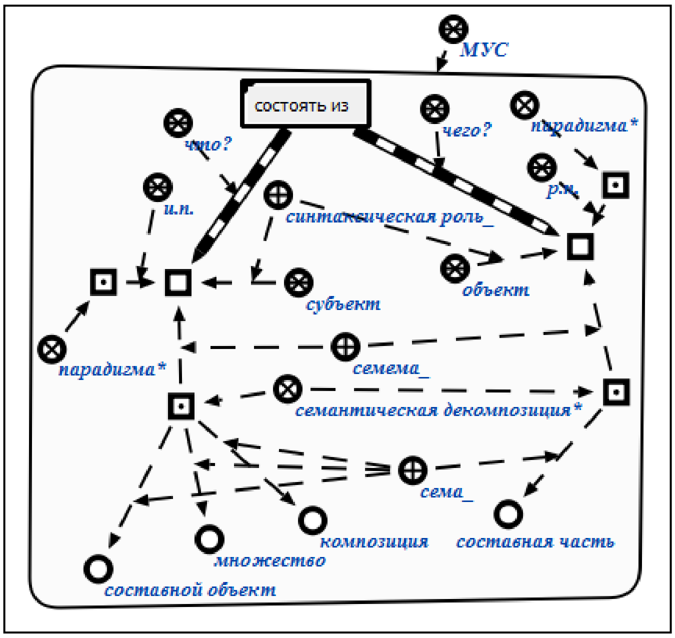


Figure 18. A description of a model of coordination of a word

To execute automatic synthesis and analysis of sentences in a natural language, sc-structure templates containing in their descriptions sc-variables should be stored in the LKB or generated in the sc-memory. Models of word coordination (MWC, [Apresyan, 1995]) seem to be good source data for creating such templates. In Fig. 18 we provide an example of a formal description of the MWC of the lexeme *состоять из* (=consist of). At the same time this MWC provides morpho-syntactic information as well as the semantics of possible lexemes that are coordinated by the lexeme *состоять из*. For the sake of simplicity we use coordination questions *что? и чего?* instead of specifying semantic valence of the lexeme *состоять из*.

In the functioning of a natural language application system formal representations of MWC can be used for analyzing input sentences as well as for generating texts in a natural language.

Conclusion

In the present work we have attempted to describe the semantic model of a linguistic knowledge base in the format of homogenous semantic networks organized in a particular way using the SC language of knowledge representation. One of the particularities of the suggested model is that it provides a uniform presentation of all language levels and stores the corresponding information in the unified knowledgebase. This has become possible because we intentionally didn't define separate levels in the process. In our opinion, this approach allows us to unify the mechanisms of automatic processing of natural language structures. Based on the present model it becomes possible to create applied natural language systems organized in accordance with the principles of knowledge-based systems. Besides, in future natural language knowledge can become the foundation for knowledge bases in other domains of knowledge [Yeliseyeva, 2011]. We also consider that the LKB will become an important tool for deep research of the natural language.

Due to the size limitations of the article we cannot provide here all the relations but we hope that we've managed to present the most general approaches to such descriptions.

For applied intelligent systems, LKB is the basis for an efficient realization of natural language interfaces. The suggested approach also enables the realization of intelligent tutoring systems for Russian language teaching. The present work uses the concept of semantic analysis of the Russian language structure with an outlook for the best practices in its teaching.

Bibliography

[Shannon, 1949] C.E.Shannon. The Mathematical theory of communication. In: The Mathematical Theory of Communication. Ed. C.E.Shannon and W.Weaver. University of Illinois Press, Urbana, 1949.

[Apresyan, 1995] J.D. Apresyan. Selected Papers, Vol. 1. Lexical Semantics: 2nd ed., corr. and add. – M.: School "Languages of Russian Culture" publishing company "Eastern Literature" RAS, 1995. (rus)

[Gavrilova, 2001] Gavrilova T.A., Horoshevsky V.F. The knowledge bases of intelligent systems. St.-Petersbourg, 2001. (rus)

[Golenkov, 2001] Golenkov V.V., Yeliseyeva O.Y., Ivashenko V.P. et al. Representation and processing of knowledge graf-dynamics associative machines: Monograph; BSUIR, 2001. (rus)

[Ivashenko, 2009] Ivashenko V.P. Semantic technology of engineering knowledge bases // BSUIR's Papers. – 2009. - No.7. – P.44-51. (rus)

[Kobozeva, 2009] Kobozeva I.M. Linguistic semantics: guide. 4th ed. – M.: Book House "LIBROKOM", 2009. (rus)

[Melchuk, 1974] Melchuk I.A. Experience of the theory of linguistic models «Meaning ⇔ Text». – M.: Science, 1974. (rus)

[Popov, 2004] Popov, E. V. Natural language interaction with the PC / E. V. Popov. – Moscow. : Yeditorial URSS, 2004. (rus)

[Yeliseyeva, 2011] Yeliseyeva O.Y. The use of artificial intelligence technologies in language teaching // Open Semantic Technologies for Intelligent Systems (OSTIS-2011) – Minsk, BSUIR, 2011. – P. 363 – 370. (rus)

[Yeliseyeva, 2012] Yeliseyeva O., Kim Y. Intelligent Tutoring System for Belarusian as a Foreign Language // Artificial Intelligence Driven Solutions to Business and Engineering Problems : Galina Setlak, Mikhail Alexandrov, Krassimir Markov (Eds.) - I T H E A®, Rzeszow, Poland – Sofia, 2012, Bulgaria. – PP. 93 – 101.

[Yeliseyeva, 2014] Yeliseyeva O.Y. Semantic conceptual design of natural language interface of intelligent system // Open Semantic Technologies for Intelligent Systems (OSTIS-2014) : V.V.Golenkov at al (ed.) – Minsk: BSUIR, 2014.

[Shvedova, 1998, 2002] Russian semantic dictionary. Dictionary, systematized on the word classes and values: In 6 vol. / N.Y. Shvedova (ed.). - RAS. Russian Language Institute, M.: Azbukovnik, 1998, 2002. (rus)

[http://www.ostis.net] URL: http://www.ostis.net – an open source project Open Semantic Technology for Intelligent Systems.

Authors' Information

Olga Yeliseyeva – *PhD, associate professor of the Department of Applied Linguistics, Faculty of Philology at the Belarusian State University, 51 Napoleona Ordy street, appt.118 Minsk, P.O. Box: 220045, Belarus, e-mail: volga.eliseeva@gmail.com*

Major Fields of Scientific Research: E-learning, Artificial Intelligence, Natural Language Processing

Kim Yury – *Master course student at Yonsei University, Seoul, 28 Tikotskogo street, appt. 83 Minsk, P.O. Box 220119, Belarus e-mail: 6rikim@gmail.com*

Major Fields of Scientific Research: Corpus Linguistics, Information Retrieval.

DYNAMIC VOCABULARIES FOR STUDYING INTERNET NEWS[4]

Mikhail Alexandrov, Daria Beresneva, Alexander Makarov

Abstract: *Nowadays there are many toolkits (methods and software) for automatic topic identification of documents. However economists, sociologists, politicians need tools not only for topic identification but also for analysis of changes in given topics related to time. In the paper we propose a simple technology, which could help to solve such a problem. For this: selected publications are distributed on sets associated with consequent time intervals, keywords are extracted from each document set, and these keywords are combined to reflect their dynamics. These combinations are named 'dynamic vocabularies'. We present two programs for building dynamic vocabularies and then we demonstrate our technology on real example related to the topic of Euro integration of Ukraine (2013-2014)*

Keywords*: dynamic vocabularies, topic identification, Internet-sociology.*

ACM Classification Keywords: *I.2 Artificial Intelligence.*

Introduction

Dynamic of social-economical and social-political processes is an object of consideration of economists, sociologists and politicians. To study such a dynamics the mentioned specialists have to read many materials circulating on the Internet and distributed in time. These efforts can be essentially reduced when one has initial knowledge (information) about topic(s) under consideration. Just for this case we propose a technology based on so-called 'dynamic vocabularies'. Such vocabularies are keywords lists, which should be extracted from publications and then combined by a special way. The publications are supposed to refer to a certain period of time. The proposed technology is demonstrated on the real example related to protests in Ukraine (2013-2014).

Dynamic vocabularies were considered in [Alexandrov, 2001; Makagonov, 2006]. In this paper we use new tools and propose new realization of dynamic vocabularies.

In section 2 we present tools for building dynamic vocabularies. Section 3 describes the results of the experiments. Section 4 contains conclusions.

Tools

Source of information

Initially several reliable Internet resources are chosen. One should not change these resources in order to keep the quality of information. We also fix the time of analysis, time step, and therefore we obtain a series of intervals on the time axis. Then all the collected documents are distributed between these intervals. Therefore, for *n* intervals we have just *n* corresponding document sets. To find documents one can use any usual search engines (Google, etc.). In our work we use our own crawler, which takes into account the title of topic with its keywords, sources of information, and time period.

Keyword extraction

Keyword lists are built for each document set. To select keywords we use here so-called criterion of specificity.

<u>Definition</u>: The level of specificity of a given word $\boldsymbol{w}$ in a given document corpus is a number $K \geq 1$, which shows how much its frequency in the document corpus $f(\boldsymbol{w})$ exceeds its frequency in the General Lexis $F(\boldsymbol{w})$:

[4] Work done under partial support of the Catholic University of San Pablo (grant FINCyT-PERU)

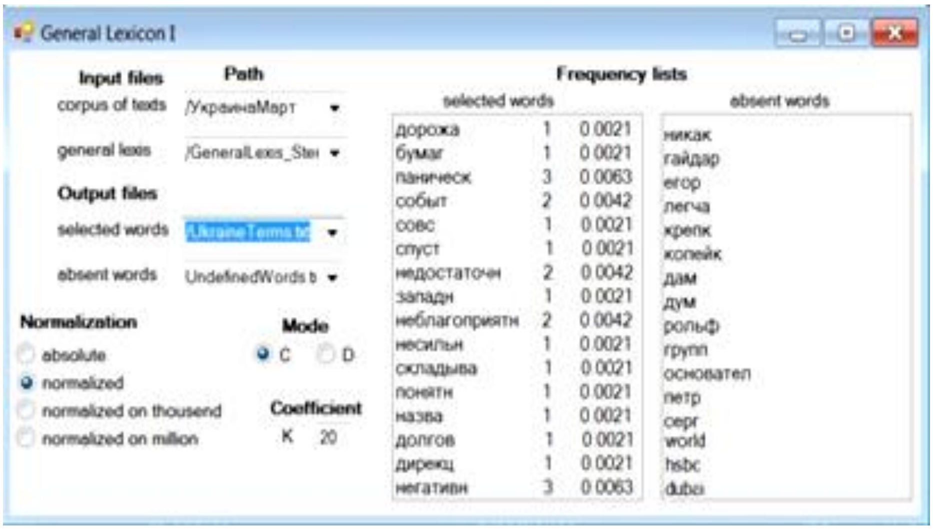
K = *f*(**w**) /*F*(**w**). Speaking 'General Lexis' we mean General Lexis of a given language (it is Russian in our case). We use here the Lexis from [Sharoff, http].

Example: Let we have the following data for the word 'protest': *f(protest)*=0,8*10^{-6}, *F(protest)*= 0,5*10^{-8}. Let the critical value *K*=100. It is easy to see that *f*(.) > *K***F*(.). It means that the word 'protest' has the level of specificity more than 100 and therefore this word will be selected. The threshold 100 is taken here only as an example.

In the real practice this value is determined experimentally: we should not lose many useful words and from the other hand we should not have the many unnecessary ones. In particularly, in our experiments we used *K*=50.

The criterion of specificity is calculated by the program LexisTerm, which is a free software developed in Peru [Lopez, 2011]. It should note that LexisTerm can use two modes: the corpus mode and the document mode. In the first case the program considers all documents as one large document. The definition presented above refers to this mode. In the corpus mode LexisTerm selects words being specific for the whole corpus but it loses words being specific for individual documents. In the second case the program considers each document independently. In the document mode all specific words are selected.

In our work we use a new version of the mentioned program - LexisTerm-I (International). Unlike LexisTerm, this program: (a) allows to process both English and Russian texts; (b) selects those words that satisfy the criterion of specificity and those that are absent in the General Lexis. A screenshot of the LexisTerm-I interface is shown in Figure 1.



General Lexicon I

Input files — Path

corpus of texts: /УкраинаМарт

general lexis: /GeneralLexis_Ste

Output files

selected words: /UkraineTerms.tx

absent words: UndefinedWords.t

Normalization

○ absolute

● normalized

○ normalized on thousend

○ normalized on million

Mode

● C ○ D

Coefficient

K 20

Frequency lists

selected words			absent words
дорожа	1	0.0021	никак
бумаг	1	0.0021	гайдар
паническ	3	0.0063	егор
событ	2	0.0042	легча
совс	1	0.0021	крепк
спуст	1	0.0021	копейк
недостаточн	2	0.0042	дам
западн	1	0.0021	дум
неблагоприятн	2	0.0042	рольф
несильн	1	0.0021	групп
складыва	1	0.0021	основател
понятн	1	0.0021	петр
назва	1	0.0021	серг
долгов	1	0.0021	world
дирекц	1	0.0021	hsbc
негативн	3	0.0063	dubai

Figure 1. Interface of the program LexisTerm-I

Keyword are selected in the corpus mode and not in the document mode, because the individual documents prove to be very subjective with respect to events under consideration. At the final stage all keyword lists are manually corrected by an expert.

Dynamic vocabularies

Keywords are selected on a weekly basis. Using the resulting lists we can combine them in order to form 4 types of dynamic vocabularies. They concern all weeks and each week:

1. the common words for all weeks or for a given part of all weeks,
2. words that belong to the current week and don't belong the previous week, we name them 'new words',
3. words that belong to the current week and the previous week, we name them 'repeated words',
4. words that don't belong to the current week but belong to the previous week, we name them 'old words'.

Speaking 'common words' we mean the words that appear in *m* weekly keyword lists where $m \geq M$ and *M* is a given threshold. The Figure 2 shows the distribution of keywords between weeks.

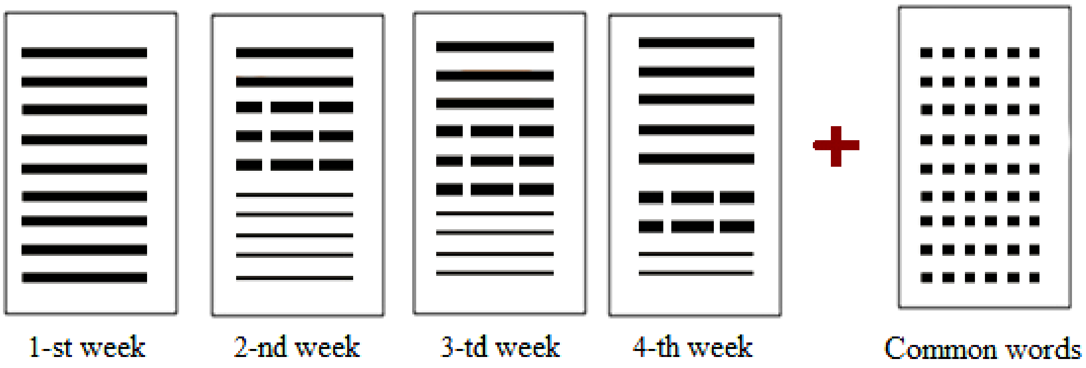
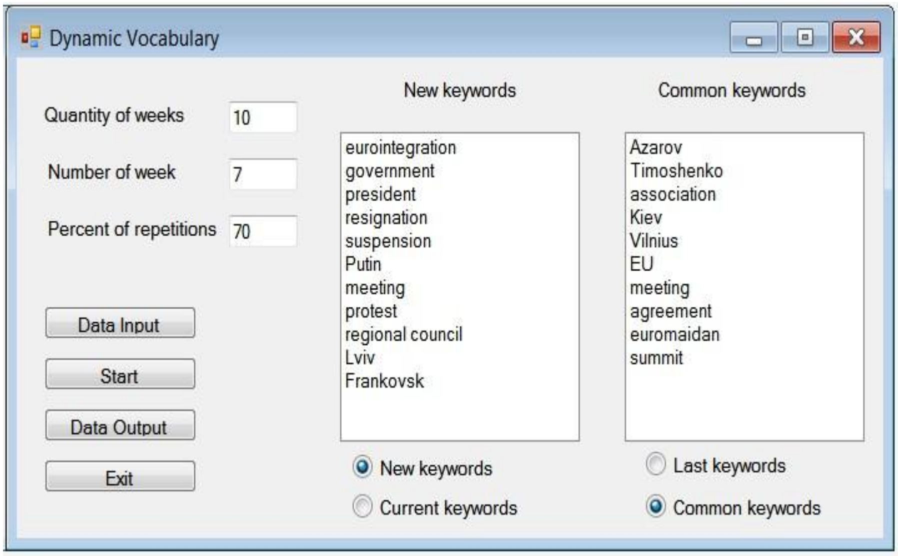


Figure 2. Dynamic vocabularies

Here: thick lines are the new words, dashed thick lines are the repeated words, thin lines are the old words, and dotted thick lines are common words.



Quantity of weeks 10

Number of week 7

Percent of repetitions 70

New keywords	Common keywords
eurointegration	Azarov
government	Timoshenko
president	association
resignation	Kiev
suspension	Vilnius
Putin	EU
meeting	meeting
protest	agreement
regional council	euromaidan
Lviv	summit
Frankovsk	

● New keywords ○ Last keywords

○ Current keywords ● Common keywords

Figure 3. Interface of the program DynVoc (keyword lists are translated to English)

Dynamic vocabularies are built automatically using the program DynVoc (Dynamic Vocabularies). The input of the program is a set of weekly keyword lists and the output of the program is a set of dynamic vocabularies. The interface of the program is presented on Figure 3.

Experiment

In the experiment we studied publications related to mass protests in Ukraine. The principal topic of the protests was Ukrainian Eurointegration. The protests lasted approximately 23 weeks (October 2013 – March 2014). We considered only the first 12 weeks (October 2013 – December 2013), which defined the further development of the Ukrainian events. We used 4 popular Russian Internet editions "Arguments and Facts", "Russia Today", "RIA Novosti", "Gazeta.ru". All papers were divided on weeks. On average we had 15 papers per one week.

Figure 4 shows vocabularies concerning the 7-th week (November 8-15, 2013). By that moment the suspension of the process of European integration had led to the numerous protests in Kiev, Lviv and other Ukrainian cities. The protesters demand the resignation of the Ukrainian Prime Minister Azarov. The Russian President Putin asks the EU to depoliticize the topic of Ukrainian Eurointegration. From other hand the perspectives of integration or partnership with the ex-Soviet countries as Kazakhstan, Armenia etc. are not already discussed.

New	***Repeated***	***Old***	***Common***
State	Azarov	Armenia	association
depoliticized	agreement	EuroSummit	power
Eurointegration	cooperation	Kazakhstan	East
Evromaydan		chancellor	European Union
compensation		conflict	Kiev
credit		Moscow	agreement
Lviv		dishonest	Ukraine
international		required	membership
meeting		partnership	Yanukovych
Regional Council		justice	Azarov
opposition		opposed	Tymoshenko
resignation		ratification	
demanded		Rogozin	
government		Russia	
president		marketing	
premier		speculation	
Suspend		customs	
protest		technology	
Wrongful		criminal	
Putin		economic	
resolution		electric power	
sovereignty			
Frankivsk			

Figure 4. Example of dynamic vocabulary (keyword lists are translated to English)

One who knows the situation in Ukraine may agree that in totally this dynamic vocabulary reflects the contents of protests at that moment in November. Here: Lviv and Frankivsk are the regional centers; Armenia and Kazakhstan are ex-Soviet countries; sovereignty, compensation and credits are related to economical aspects of Eurointegration; President Putin (Russia) says about depoliticization; suspended agreement, protests, Euromaydan and Eurosummit are in one chain; the other chain is President Yanukovich (Ukraine), dishonest activity and criminal environment; the third chain is Mr. Rogozin from the Russian Government, Ukrainian market, and cooperation with Russia; etc.

Note. Common words here are the words that appear in more than 50% of weekly keyword lists. The threshold 50% was assigned by user (one of the authors). So, these words may be absent in some weakly keyword lists. Just for these reason some common words can be included both in the last column and simultaneously in one of the first three columns, see the words Azarov and agreement

Conclusions

The results of completed work are:

- the simple method for studying dynamics of topics;
- software for building dynamic vocabularies;
- experiment with the real data.

The proposed approach 'works' well when user (expert) has an initial knowledge about events, objects, persons related to the topic of interest and he/she wants only to meet with dynamics of this topic. Otherwise it is necessary to use the other ways as document annotation or selection of representative documents from given document sets.

In future we suppose:

- to automate the process of keyword lists correction;
- to include the procedures of visualization to the program DynVoc.

Bibliography

[Alexandrov, 2001] Alexandrov, M: Dynamic domain-oriented dictionaries as a tool for revealing tendencies in development of some scientific and technological disciplines and interaction between them. // Mexican National Council on Sciences and Technologies (CONACyT), Reg.No. 39011-a.

[Lopez, 2011] Lopez, R., Alexandrov, M.:Tejada, J: LexisTerm - The Program for Term Selection by the Criterion of Specificity, // Proc. of 4-th Intern. Conf. on Intelligent Inform and Engineering Systems, ITHEA Publ, 2011, p. 8-15

[Makagonov, P., 2006] Makagonov. P., Figueroa, A., Gelbukh, A. Studying Evolution of a Branch of Knowledge by Constructing and Analyzing its Ontology.// Springer, LNCS, 2006, 10 pp.

[Sharoff, http] http:// www.artint.ru/projects/frqlist.php, General Lexis of Russian

Authors' Information

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

Daria Beresneva – *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, Moscow Region, 141700, Russia*

e-mail: dejame@ yandex.ru

Major Fields of Scientific Research: mathematical modeling, world economy

Alexander Makarov – *Researcher, Russian Presidential Academy of national economy and public administration; Prosp. Vernadskogo 82, bld. 1, Moscow, 119571, Russia;*

e-mail: mackarov54@ gmail.com

Major Fields of Scientific Research: data mining, Internet sociology

CONCEPTS IDENTIFICATION IN A DOCUMENT WITH NETWORK ACTIVATION METHOD[5]

Dmitry Stefanovskiy, Mikhail Alexandrov, Ales Bourek, Tomas Hala

***Abstract**: In the paper concepts are considered as groups of semantically related keywords. To reveal such groups we propose the technology based on a) constructing network of terms on the basis of procedure of segmentation; b) revealing groups of terms using network activation method. We demonstrate the proposed technology on two examples related to medical problems and social problems. The results proved to be very promising with the point of view of experts. The presented work is a pilot study.*

***Keywords**: topic identification, text classification, network activation.*

***ACM Classification Keywords**: I.2.7 Natural Language Processing.*

Introduction

Concepts identification is one of the most important elements of natural language processing whose results are used in ontology construction, topic presentation, document classification, etc. The famous manuals on Information Retrieval [Baeza, 1999; Manning, 2008] contain descriptions of different forms of concepts presentation and different methods for their formation. In the paper we use less-known procedure: segmentation and network activation.

The paper is built by the following way. In section 2 we describe steps of the proposed algorithm. Section 3 presents the results of experiments. Section 4 contains conclusions

Algorithms

Keyword selection

The initial stage of text processing is a typical one: we exclude all stop words and general lexis. Here the criterion of term specificity proves to be very useful. This criterion was described and studied in [Lopez, 2011]. Term specificity with respect to a document is the relation $K = f(\mathbf{w}) / F(\mathbf{w})$, where $f(\mathbf{w})$ is frequency of a given word $\mathbf{w}$ in a document and $F(\mathbf{w})$ is frequency of this word in some basic corpus. Usually the National corpus of a given language is used. As a rule the threshold for word specificity K = 5-7 provides good results.

Building network

To build a network of selected terms (hereinafter we will name them 'keywords') one should determine the pairwise relations between them. The key-position here is text fragmentation, which allows to reveal the joint term occurrences in each fragment and calculate the correlation between terms. The problem of fragmentation was the subject of consideration in a few number of papers. One of the promising ways consists in using external information as the Word Net [Aung, 2013]. In our work we try to use only the internal resources

Option 1

Fragmentation is realized by means of running window. Here is some ways to determine the width of this window:

1.1. The width of window should be fixed and equal 2-5 sentences. These values take into account well-known linguists opinion that: a) two terms are strong related when they both are collocated in the same sentence; b) the relation between terms are still essential when these terms are taken from adjacent sentences; c) the relation

[5] Work done under the partial support of the Catholic University of San Pablo (grant FINCyT-PERU)

between terms is weak or absent when there are one or more sentences between them. Such an approach is enough rigid. It does not take into account a priori information concerning density of keywords, etc.

1.2. Let *N* is a number of keywords, *n* is a number of sentences, and *m* is a number of expected concepts. Therefore on average one concept is reflected by *N/m* keywords and the density of these keywords is equal *k*=*N*/(*m***n*). The width of window measured in sentences should be more then 1/*k* to provide at least one occurrence of keyword related to a concept. For example, if *N*=50, *n*=200, and *m*=5, then the width of window is equal 20 sentences or more.

Option 2

Fragmentation is based on document structure. Speaking 'structure' we mean paragraphs (indentions) formed by author(s) of these documents, or cells of tables when we deal with textual tables, etc.

Network activation method

This method is a simplified variant of the spreading activation method proposed by A. Troussov and studied in detail in his publications [Troussov, 2008; Troussov, 2009]. Initial information for this method is a network: its nodes are keywords and its arcs reflect the relations between nodes. The weight of arc between *i*-th node and *j*-th node is accepted to be equal the correlation between *i*-th keyword and *j*-th keyword.

On the stage of preprocessing all weak connections are eliminated. The threshold for these weak connections is assigned by a user. Usually it is equal to 10%-20% of the maximum value of correlation between nodes.

Then the iterative procedure is implemented using one of two options

Option 1

Each node is a source of heat. This heat is transferred to other nodes over arcs. If *i*-th node and *j*-th node are directly connected then the heat coming from *i*-th node to *j*-th node is equal w_{ij} , $0\leq w_{ij} \leq 1$. It takes one iteration. On the next iteration this heat diffusion continues

Option 2

A user himself/herself selects the sources of heat. Usually it is the most interesting nodes being the foci of concepts. Then the process of heat diffusion continues as it were described above.

The number of iterations is determined subjectively. For example, it is possible to set a threshold related to maximum difference in heats on the network. When this difference achieves this threshold then the iterative process finishes.

The simplified version of this method consists in the following:

- all weights are equal 1,
- only one iteration is implemented.

To decompose a heated network on its components we use the typical way: a threshold is set and then all arcs having the weight less then this threshold are eliminated. As a result we have a certain number of isolated groups of nodes. Just these groups define concept description. One should remind that the threshold here refers to any function of closeness, for example, coefficient of correlation.

To find this threshold we use the criterion of stability. Namely, the program automatically changes the threshold and calculates the number of groups. The jump of this number is an indicator of possible threshold. Figure 1 demonstrates the typical dependence between the thresholds and the number of groups.

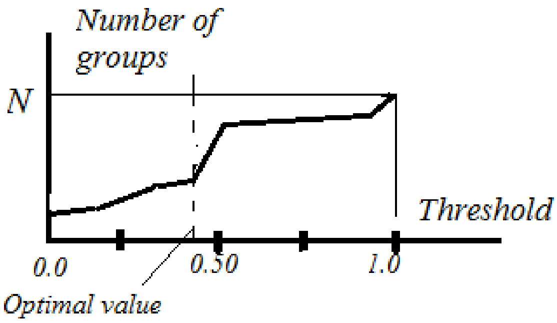


Figure 1. Thresholds and number of groups

Experiments

Medical problems

The first corpus of documents (20 documents in English) is the materials of one EU-project related to some problems of healthcare. These documents contain answers on two questions: "Which aspects or characteristics of the current healthcare facilitate patient empowerment?" and "Which aspects or characteristics of the current healthcare do not facilitate patient empowerment?". We gathered together all answers on the first question and titled this large document as Advantages. We did the same with the answers on the second question and titled this large document as Barriers. Then we applied the technology described above to both documents. Conditions of the experiment are:

- segmentation is done using option 2,
- activation is done using option 2.

The figure 2 and figure 3 demonstrates the part of results.

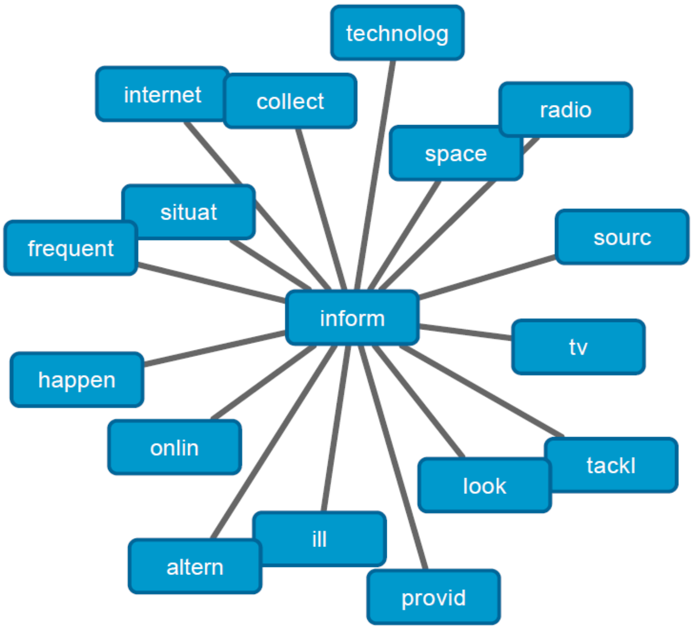


Figure 2. One of the concepts related to Advantages

Here is the expert opinion about the contents of concept presented on figure 2: “This concept suggests the important role of communication technologies visible from the words source, online, Internet, TV, radio, that is necessary to use in order to collect, tackle and provide information”.

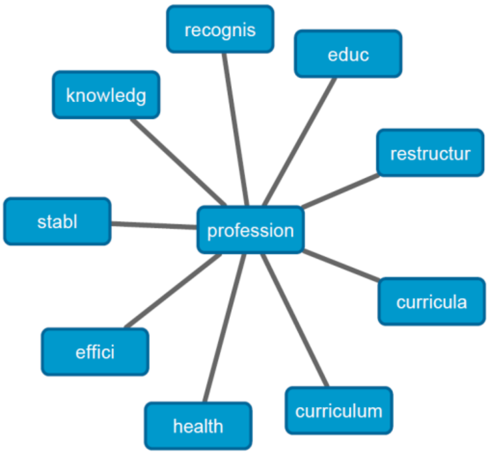


Figure 3. One of the concepts related to Barriers

Here is the expert opinion concerning the contents of concept presented on figure 3: “This concept suggests the need to address the educational process in forming new healthcare professionals that would not become barriers to patient empowerment. It shows the need to restructure HC professional educational curricula in order to assure a stable, efficient, knowledge based environment recognizing and addressing the health issues”.

Social problems

The second corpus of documents (10 documents in Czech) was downloaded from the Internet. The principal topic of all documents is corruption. Conditions of the experiment are:

- segmentation is done using option 1, the first way,
- activation is done using option 1.

The most interesting concept is presented on figure 4.

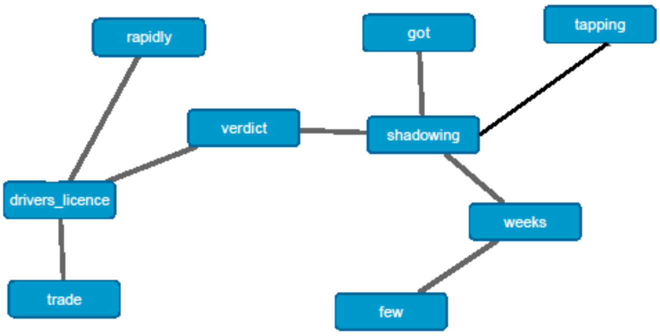


Figure 4. One of the concepts related to the problem of abuse with driver licences

The contents of this group obvioiusly says about a corruption related to driver licences. To prove the crime it is necessary to organize shadowing and tapping. All this takes a certain time up to few weeks. When the results of shadowing had been got then the materials were sent to a court and the verdict of the court was rapidly obtained

Conclusion

In the paper we demonstrate possibilities of network activation method to reveal concepts from documents in the form of grouped terms. This method uses physical analogies, which promote manifestation of relations between terms. To obtain a network of terms the procedure of segmentation is used. The results of experiments with two document sets show good results.

In future we suppose to use physical and mathematical analogies in the problem of concept identification not only for grouping terms but also for selecting terms and building network. In particularly, we intend to use algorithm described in [Carpena, 2009]. The authors of this work use some principles of quantum mathematics for term selection.

Bibliography

[Aung, 2013] Aung N.M.M., Maung S.S.: Semabtic-based text block segmentation using word net. // Intern. Journ. of Computer and Communication Engineering, vol.2, No.5, 2013, 4 pp.

[Baeza-Yates, 1999] Baeza-Yates, R., Ribero-Neto, B.: Modern Information Retrieval. Addison Wesley, 1999

[Carpena, 2009] Carpena P., et al: Level statistics of words: finding keywords in literary texts and symbolic sequences // In: Physical Reveiew, E-79, 035102(R), 2009, 4 pp.

[Lopez, 2011] Lopez, R., Alexandrov, M.:Tejada, J.: LexisTerm - The Program for Term Selection by the Criterion of Specificity, // Proc. of 4-th Intern. Conf. on Intelligent Inform and Engineering Systems, ITHEA Publ, 2011, p. 8-15

[Manning, 2008] Manning C.D., Raghavan P. Schutze H.: Introduction to Information Retrieval, Cambridge University Press. 2008

[Troussov, 2008] Troussov, A., et al : Mining Socio-Semantic Networks Using Spreading Activation Technique //. Proc. of I-KNOW -2008 and I-MEDIA 2008, Graz, Austria, 2008, pp. 405-412

[Troussov, 2009] Troussov, A., et al : Spreading Activation Methods. // In: Dynamic and Advanced Data Mining for Progressing Technological Development, IGI Global, USA, 2009, 31 pp.

Authors' Information

Dmitry Stefanovskyi – *Assoc. Prof., Ph.D, 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

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

Ales Bourek – *Senior lecturer, Masaryk University, Brno, Czech Republic; Head of Center for Healthcare Quality, Masaryk University. Kamenice 126/3, 62500 Brno, CZ.*

e-mail: bourek@ med.muni.cz

Major fields of interest: reproductive medicine – gynecology, health informatics, healthcare quality improvement, health systems

Tomas Hala – *Senior lecturer, Mendel University in Brno, DI FBE, Zemědělská 1, 61300 Brno, Czech Republic;*

e-mail: thala@ pef.mendelu.cz

Major Fields of Scientific Research: text processing, typesetting, typography.

COMPUTER AIDED ENGINEERING AND SIMULATION

DIRECTED MOVEMENT OF A FINGER MECHATRONIC TO IMPROVE THE VISIBILITY OF ARGOPECTEN PURPURATUS'S KIDNEY USING COMPUTER VISION

Sonia Castelo-Quispe, Roxana Flores-Quispe, Yuber Velazco-Paredes, Raquel Esperanza Patiño-Escarcina and Dennis Barrios-Aranibar

***Abstract**: The shucking of the Argopecten Purpuratus (scallop) is a process where the viscera and the kidney are removed. The use of computer vision is an alternative to automate this process, however the extraction of the kidney is the most difficult task because in many cases the scallop's stem covers the kidney avoiding the kidney extraction. Due to this circumstance the whole shucking process fails. This paper proposes a method of computer vision to determine the movement direction of a finger mechatronic, which move the stem in the opposite direction of the kidney improving its visibility. In order to achieve this improvement we use image segmentation, logical operations on images, geometric and mathematical calculations, etc. The proposed method provides the efficiency 99.38% for removing Argopecten Purpuratus's kidney.*

***Keywords**: segmentation, computer vision, image processing, colors spaces, robotic finger, Argopecten Purpuratus.*

***ACM Classification Keywords**: J.7 Computers in other systems.*

Introduction

The exportation of Argopecten Purpuratus (scallop) has increased significantly over the past few years. This increase is due to high demand that exists in countries like United States and France, not forgetting the Asian Continent, which is the biggest importer of Argopecten Purpuratus in the world. In factories, this increase means additional number of human inspectors working on processing lines. In this regard, process automation has become a crucial factor in industries [Narendra, 2010] [Gumus, 2011] [Yachida, 1980].

Many researches have been conducted with computer vision systems in order to automate processes in field of foodstuffs. In [Arnarson, 1994] a shape classification technique in computer vision called PDL-HM is proposed. It combines the selection of morphological structuring elements from contour description languages and the extraction of shape characteristics to fish species classification in an industrial environment. In [Misimi, 2007] and [Misimi, 2008] computer vision methods are proposed to evaluate the color of Atlantic salmon fillets. In [Rodriguez, 2011] a method is developed to classify the Argopecten Purpuratus based on determination of weights by conversion and adjustment factors.

Similarly, researches about manipulator arms have achieved new progress in conjunction with computer vision systems, which allows controlling the movements with visual feedback [Li, 1996]. A robotic apparatus is presented in [Lefebvre, 1993], it aims to automate pulp sampling of potatoes in order to detect viral diseases using two computer vision approaches. In [Martinez-de Dios, 2003] is presented a robotic system for fish feeding and an underwater robot for autonomous pond cleaning.

The rest of this paper is organized as follows. Section 2 describes the method to determinate of movement direction of a finger mechatronic. Section 3 shows the results of the method. Finally, Section 4 presents conclusions.

Approach

This paper proposes a technique to determinate the movement direction of a finger mechatronic in order to improve the visibility of the Argopecten Purpuratus's kidney. It will facilitate the full removal of the kidney, automating overall process of shucking of the Argopecten Purpuratus (Scallop).

The Argopecten Purpuratus's stem covers the kidney avoiding the kidney extraction by automated methods. Due to it the whole shucking process fails. In this context the aim of finger mechatronic moves the Argopecten Purpuratus's stem in the opposite direction to the localization of kidney through a digital images analysis using computer vision techniques. The Fig. 1 shows Stem and Coral of Argopecten Purpuratus.

The difficulty of this problem increases due to the high variability of scallops, such as their shape, colour or size. To achieve the objective proposed in this paper, it is necessary to perform various processes, which are desribed in detail below.

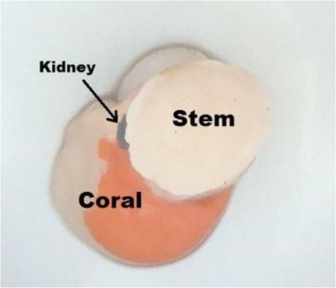


Figure 1. Scallop (original image): Stem, Coral and Kidney

Segmentation of the Region of Interest (ROI)

The ROI includes the area of the image where Scallop is presented. The segmentation process starts by converting the image from RGB to YCrCb color space. Then, the Cb channel is used to segment the images, with a dynamic threshold based on color histogram for binarization. The threshold is computed using the average of the color histogram. This value will be different in each image because each image is affected by external factors such as luminosity, which cause changes in the color and tonality.

Then, the noise from image is deleted by discarding of areas. Small areas which are outside or inside the ROI (Scallop) are deleted. Finally, an AND operation was applied between the original and segmented image. Thus, subsequent processes will work with the ROI of original image, as shown in figure 2 on the right.

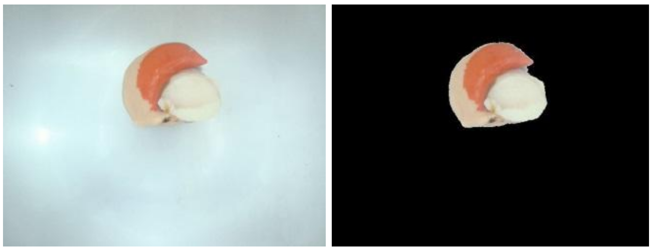


Figure 2. Original and segmented Image (Scallop)

Segmentation of the Stem and Red Coral

Segmentation of the Stem is important because the finger mechatronic shall be positioned over the Stem. The Stem was segmented using the B channel of RGB color space. Then, it is binarized with an experimental threshold, which was calculated by measuring the brightness in the background original image. Finally, a discarding of areas is performed, due to it is possible to detect areas which do not correspond to the Stem. The figure 3 on the right showed a stem segmented.

The images of database shows that the kidney is always located at one end of Red Coral (RC). In addition, segmentation of the Red Coral will help us later to find the region closest to kidney. Segmentation of the Red Coral was performed using the Cb channel of YCrCb color space. Then, it is binarized with experimental threshold of 56 and finally a discarding of areas is performed. The figure 4 on the right showed a segmented Red Coral.

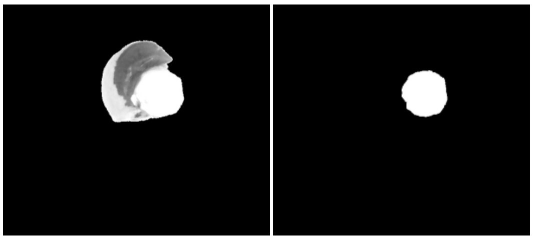


Figure 3. Segmentation of Stem

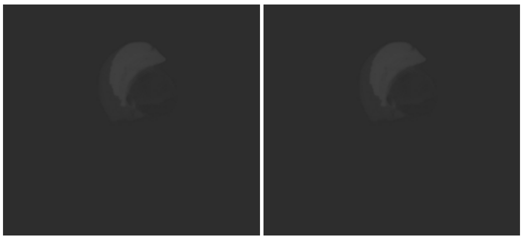


Figure 4. Segmentation of Red Coral

Determination of the movement direction of finger mechatronic

To determine the movement direction, it is necessary to determine the location of two specific points. The first point will be the center of mass of the Stem, and the second point will be a point on Red Coral's surface that is closest to the kidney.

Obtaining center area of the Stem: To obtain the location of this point is used the image of stem segmented. It is computed calculating the vertical and horizontal average of segmented image, which will represent *x, y* coordinates of the center of the Stem.

Obtaining Red Coral Point nearest to kidney: As it can be seen in the image database, the position and shape of the scallop are not always the same one. So, it is important to take into account the location of the Stem with respect to the Red Coral, since the position of scallop in the picture is not always the same.

One of the four extreme points of Red Coral (top, bottom, right and left) will be identified as the nearest point to the kidney, as shown in figure 5 (blues points). In order to determinate this position, the center point of scallop (red point in figure 5) and the center point of Red Coral (brown point in figure 5) will be attached to trace a straight line R1, as shown in figure 6.

Then a Cartesian plane is projected on the center point of Red Coral as it is shown in figure 6. According to the location of line R1 in the Cartesian plane, the position of the Stem with respect to Red Coral on the image can be determined. In the figure 6, the line R1 is localized in the fourth quadrant of the Cartesian plane.

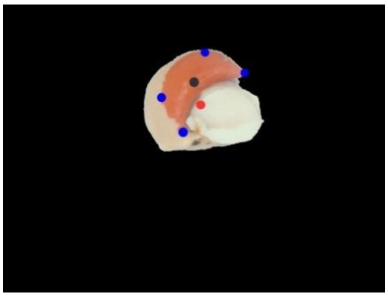


Figure 5. Highlights: Extreme points of RC (blue), center point of the Scallop (red) and center point of the RC (brown).

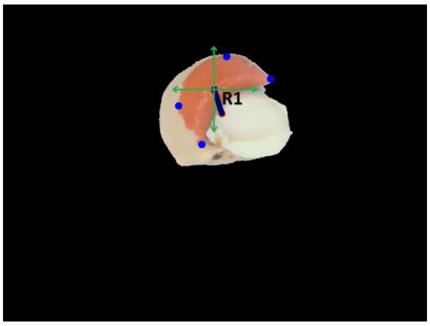


Figure 6. Localization of the straight line R1 in the Cartesian plane.

After that we determine a quadrant of the Cartesian plane, which contains the straight line. Then the two extreme points being the closest to this quadrant are chosen. After these operations we draw other three straight lines: the first straight line *D1* (equation 1) is formed by the two points chosen above, the second straight line *D2* (equation 2) and the third straight line *D3* (equation 3) are formed together with each point (points chosen above) having the closest neighboring extreme points of the Red Coral as it is shown in figure 7.

$$D1: m_{slc} = \frac{y2 - y1}{x2 - x1} \tag{1}$$

where m_{slc} represents the slope of the straight line *D1*, and (*x*, *y*) represents the location of points in the image on the *x-axis* and *y-axis* respectively.

$$D2: m_{slc1} = \frac{y2 - y3}{x2 - x3} \tag{2}$$

where m_{slc1} represents the slope of the straight line *D2*, and (*x*, *y*) represents the location of points in the image on the *x-axis* and *y-axis* respectively.

$$D3: m_{slc2} = \frac{y2 - y4}{x2 - x4} \tag{3}$$

where m_{slc2} represents the slope of the straight line *D3*, and (*x*, *y*) represents the location of points in the image on the *x-axis* and *y-axis* respectively.

Then we calculate the distances of these three straight lines *D1*, *D2* and *D3*, and the straight line with the greatest distance is chosen as it is shown in figure 8. One of two endpoints of this chosen line could be selected as the point of Red Coral being the nearest one to the kidney.

Having obtained these two extreme points (blue and green points in figure 8) the distance of each extreme point towards the center point of the scallop (brown point in figure 8) , is calculated. Finally, the closest point (less distance) to the center of the coral is chosen. It is considered as the point being the nearest one to the kidney as it is shown in figure 9.

The geometric and mathematical calculations have been done because the position of the scallop in the image is not always the same one. The additional circumstance is irregular shapes of the scallop.

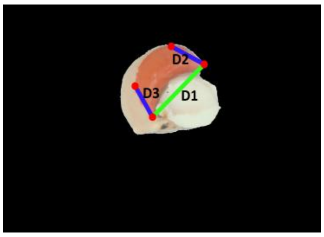


Figure 7. Trace of lines D1, D2 and D3

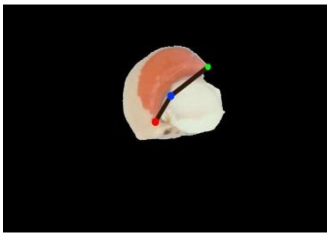


Figure 8. Two extreme points of Red Coral

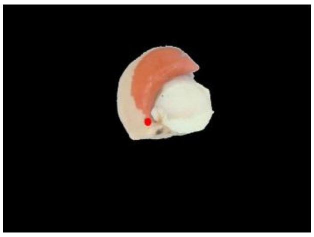


Figure 9. Point on Red Coral's surface that is closest to the Kidney

Movement direction of finger mechatronic. We obtain the two points: the center point of the Stem (green point in figure 10) and the point on Red Coral's surface that is the closest one to the kidney (red point in figure 10). With these two points we can determine the movement direction of the finger mechatronic. For this we draw a straight line *R2* containing these two points as it is shown in figure 10.

Then we project the straight line *R*2, which will have: a) the center of Stem as the starting point and b) the distance being equivalent to the length of the straight line to segment *R2.*

Thus, the movement of the finger mechatronic starts in the center of the Stem Area in opposite directions to the localization of the nearest point to the kidney. It follows the path of the projected straight line *R*2, as it is shown in figure 11.

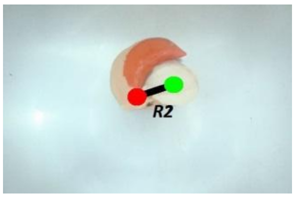


Figure 10. Trace of the straight line R2

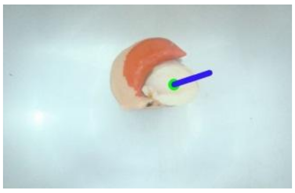
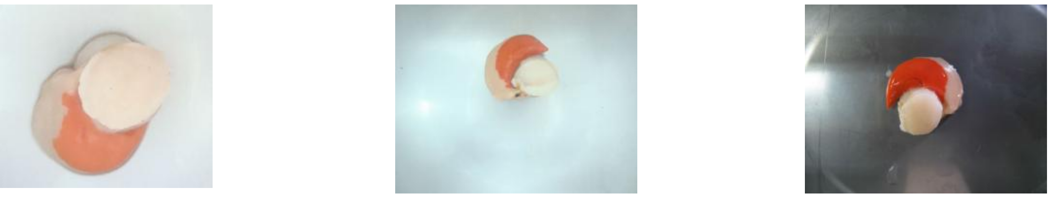


Figure 11. Movement direction

Results

The images used for the experiments were taken from the SSCCA-5 image database of the Research and Software Development Center of the San Agustin National University in Peru. This image database was developed in order to test the Shucking and Codification Process of the Argopecten Purpuratus. For these experiments a total number of 969 images were used: 693 imagens with a dark blue background, 176 dark lead background, and 100 of a light background. A sample of each image can be seen in figure 12. The dimension of the images is 1920x1080 pixels.



(a) Light background

(b) Dark Blue(DB) background

(c) Dark Lead(DL) background

Figure 12. Image Database according to the background

Tests to evaluate the efficiency. A result is considered as the "Good" one when the direction taken by the finger mechatronic is contrary to the position of the kidney. It makes the kidney to be visible. On the other hand, a result is "Bad" if the finger movement is directed towards any other direction.

The quantitative results of experiments are summarized in tables 1 and 2. The results show the level of efficiency achieved according to the background image. Thus, an overall performance of 99.38 %was achieved on the test set.

Table 1. Results of experiments with light background, Dark Blue(DB) background, Dark Lead(DL) background.

CODE	light background		Total	Accuracy (%)	DB background		Total	Accuracy (%)	DL background		Total	Accuracy (%)
	Good	Bad			Good	Bad			Good	Bad		
Code 1	26	0	26	100	104	2	106	98.11	0	0	0	100
Code 2	64	0	64	100	336	0	336	100	99	0	99	100
Code 3	9	0	9	100	156	0	156	100	71	0	71	100
Code 4	1	0	1	100	88	4	92	95.65	6	0	6	100
Code 5	0	0	0	100	3	0	3	100	0	0	0	100
Total	100	0	100	100	687	6	693	99.13	176	0	176	100

Table 2. Efficiency in groups according to the background

BRACKGROUND	Efficiency in Established Groups			
	Total Images	Good	Bad	Efficiency (%)
Light	100	100	0	100
Dark Blue(DB)	693	687	6	99.13
Dark Lead(DL)	176	176	0	100
Total	969	963	6	99.38

Conclusion

In this work we propose a method that allows to determine the movement direction of a finger mechatronic and therefore to improve the visibility of the Argopecten Purpuratus's kidney. This method is based on location of the point on Red Coral's surface being the closest one to the kidney even if this point is not observed in the image. The results show the efficiency 99.38% for removing Argopecten Purpuratus's kidney.

Bibliography

[Narendra, 2010] V. G. Narendra and K. S. Hareesh, "Prospects of computer vision automated grading and sorting systems in agricultural and food products for quality evaluation," International Journal of Computer Applications, vol. 1, no. 4, pp. 1–9, February 2010, Published By Foundation of Computer Science.

[Gumus, 2011] B. Gumus, M. Balaban, and M. Unlusayin, "Machine vision applications to aquatic foods: A review," Turkish Journal of Fisheries and Aquatic Sciences, vol. 11, pp. 171–181, 2011.

[Yachida, 1980] M. Yachida and S. Tsuji, "Industrial computer vision in japan," Computer, vol. 13, no. 5, pp. 50 –63, May 1980.

[Arnarson, 1994] H. Arnarson and L.F. Pau, "Pdl-hm morphological and syntactic shape classification algorithm: Real-time application to fish species classification," Machine Vision and Applications, vol. 7, no. 2, pp. 59–68, 1994.

[Misimi, 2007] E. Misimi, J. Mathiassen, and U. Erikson, “Computer vision - based sorting of atlantic salmon (salmo salar) fillets according to their color level,” Journal of food science ,Wiley Online Library, vol. 72, no. 1, 2007.

[Misimi, 2008] E. Misimi, U. Erikson, and A. Skavhaug, “Quality grading of atlantic salmon (salmo salar) by computer vision,” Journal of food science ,Wiley Online Library, vol. 73, no. 5, pp. E211 E217, 2008.

[Rodriguez, 2011] A. Rodriguez, A. Diaz-Zea, R. Flores, M. Delgado, D. Barrios-Aranibar, and R. Patiño, “Argopecten purpuratus codification based on determination of weight by conversion and adjustment factors,” JCC’2011. XXX International Conference of the Chilean Computer Science Society, 2011, November 2011.

[Li, 1996] Y.F. Li and M.H. Lee, “Applying vision guidance in robotic food handling,” Robotics Automation Magazine, IEEE, vol. 3, no. 1, pp. 4 –12, Mar 1996.

[Lefebvre, 1993] M. Lefebvre, S. Gil, D. Brunet, E. Natonek, C. Baur, P. Gugerli, and T. Pun, “Computer vision and agricultural robotics for disease control: The potato operation,” Computers and Electronics in Agriculture, vol. 9, no. 1, pp. 85 – 102, 1993.

[Martinez-de Dios, 2003] J. R. Martinez-de Dios, C. Serna, and A. Ollero, “Computer vision and robotics techniques in fish farms,” Robotica, vol. 21, pp. 233–243, June 2003.

Authors' Information

Sonia Castelo Quispe – *Master Student in Computer Science, Universidade de São Paulo, Av. Trabalhador São-carlense, 400 – São Carlos, São Paulo, Brazil; e-mail: scastelo2@gmail.com*

Major Fields of Scientific Research: Information Visualization, Image processing, Computer vision and Machine learning.

Roxana Flores Quispe – *Professor and researcher at the San Pablo Catholic University and in the Research and Software Development Center of the Saint Agustín National University. Urb. Campiña Paisajista s/n Quinta Vivanco, Arequipa, Perú; e-mail: rfloresq@ucsp.edu.pe*

Major Fields of Scientific Research: Image processing, Computer vision and Artificial intelligence.

Yuber Velazco Paredes – *Professor and researcher at the San Pablo Catholic University and in the Research and Software Development Center of the Saint Agustín National University. Urb. Campiña Paisajista s/n Quinta Vivanco, Arequipa, Perú; e-mail: yvelazco@ucsp.edu.pe*

Major Fields of Scientific Research: Image processing, Computer vision and Artificial intelligence.

Raquel Patiño Escarcina – *Professor and researcher at the San Pablo Catholic University and in the Research and Software Development Center of the Saint Agustín National University. Campiña Paisajista s/n Quinta Vivanco, Arequipa, Perú; e-mail: raquel.patino@gmail.com*

Major Fields of Scientific Research: Computer vision, Image processing and Artificial intelligence.

Dennis Barrios Aranibar – *Professor and researcher at the San Pablo Catholic University and in the Research and Software Development Center of the Saint Agustín National University. Urb. Campiña Paisajista s/n Quinta Vivanco, Arequipa, Perú; e-mail: dennisbarrios@gmail.com*

Major Fields of Scientific Research: Robotic and automation applied to industry, Multiagent systems, and Machine learning.

MANAGING SIZE AND POSSIBLE GEOMETRY BEHAVIOR USING PARAMETERIZATION OF SKETCH GEOMETRY WITH A TOOLKIT APPLICATION IN THE AUTODESK INVENTOR PACKAGE

Evgeniy Ivanov, Zoya Ivanova, Olga Kalynychenko, Kostyantyn Lagosha

Abstract: *There was introduced three ways to create the chamfer element in package Autodesk Inventor applying the parameterization of sketch geometry with dependences for size management and possible behavior of geometry. Analysis and mathematical processing of reference data allowed using the reference value "Width Across Fields S" as the source parameter of chamfer element.*

Keywords: *chamfer, width across flats, parameterization, Autodesk Inventor package.*

ACM Classification Keywords: *I.3.5. Computational Geometry and Object Modeling – Boundary representations Constructive solid geometry (CSG) Curve.*

Introduction

The rapid development of computer technology and the creation of a multi-faceted automated system of computer graphics on its basis is the outstanding achievement in the last decades.

Modern computer graphics has a number of different applications, one of which is the creation of dynamic, virtual, multimedia environments and three-dimensional solid models.

The need to create and develop interactive graphical modeling of three-dimensional objects for various functional purposes has caused the development of hardware-independent software packages of parametric 3D-modeling of parts, surfaces, assemblies and drawings for machine builders. The most popular computer-aided design environment is the Autodesk Inventor package -- the Autodesk Company's product. The graphics system with OpenGL support used by the company allows you to work with three-dimensional assembly drawings that contain a lot of components. The Autodesk Inventor package has an intuitive user interface, the DSS design support system, and an excellent EESW multimedia aid system. To create 3D-models of the elements of part designs and structural elements of parts, assemblies and mechanisms, the Autodesk Inventor package uses a large set of tools. The conditions for the development of adaptive and parametric 3D- models have also been created.

The Structural Chamfer Element

In this paper we consider the issue of enhancing the tool options to create a structural chamfer element.

The word "chamfer" originates from the French word "faccete", which means the beveled portions of assemblies, edges, and so on. The material is cut by way of a plane or conical surface. The bulk of chamfers are designed for blunting sharp edges to ensure the safety of these manufacturing operations or use of products and mechanisms [Anurjev, 2006].

In the technical drawings the chamfers and their geometric parameters are indicated in those cases where it is necessary to clearly indicate their presence due to the technical solution. In other cases, chamfers are not specified, but they should be considered in the manufacturing process.

Preferably, as it was mentioned above, chamfers are designed to ensure the safety of interaction between people and products of their industrial activity, but in some cases chamfers are needed as decorative items that are used by designers as a part of the product.

There are chamfers with quite a flat bevel [Anurjev, 2006] that enable the parts to perform functions providing a guaranteed engagement with the corresponding components of assemblies and mechanisms.

If conical chamfers are cut off from faceted surfaces (squares, hexagons), then they automatically show line crossings which are conventionally depicted in drawings by arcs of circles [Anurjev, 2006], [Mikhaylenko , 2001].

Problem Statement

To create a chamfer element in the Autodesk Inventor package, three methods [Guznenkov, 2009] are used: 1. Distance; 2. Distance and angle; 3. Two distances.

When creating a chamfer on hexagonal surfaces, it is convenient to use the reference data "width across flats S" [Anurjev, 2006] as the initial parameter for the chamfer element in the Autodesk Inventor package.

After the analysis and mathematical treatment of the reference data for creating a chamfer element in the Autodesk Inventor package by the second method (distance and angle), we propose three ways of building up chamfers:

1. *0,27 S, α=30^0;* **2.** *0,475 S, α=30^0;* **3.** *0,1 S, α=30^0.*

The First Method to Build up a Chamfer Element in the Autodesk Inventor Package

First, we make a basic sketch using the tools to create a sketch profile – a regular hexagon (Fig. 1). The hexagon is circumscribed around a circle with a diameter specified by the parameter taken from the reference data [Anurjev, 2006] "width across flats S" (d0=S).

Then we use the tool "Extrude" to create the extruded element on the basis of the existing sketch profile (Fig. 2).

Next, we make an additional sketch using the tools to create a sketch profile – a triangle (Fig. 3). The additional sketch is created in the symmetry plane of the extruded element and perpendicular to the plane of the base sketch (Fig. 1). In building up the triangle it is necessary to observe the conditions listed below:

- The angle between the sides at the top of the triangle is equal to 300 [Anurjev, 2006];
- The vertex of the triangle at an angle of 300 is to be on the axis passing through the center of the inscribed circle d0 = S (Fig. 1, 3);
- One side of the triangle is to be perpendicular to the axis passing through the center of the inscribed circle d0 = S (Fig. 3);
- The two sides of the triangle at an angle of 300 are to be at least half the diameter of the circum circle of the extruded element (Fig. 3).

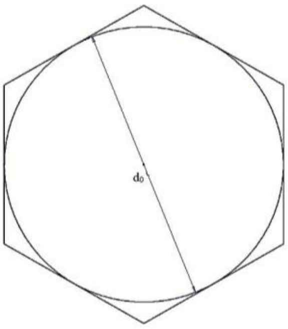


Figure 1. Creating a basic sketch

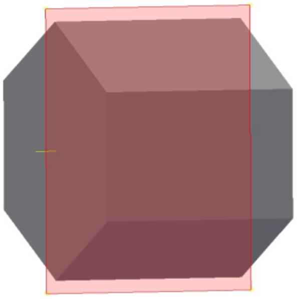
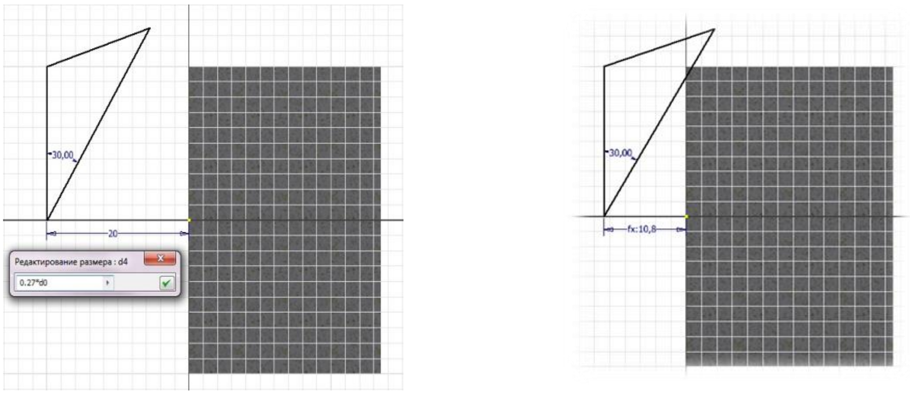


Figure 2. Creating the extruded element

The next step is to place the triangle at a distance from the top (bottom) base of the extruded element at a distance of 0,27 d0 (Fig. 3)



a b

Figure 3. Creating a sketch profile – triangle

For the termination of chamfer building we use the tool "Revolve" to create the element of rotation based on the existing sketch profile (Fig. 3b). The profile is rotated at an angle of 3600. The tool "Revolve" is used with the option "Delete" (Fig. 4).

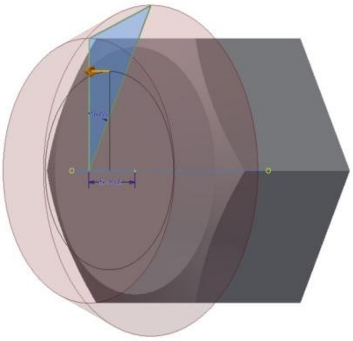
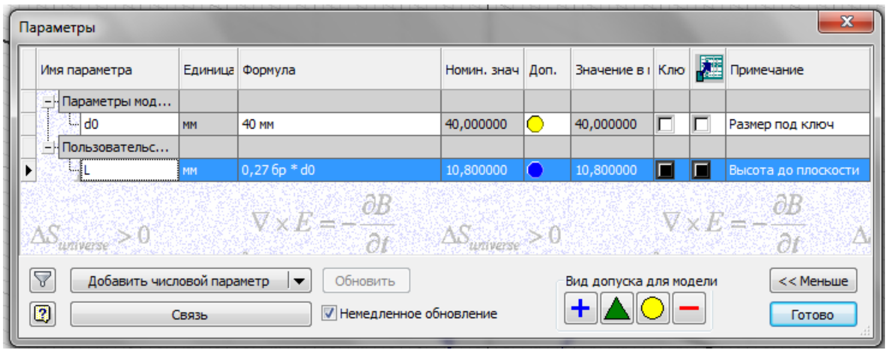


Figure 4. Termination of creating a chamfer element

For creating a sketch, the Autodesk Inventor package provides parameterization, i.e. the sketch geometry with superimposed constrains for managing size and possible geometry behavior. These features allow you to use parameterization for creating a chamfer element. Then, when you create a basic sketch expressing the distance *L* as the parameter "width across flats S" (Fig. 5) in a sub-sketch, with the given distance of the triangle location from the top (bottom) base of the extruded element, it is sufficient to enter the parameter *L* in the dialog box "Edit the size" (Fig. 3a).



Параметры

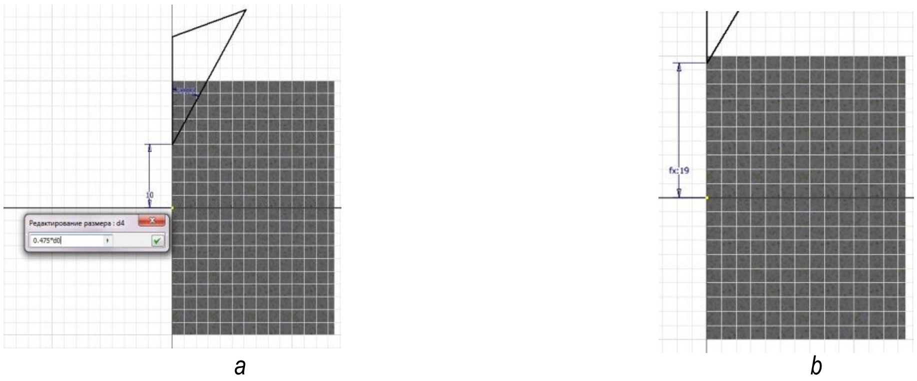
Имя параметра	Единица	Формула	Номин. знач	Доп.	Значение в	Клю		Примечание
Параметры мод…								
d0	мм	40 мм	40,000000	●	40,000000	□	□	Размер под ключ
Пользовательс…								
L	мм	0,27 бр * d0	10,800000	●	10,800000	■	■	Высота до плоскости

Figure 5. Parameterization in the basic sketch

The Second Method to Build up a Chamfer Element in the Autodesk Inventor Package

First, as it was the case with the first method of building up a chamfer, we create a basic sketch (Fig. 1) using the parameter "width across flats S" and the extruded element (Fig. 2). Then, we create an additional sketch (Fig. 6). In building up the triangle the following conditions must be observed:

- The angle between the sides at the top of the triangle must be equal to 300 [Anurjev, 2006];
- The vertex of the triangle at an angle of 300 must lie on the top (bottom) base (Fig. 6);
- One side of the triangle must be perpendicular to the axis passing through the center of the inscribed circle d0= S (Fig. 1, 6);
- The two sides of the triangle at an angle of 300 must be outside the contour line of the circumcircle of the extruded element (Fig. 6).


Редактирование размера : d4

0.475*d0

a

b

Figure 6. Creating a sketch profile – triangle

The next step is to place the triangle at a distance from the axis passing through the center of the inscribed circle d0 = S (Fig. 1, 6) on the top (bottom) base of the extruded element, at a distance of 0,475 d0 (Fig. 6).

For the termination of building up the chamfer, we use the tool "Revolve" to create the element of rotation on the basis of the existing sketch profile (Fig. 6 b). The profile is rotated at an angle of 360^0. The tool "Revolve" is used with the option "Delete" (Fig. 7).

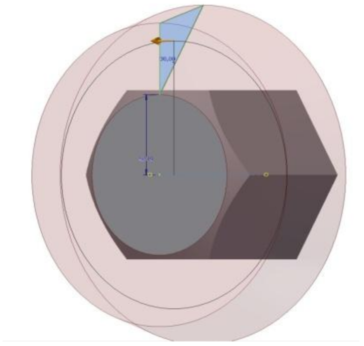
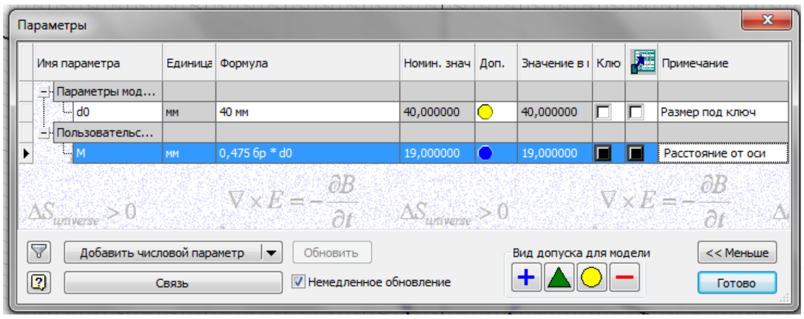


Figure 7. Termination of creating a chamfer element

The second method to build a chamfer in the Autodesk Inventor package also enables the use of parameterization for creating a chamfer element.

Then, when you create a basic sketch expressing the distance M as the parameter "width across flats S» (Fig. 8) in the sub-sketch, with the given distance of the triangle location from the axis passing through the center of the inscribed circle d0 = S, it is sufficient to enter the parameter M in the dialog box "Edit the size" (Fig. 6a).



Параметры

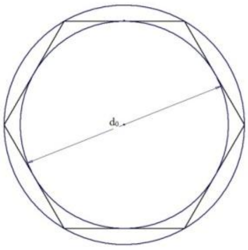
	Имя параметра	Единица	Формула	Номин. знач	Доп.	Значение в і	Клю		Примечание
	Параметры мод...								
	d0	мм	40 мм	40,000000	●	40,000000	□	□	Размер под ключ
	Пользовательс...								
▶	M	мм	0,475 бр * d0	19,000000	●	19,000000	■	■	Расстояние от оси

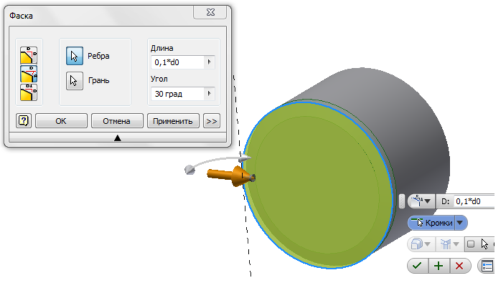
Figure 8. Parameterization in the basic sketch

The Third Method to Build up a Chamfer Element in the Autodesk Inventor Package

The third method has significant differences from the previous ones. Firstly, in the base sketch a circle is built up and described (Fig. 9), secondly, a cylindrical surface is extruded (Fig. 10), and thirdly, there is no need for an additional sketch.

To build up a chamfer by using standard built-in tools provided by the Autodesk Inventor package, we select the tool "Chamfer", the method "distance and angle." In the field "Length" we indicate 0,1 d0,, and in the field "Angle" we indicate the angle of 300 (Fig. 10).





Фаска

Ребра
Грань
Длина: 0,1*d0
Угол: 30 град

Figure 9. Creating a basic sketch

Figure 10. Creating the extruded element

To finalize we use the tool "Extrude" with the option "Delete" (Fig. 11).

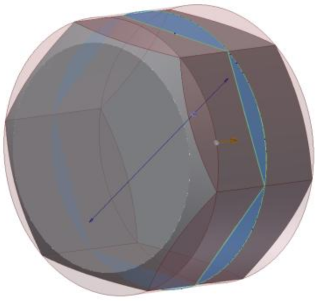
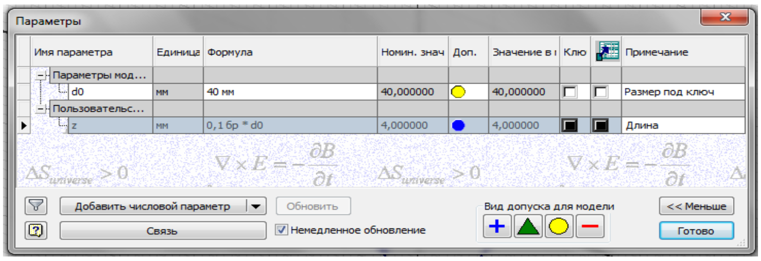


Figure 11. Termination of creating a chamfer element

The third method to build up a chamfer in the Autodesk Inventor package also enables the use of parameterization for creating a chamfer element.

In this case, in creating a basic sketch "Length», z acts as the parameter "width across flats S" (Fig. 12).



Имя параметра	Единица	Формула	Номин. знач	Доп.	Значение в	Клю		Примечание
Параметры мод...								
d0	мм	40 мм	40,000000	○	40,000000	□	□	Размер под ключ
Пользовательс...								
z	мм	0,1 бр * d0	4,000000	●	4,000000	■	■	Длина

Figure 12. Parameterization in the basic sketch

Next, when we build up a chamfer using standard built-in tools of the Autodesk Inventor package by means of the tool "Chamfer", the method "distance and angle", we indicate the parameter z in the field "Length" (Fig. 10).

Conclusion

The analysis of literary sources has shown that in some cases, when you build up a chamfer element in the Autodesk Inventor package, it is convenient to use the parameter "width across flats S".

None of the options for creating a chamfer element in the Autodesk Inventor package are universal, and they are limited in their application.

The three methods proposed for building up a chamfer element in the Autodesk Inventor package are simple and accessible in creating 3D-models of parts, assemblies and mechanisms;

A method to create a chamfer element in the Autodesk Inventor package is chosen based on the geometric characteristics of parts.

The third method of creating a chamfer element in the Autodesk Inventor package with easy parameterization (z = 0,1 S) and a smaller number of additional constructions is advantageous among all the considered methods.

References

[Anurjev, 2006] V.I. Anurjev. Reference Design-mechanic. In: Engineering, 2006.

[Banach, 2006] Daniel T. Banach, Travis Jones, Alan J. Kalameja. Autodesk Inventor Essentials, New York, 2006.

[Guznenkov,2009] V.N. Guznenkov. Autodesk Inventor aware of engineering graphics. A manual for schools, 2009.

[Koncevich, 2007] V.G. Koncevich. Solid modeling of engineering products in the Autodesk Inventor, 2007.

[Mikhaylenko, 2001] V. Mikhaylenko, V. Naidu, A. Pidkorytov, I. Skidan. Engineering and Computer Graphics: Tutorial, 2001.

Authors' Information

Evgeniy Ivanov - *candidate of technical sciences, associate professor of engineering and computer graphics, Kharkov National automobile-road university, Petrovskogo str., 25, 61002 Kharkov, Ukraine; e-mail: repositiv@mail.ru*

Zoya Ivanova - *candidate of technical sciences, associate professor, senior researcher of the department of reliability and vibration testing of high-speed machines, Institute of problems of mechanical engineering of them A.N. Podgorny the National academy of sciences of Ukraine, Dm. Pozharsky str., 2/10, 61046 Kharkov, Ukraine; e-mail: bozhko@ipmach.kharkov.ua*

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*

Kostyantyn Lagosha - student, Kharkov National automobile-road university, Petrovskogo str., 25, 61002 Kharkov, Ukraine; e-mail: sheriff_187@mail.ru

DECISION SUPPORT SYSTEM IN RAPID PROTOTYPING TECHNOLOGY

Arkadiusz Rzucidło, Grzegorz Budzik, Łukasz Przeszłowski

***Abstract**: Article describes building a base system as a supporting tool in rapid prototyping (RP) technology. The main task of this system is a supporting in decision making of choosing right technology for build a prototype. The base system means a initial structure of system which contains a particular technical values for each RP process and used materials. The operator can use them for initial selection of technology which is the best choice for projected and generated CAM model. It is a base for further developing in this topic.*

***Keywords**: Decision supporting, Rapid prototyping.*

***ACM Classification Keywords**: A.0 General Literature - Conference proceedings, I. Computing Methodologies, I.2.1 Applications and Expert Systems, J. Computer Applications.*

Introduction

Rapid Prototyping technology is a part of the computer-aided design (CAD). It is usually associated with modelling of machine parts. It is commonly accepted to call this type of production – 3D printing. This name is associated with the method used in this process of the incremental construction of the designed detail.

Undoubtedly, it is one of the fastest methods of creating machine parts. Manufactured parts can be used in different ways. Models can have a demonstrative character and have the presentation role. They can also be fully functional elements of short batches of products used in greater mechanisms. Such manufacturing is expensive, hence this method is not used in the mass production.

The way of manufacturing the detail in a 3D printer is only based on the data included in the CAD file. The direct data file for the printer is the export called STL in the file. There are also used other formats of exports. Designed models are made with specific techniques and using materials described for the given method. Therefore, they have predetermined properties imposed by the manufacturing technology. The main advantage of the production of models using the RP technique is the possibility to omit the tedious stage of preparation of the detail production obtaining it directly after the design process. This allows the client to present the part, which he is interested in. The fact, which is also important, is that thanks to the development of the prototype it is possible to notice the construction errors at the machine design stage. In each of these cases we gain both time and money. Also the scientific aspect is also important involving the possibilities of the analysis of the implemented machine parts in terms of structure and durability. Studies on prototypes allow to determine the features of final items and technology, which will be optimal for the production of parts in the high volume production

Rapid Prototyping Technology (RP)

A set of methods for producing prototypes of machine parts connected with RP is very rich. However, it is possible to determine the classical division of technology taking into account both the material, from which the prototype is made, and the way of its use and binding. In [Budzik 2011] particular methods have been thoroughly characterised, from which the basis group is:

- **Stereolithography (SLA) -** that is hardening with laser the successive layers of liquid resin in accordance with the prototype model. Hardening takes place only in the particular place, hence high accuracy of this method.

- **Selective Laser Sintering (SLS) -** In this technique, successive layers of material are applied by the machine and then the laser hardens the selected points. Unhardened powder is then removed and a finished item is received. The process must take place in the vacuum.
- **Electron Beam Melting (EBM) -** The model production method using the electron beam melting metal powders. Each of the model layers is made by melting the next cross-section of the model according to data from the control file.
- **Laminated Object Manufacturing LOM -** is the creation of the model of thin paper layers, most often adhesive, forming layers of another plane creating the given prototype. Thus created forms are easily workable and machined and can serve as project verification – they do not have limitation regarding the complexity of the detail. They can serve as models for making forms.
- **Fused Deposition Modelling FDM -** Fused deposition modelling. It involves the placement of the model material by the head (e.g. ABS) according to successive horizontal sections based on the 3D model. The ready made design is practically ready for use after its creation.
- **Digital Light Processing 3D printing (DLP) -** It hardens the liquid polymer using the laser beam.
- **Three Dimensional Printing (3DP) -** This method uses powders: ceramic, polymer or metal. Powders are bonded with an adhesive, which connects them together layer by layer of the model. After creation, the model is subjected to annealing and next layers are connected together by forming, and as a result the adhesive disappears.
- **PolyJet -** The object is constructed on a special tray over which the moving heads (like in an inkjet printer) place the photopolymer. After each application, the UV rays harden the layer. Thus, there is no need to use the additional finishing of the model.

RP Decision Support System

The task of the decision support system in Rapid Prototyping is gathering the expert knowledge and using it in the selection of optimal technology to create the designed part based on the multi-criteria query of the operator. The complexity of the query depends on the references regarding the prototype. It concerns such basic parameters like: dimensions of the manufactured element, which determine the machine on which the given detail can be made as well as the material, from which it will be created. Classification also concerns more advanced requirements like the purpose of the prototype indicating the precision of mapping the details of the designed part or a complex as well as the features, like the physical-chemical properties extremely important in prototype research.

The basic classification concerning dimensions of the constructed objects and declarations of the material for the construction is made based on data included in technical descriptions. Technical specifications of machines introduced to the structure of the service database concern information both about the selection of machines compare the operator's requirements with the dimensions of the machine's working space. Each machine is at the same time associated with the group or family of materials related to the technology, in which the details are made. Hence the basic criterion is determined by dimensions of the objects indicates the range of available methods and determines the material. Designer of parts is therefore able to state at an early stage of design whether he is able to make the selected prototype according to the preferences. Technical data of materials used in RP and selected by the system indicate also the potential purpose of the created prototype as, e.g., the work part of the research subject [Sobolak Budzik 2008, Oleksy at al. 2010, Grzelka at al. 2012], or the review mock for the presentation [Budzik, 2011, Budz at al. 2013].

Precise classification, concerning the expert knowledge in the field of selection of RP techniques requires the possession of extensive information reaching beyond the framework of catalogue technical data. Response of the service to the detail query of the type: “with what method and what kind of material can we make the object characterised by the selected hardness, with holes with the selected diameter and carrying the determined loads” requires resorting to additional data obtained from research. Preliminary tests for particular RP technologies connected with the description of accuracy of the description, load characteristics and suitability for the manufactured prototypes are a must. Writing down the achieved results as information in the service in a way which enables the reference to them in the query requires the development of database with proper structure. This design is to provide not only the present use but is to become the good base for the further development of the system. The extended part of the service however is the future. The current state of the system of the RP support covers the basic catalogue data and on their basis the selections of machine lists.

RP Information Store

Database according to the rules, which determine the relational structures of data stores is divided into several basic sets. This structure is created by tables defining the information of the elementary nature. The system of tables is presented n fig.1.

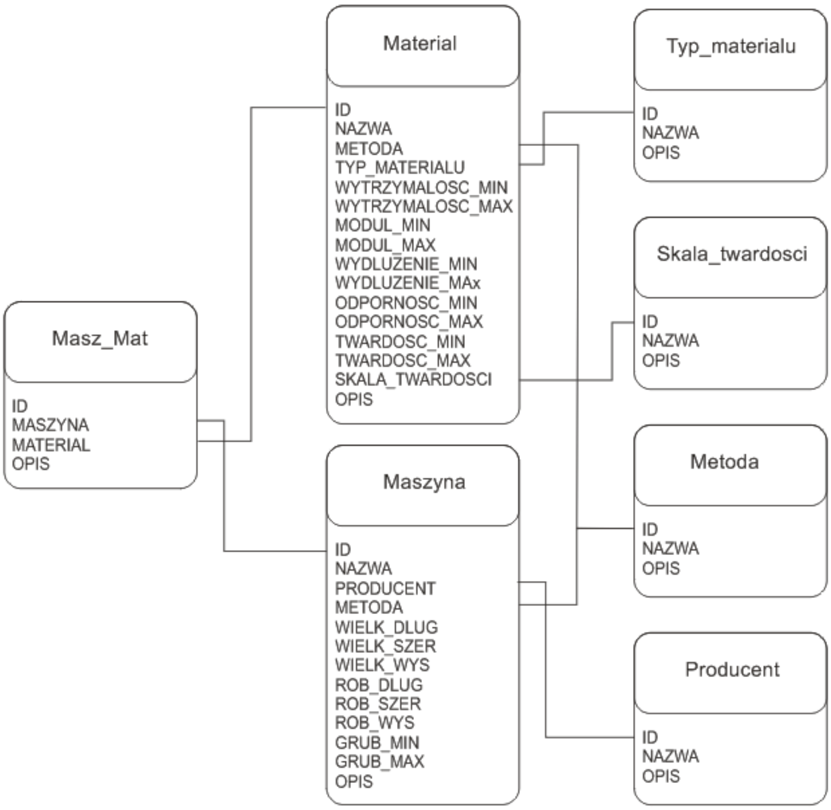


Figure 1. The system of tables in the RP support system

Basic dictionary data are included in the tables Material_type, Harndness_scale, Method and Manufacturer. They contain information which are the complement of the fields of the main selection forms for the service. Main forms supplement the Material and Machine tables with necessary data. They are main containers of data with the characteristic of the material and the machine. Data in both tables are described by the method in order to clarify the information included in them.

The transaction table and at the same time the element binding all collected data is the Masz_Mat. It includes the compositions of the Machine and Material on the basis of sets. Data included in it are used by the form of the main query in the service.

Characteristics of data in the table Material in addition to classification for the method of a particular material also concerns the numerical limit values within the scope of:

- material strength in [MPa],
- Young elastic module [MPa],
- elongation [%],
- resistance to impact [J/m2],
- hardness – as an extension there was provided the hardness scale.

Characteristics of Machines described in the service concerns, as in the case of materials, the determination of the RP method and manufacturer, as well as the numerical limit values in the scope of:

- Dimensions of the machine: length, width and height.
- Dimensions of the workplace: length, width and height.
- Limit values connected with the accuracy of the detail performance.

It should be noted that for particular machines the accuracy of the prototype performance depends on the adopted material and the next layer of reproduction, which is connected with the RP technology. In other words, it can be assumed that the used material determines the accuracy of the performance and the machine placing next layers builds the appropriately detailed prototype. Of course, at this stage of design, in CAD programs there are defined the details of the prototype and its accurate dimensions, however the finally adopted RP method allows to map the construction of the detail in the appropriate way and possible to achieve in reality (characteristic for the technology of prototype creation). This is of particular importance in case of details with a complex structure, and especially in case of such, in which there can be observed holes as elements of cooperation with other details, e.g., sliding bearing. In case of making prototypes with the purpose of review models it does not matter, however the cooperation of imprecisely made details manifests itself in the form of, e.g., vibrations. In case of prototypes which are the useful parts of machines, the quality of performance is definitely important and influences the resulting nature of the detail classifying it directly for use or the finishing. This gives it the final dimensions and the form compatible with the CAD design.

The Query of the RP Support Service

The possibility to use the collected information in the service database is the basic task of the RP support system. It essentially contains the core idea of the system. Complete data placed in tables and organised in the above-mentioned way allow to filter data placed in the form of the basic query. Here we use the classic approach of selection using the conditional instructions if…else…

The query determines such basic data like dimensions of the designed detail, which initially estimate the list of machines, which meet the requirements of the designer. At this stage it is possible to decide also what kind of material will be used for “printing” of the object. This is the basic stage, in which the query defines the criteria regarding the Machine table. In case of the determination of the Material field, the query is expanded also with the Material table. Fields of the form of the basic query within dimensions are required.

The next stage for the system is the precision of properties of requirements concerning the physical properties of the detail. This concerns the arrangement of details regarding the resistance of the parts to loads, its elastic module, elongation, resistance to impacts and hardness. These are the features of the material of which the prototype will be built and they correspond to entries to the Material table.

In case of the absence of numerical values in the form fields of the second part – “advanced” – no criteria regarding the material requirements are determined.

Supplementing the above-described steps results in the list of machines, on which there can be performed the designed part or detail. The list of proposals contains links to the models selected by the system, which are the photographic examples of what the parts manufactured in the selected technology look like.

The Evolution of the System

The above-described scenarios concerns the basic form of decision support in the RP process.

The current activity of the system consists of the catalogue data, which are given by the material and machine manufacturers. These are real data, but their theoretical nature allows only the simple qualification of machines and materials in the design.

The evolution of the system assumes the empirical approach to the mentioned classification and basing the conclusions about the selection of RP technologies on the laboratory data. This type of task requires the use of a fairly complex approach to the problem because the characteristics of details made in RP is extremely difficult. The spectrum of parts obtained thanks to 3D printing is basically unlimited, however, the requirements posed to them are connected with specific needs. These depend on properties of the material selected for the product.

Conclusion

The currently performed system is the output element for further works within the RP decision support. It serves as the dictionary base for embedding empirical data, which in the further stage of the analysis of the RP process will be acquired through experiences. The system will be developed so that in the final form it will allow the most precise mapping of the needs of the operator designing the product into a prototype made with the right technology.

Bibliography

[Budzik 2011] Budzik G.: *The Use of the Rapid Protyping Method for the Manufacture and Examination of Gear Wheels*. [w:] Advanced Applications of Rapid Prototyping Technology in modern Engineering., (pod red.) Muhammad Enamul Hoque s.339-364, Croatia, 2011 INTECH - Open Acces Publisher, Croatia 2011, s. 339-365.

[Sobolak Budzik 2008] Sobolak M., Budzik G.: *Experimental method of tooth contact analysis* (*TCA*) *with Rapid Prototyping* (*RP*) *use*, Rapid Prototyping Journal, 14, 4, 2008, s. 197-200.

[Oleksy at al. 2010] Oleksy M., Budzik G., Heneczkowski M., Markowski T.: *Kompozyty żywic poliuretanowych z dodatkiem Nanobentów*, POLIMERY 2010, 55, 3, s. 194-200.

[Grzelka at al. 2012] Grzelka M., Marciniak L., Gapiński B., Budzik G., Trafarski A., Augustyn-Pieniążek J., Gaca M.: *Accuracy of the Element Geometry Mapping Using Non-Invasive Computer Tomograpy Method*, Journal of Automation, Mobile Robotics & Intelligent Systems –JAMRIS, Vol. 6, No 3/2012, s. 23-26.

[Budzik at al. 2013] Budzik G., Pacana J., Kozik B.: *Defining influence of load conditions on distribution and value of stresses dual-power path gear wheels applying MES*, Aircraft Engineering and AerospaceTechnology Vol. 85 No 6 2013, s. 453-459.

Authors' Information

Arkadiusz Rzucidło – *Rzeszow University of Technology, Ph.D. in Departament of Computer Science. Al. Powstancow Warszawy 8, 35-959 Rzeszow, Poland. Email: arzucidl@prz.edu.pl*

Major Fields of Scientific Research: Internet Technology, Data visualizations, Decision supporting in production, Software technologies, Information systems.

Grzegorz Budzik - *Rzeszow University of Technology, Professor in Departament of Machnine Construction. Al. Powstancow Warszawy 8, 35-959 Rzeszow, Poland. Email: gbudzik@prz.edu.pl*

Major Fields of Scientific Research: Rapid Prototyping Technology, CAD-CAM Systems

Łukasz Przeszłowski - *Rzeszow University of Technology, Msc. in Departament of Machnine Construction. Al. Powstancow Warszawy 8, 35-959 Rzeszow, Poland. Email: lprzeszl@prz.edu.pl*

Major Fields of Scientific Research: Rapid Prototyping Technology, CAD-CAM Systems, CAE Systems,

THE PROBLEM OF CORRECT TECHNOLOGY SELECTION IN RAPID PROTOTYPING

Arkadiusz Rzucidło, Grzegorz Budzik, Łukasz Przeszłowski

Abstract: *Rapid Prototyping (RP) Technology gives a possibility fast projection of data saved in files by CAD programs on real object. Rapid Prototyping Technology gives a possibility fast projection of data saved in files by CAD programs on real object. RP allows on "3D printing" for objects even with complicated structure. This is not a machining. Prototypes are built layer by layer. Producing is faster than in machining. Numbers of available technologies RP and used materials is large. Technologies have own preferred materials. The materials have different physical and chemical properties. Choosing of right technology for particular object is not easy. We should take into consideration many features, which affect on produced object. In process of choosing we can use collected properties from catalogues related to the algorithm in software of decision support. This article presents a few RP technologies and materials. Contains characterization of known decision support systems in RP processes. The main topic is a new concept of base system for RP. The first part is system based on catalogue data for machines and materials. It provides first selection in decision making process. Next step will be using empirical data. There is a purpose to use a neural network structures to determine right technology.*

Keywords*: Decision supporting, Rapid prototyping.*

ACM Classification Keywords*: A.0 General Literature - Conference proceedings, I. Computing Methodologies, I.2.1 Applications and Expert Systems, J. Computer Applications.*

Introduction

Computer-aided machine design (CAD) is the basic stage in the modern digital design of machines. It is applicable in industrial sectors, but not only. This technology, due to its possibilities can successfully be used also for modelling other structures apart from machine parts, e.g., in medicine for the reconstruction of the skeletal system. It involves the mapping of the physical geometrical three-dimensional model into the digital model. Such action can be used in order to write down the structure of detail, which can be the part of the object or the whole object. This way there is created the digital prototype of the designed object.

Thanks to Rapid Prototyping (RP) technology, there is a possibility of the quick transition of the digital CAD model to the physical object. In contrast to the machining, RP is the incremental technology. Thus, creating the physical model from digital STL data after the design CAD stage does not take place through removing the material like in the machining, but by adding material or by the phase change from the liquid into solid state. The object is built in layers. This allows to make objects with very complex structure, both internal and external.

In a typical process of the computer design support there can be specified several characteristic steps:

1. Establishment of the concept of the built object or its parts.
2. Implementation of the model of the main element in the digital form.
3. Implementation of the digital mock of the object with the determination of the technical and technological specification for the object.
4. Performing durability calculations of the elements of the object, checking the correctness of operation (analysis of collision), selection of material for the object.
5. Making the prototype (among others Rapid Prototyping Technology)
6. Implementation of the necessary documentation.

Rapid Prototyping technology combines the digital stage, conceptual with the Real physical model of the constructed object. It allows the quick mapping of the design as the prototype, which in case of some RP methods may have physical-chemical parameters close or even identical to the design. Currently, the spectrum of available RP methods is quite large, and therefore, there appears the need of decision, which of the methods is the right one in case of the particular design. The topic of decision support in the selection of the optimal RP technology for prototypes is not new. Many academic centres dealt with this topic. So far the complex system has not been developed, which would support this type of actions. Decisions about the selection of technology apart from support using computer programs are still performed based on experience of designers specialising in the RP technology.

The accurate selection of the RP technology is important because it favours the growth of the product quality [Sobolak Budzik 2008]. Already at an early stage of designing we can eliminate all errors accompanying the construction or modify the object before it goes to the mass production. In some cases there can be made the missing element of the object, which cannot be repeated. The example is the area of medicine, where the modelled part of bones, e.g., skull can be made in the RP technology in almost 100% compliance with the need without performing the finishing. The prototypes made by Rapid Prototyping technology are used in many cases. The objects or parts are used in many measurements in researches. There are a researches of cooperation between parts of machine [Sobolak Budzik 2008, Budzik at al. 2013] especially in machine gear boxes, accuracy of geometry a machine parts [Grzelka at al. 2012], visualizations of mechanism and the others.

Rapid Prototyping Technology

The range of creating real models as prototypes of machine parts in RP technology is very rich. We can distinguish the classical division of technologies taking into account both the material, of which the prototype is made, and the way of its use.

In Fig.1. we can see the diagram describing the current methods of rapid prototyping with regard to materials for the produced details.

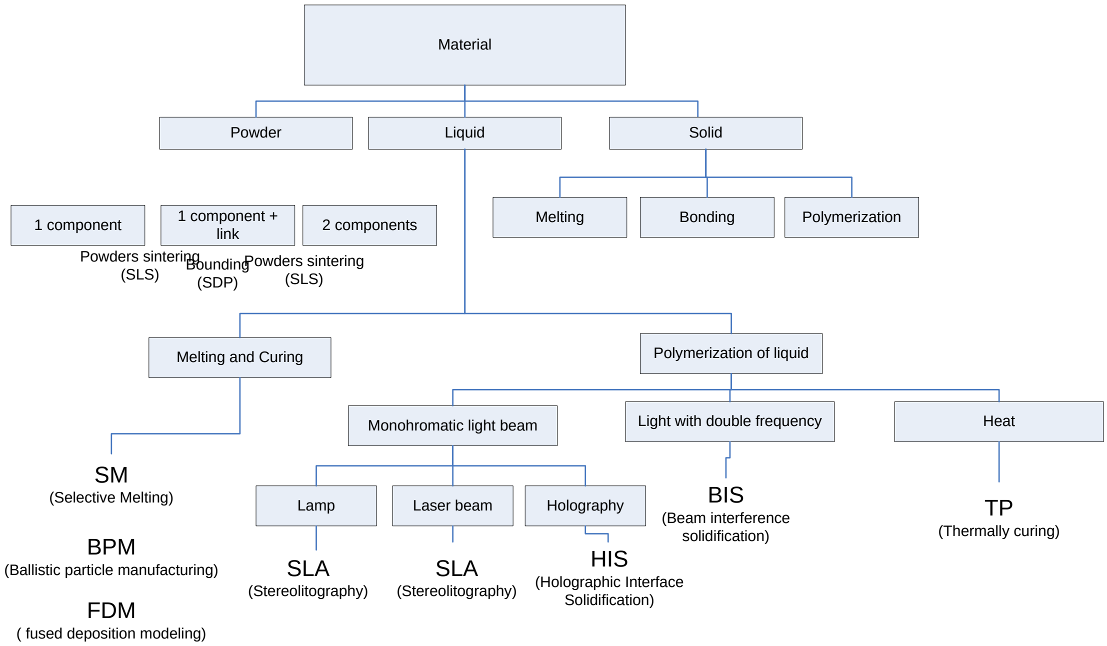


Figure 1. Diagram of methods in Rapid Prototyping technology

[Budzik 2011] characterises precisely the particular methods, of which the basic group considered in the content is:

- **Stereolithography (SLA) -** That is hardening next layers of resin by the laser in the liquid form according to the prototype model. Hardening takes place only in the particular place, hence the high accuracy of this method .A example of SLA object is set on fig.2.

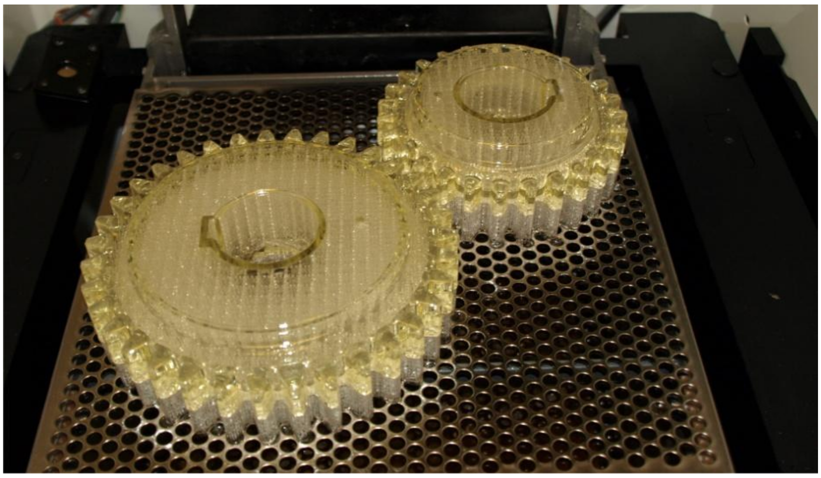


Figure 2. Gear wheels made by SLA method

Selective Laser Sintering (SLS) - In this technique next layers of material are placed by the machine, and then the laser hardens the selected points. Unhardened powder is then removed and we obtain the finished item. The process must take place in the vacuum. Example on fig.3.

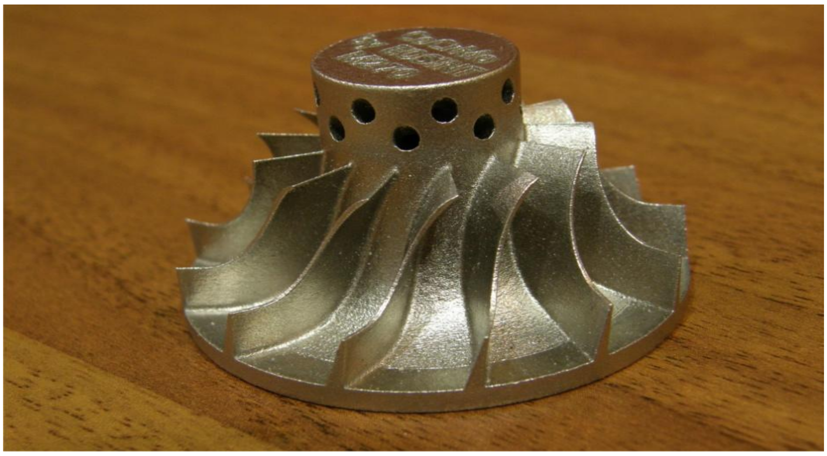


Figure. 3. Rotor made by SLS method

Electron Beam Melting (EBM) - Method of making the model using the electron beam melting metal powders. Each layer of the model is made by melting the next section of the model according to data from the control file.

Laminated Object Manufacturing LOM - Is the creation of the model of thin layers of paper, most often adhesive, creating layers of another section consisting the given prototype. Such made parts are easily workable and may serve as the design verification – they have no limitation in terms of complexity of detail. They may serve as models for making forms.

Fused Deposition Modelling FDM - Fused deposition modelling. It involves placing the model material by the head (e.g. ABS) according to another levels of sections based on the 3D model. The ready project is practically ready for use after its creation (fig. 4).

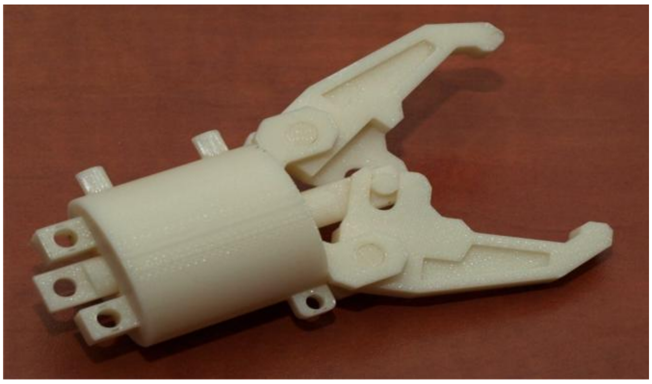


Figure 4. Gripper made by FDM methods

Digital Light Processing 3D printing (DLP) - The liquid polymer is hardened with the laser beam.

Three Dimensional Printing (3DP) - This method uses powders: ceramic, polymer or metal. Powders are connected with an adhesive, which links them together layer by layer of the model. After creation the model is heated and next layers are connected together linking, and as a result the adhesive disappears (fig. 5).

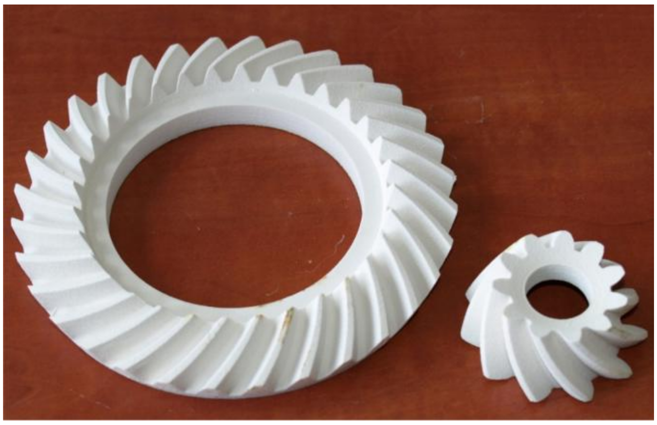


Figure 5. Models produced by 3DP method - gear wheels

Digital Light Processing 3D printing (DLP) - The liquid polymer is hardened with the laser beam.

Three Dimensional Printing (3DP) - This method uses powders: ceramic, polymer or metal. Powders are connected with an adhesive, which links them together layer by layer of the model. After creation the model is heated and next layers are connected together linking, and as a result the adhesive disappears (example on fig. 6).

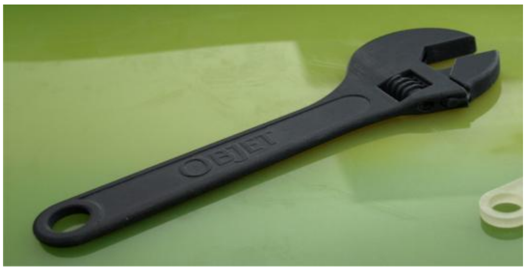


Figure 6. Element produced by PolyJet method

The above classification is the presentation only of the selected methods and examples of the created prototypes. It does not include all ways of prototyping. As it can be easily observed, already in case of this group the decision about the selection of the particular method may constitute a problem, especially in case of people not being experts in the RP technology.

Decision in the Method Selection

Decisions about the selection of the appropriate technology in creation of a prototype in the context of a number of presented methods are a difficult stage. It requires insight into the general range of technologies from the designer. This is not always possible, even due to the number of their number and diversity of materials used for the selection of prototypes. Each material has different physical-chemical properties, which determine the performance of the designed part in some group, e.g., endurance. Selection criteria of the optimal material, and at the same time technology, can be based on determinants of the material properties, machines as well as tests conducted on the created details. The scope and accuracy of the selection and of the undertaken decision concerning technology will depend on the area of knowledge that will be built for the algorithm supporting the choice.

The task of the decision support system would be gathering the expert knowledge for the further use on the path of the optimal RP technology for the constructed detail. Knowledge gathered by the support system would have to be so extensive so that it could in a simple way answer the simple, in terms of language, and at the same time extremely difficult, in technical terms, question: which method is suitable in our case?

Multi-criteria query for the system in assumption should be based on the basic “catalogue” data of both the machine and material, as well as some component, which is obtained on the path of studies or observations, and constitutes the expert element. To include it is the system database there should be prepared the appropriate structure of databases and find the right algorithm for inference. The algorithm based on the components determined by the designer (properties required for the detail) would indicate then the range of the most appropriate solutions in the given case.

Over the last decade many people were involved in the RP decision support using various techniques. Many of them were the composition of attempts to use IT and mathematical solutions for the most precise determination of criteria of the RP method among the selected parameters. So far, there has only been observed that the problem of decisions in this regard is so complex and complicated that it cannot be solved in a simple way, which would give 100% certainty of the right choice of technology. Among the solutions there can be noticed the simple implementation of databases and the use of more sublime methods, like the fuzzy logic. Supporting decisions in the Rapid Prototyping technology has so far been implemented several times. There have been developed a few systems supporting the selection of the right printing method. A few of the implemented solutions were described below [Byu Lee 2005].

Phillipson developed the RP Advisor. The program, which task was the selection among six available RP technologies of the right one based on time parameters of the performance of detail, cost of making and accuracy. The system worked in the MS Access database and used theories of multi-criteria optimisation, however it did not take into account other criteria, like material properties.

Bibb has developed another system. Its task was the decision support, consulting for small businesses, which wanted to use the RP technology in developing their products. In this system there were used two rules: decision-making and calculation. Decision-making rules were used for the selection of the right method based on data about quality of performance included in the STL file describing the object.

In the Institute of Industrial Technology and Applied Work Science (BIBA) there has been developed a method supporting RP decisions based on the selection of technologies from the relational database. MS Access was used for the construction of the mechanism. The system supported the decision on the basis of selection of fundamental values in catalogues and limitations of the list of the right ones for the design assumptions of machines and materials. The tool contained the combination of machines and materials and required the clarification of needs.

In the Industrial Research Institute Swinburn (IRIS) Masoon and Soo created the expert system for the selection of RP methods. The goal of the tool was the aid of the not yet implemented users, in the adjustment of the right RP methods for the performed designs and purchase of the right machine. The program included many available technologies. Compared to the previously described system it had the possibility to determine criteria regarding the price of machines, accuracy of the performance of the details, quality of the output surfaces, implementation time of the prototype, type of material and printing speed. The program allowed the user to select one out of four options: a rapid selection of parameters, careful selection of parameters, RP technology and machine type to determine the satisfactory 3D printing technique. The system was not designed to select the best technology for printing the proper part, and it was more directed towards the selection of the appropriate one for the selected parameters by the user to purchase the optimal machine.

The system described by H.S. Byun and K.H. Lee. It is the solutions of RP decision support with the multi-criteria approach. The TOPSIS algorithm is used here (Technique for Order of Preference by Similarity to Ideal Solution). TOPSIS is based on the concept of alternative, which should have the shortest geometrical distance from the perfect solution and the longest geometrical distance from the negative solution. In this case the perfect solution is the creation of the model with the assumed properties.

Table 1. Summary of RP support systems. Based on [Byu Lee 2005]

Name	**Criteria for data selection**	**Used Method**
RP Advisor (RPA)	Cost, Time, Quality	Multi-criteria optimization (MS Access)
Rapid Prototyping System Selector (RPSS)	Accuracy, Wall thickness, Material properties	Rule based (MS Access)
Muller system	Material Properties	Benefit value analysis
Intelligent RP System Selector (IRIS)	Price of the RP system Accuracy,	Rule based, Surface finish, type of material, Range of layer thickness, Building envelope, Building Speed
Multiple-attribute decision making (MADM)	Dimensional accuracy, Surface roughness, Part cost, Build time, Material properties	Fuzzy multiple-attribute TOPSIS Method

As it can be seen in table 1., there are many methods supporting RP decisions. In a more or less accurate way they present the range of available technologies limiting it to several (or one of) techniques, which meet the design requirements.

In the further part of the paper there is proposed the look at the topic of RP support from a slightly different perspective. The common fragment in relation to the mentioned techniques is the gathering of the appropriate knowledge used for the new concept. This will be the “catalogue” knowledge, widely available, recognised in the framework of database created for the needs of the idea. The system proposed as the concept is to be enriched with a research component, which will extend the bases of the system’s knowledge with the empirical element. Catalogue and experience data will allow the comprehensive recognition of the problem of the selection of the appropriate RP technique in the context of the requirements for the designed part.

RP Decision System in the Scope of Catalogue Data

For the purposes of works at the construction of the RP support system using Artificial Intelligence techniques, there was developed the base system containing catalogue data, which describe the fundamental available RP technologies. The system uses the cooperation with database with the mechanism of queries using the conditional statements “if … else”. It can compile a list of available RP machines basing on the selection criteria determined by the operator. The construction contains machine parameters made available by manufacturers, like e.g., machine dimensions, of workplaces. For expanding the criterion there was also determined not only the mere material for the constructed detail, but also its physical and chemical properties. Because of that, the system while determining the border values like, e.g., impact strength, elongation hardness selects the machines also in terms of requirements concerning the mere designed part. As an additional option there was also used the classification category of parts, which determines the usefulness of the printed details in the specified class of solutions, e.g.: presentation, work demonstration, machine part, etc.

This system is the foundation for the further research part, which task is to add the empirical value. This will supplement the system with compete information about the usefulness of the technique for the creation of the part of a characteristic construction, structure and mechanical properties. Due to the adopted concept of using the technique of neural networks, there is required the gathering of a large amount of data for teaching neuron structures as well as the verification of the correctness of operation (inference).

Neural Networks in the Application of RP Support

The nature of neural networks as the part of the main algorithm of the RP support system may turn out to be the best solution. However, for their use to bring good results, it is necessary to have the sufficiently large set of data for teaching the modelled constructions as well as to verify the adopted settings. It is also important to determine the structure of the set of input data despite that the constructed neuron structure is assumed with the universality feature in the context of RP methods. As the elements and details created in RP are characterised by the high complexity, describing the common features will be a challenge. Therefore, it is planned to perform, at first, of the less complex details, designed so that there are used the single, characteristic for solids and spatial structures, features. This concerns such details, like: holes, passages between planes (rounded, sharp), detail edges, fragments with small spatial sections, etc. Each of these technological detail is described with other parameters. Incorporating them into a single shared data vector is accepted for the network is already difficult at this stage. In case of the constructed system, however, this is necessary.

Planning the structure of the input vector will be realized as the direct stage with the performance of the dimensions of accuracy for the making of the selected technological details printed with the RP method. Measurement will concern details of the same block, made in different technologies and materials appropriate for the technology.

In order to determined the usefulness of neural networks for the support of the optimal selection of the RP method for the selected design, it is necessary to check several neuron structures. It is planned to initially subject, among others: multi-layer percepton, Kohonen networks, RBF networks to tests.

Conclusion

The presented concept of the decision support system concerning the Rapid Prototyping technology in the final form assumes the two-stage approach to the decision-making. The first stage is the initial selection of technologies based on catalogue data, which determine basic parameters, both of:

- the machine (e.g. dimensions of the workplace),
- category of using the detail (e.g. presentation of the mechanism, working part),
- material features (physical-chemical properties connected with the material selected for the technology) and,
- relative accuracy, which is imposed by the mechanism of the detail implementation.

For some details, this stage of support can be sufficient to assess and select the right RP technology. In more precise cases, details characterised by greater requirements, the decision-making process will require the use of a further stage, that is the algorithm included in the neuron structure built for the needs of the system. Neuron network taught the correct selection based on the selected parameters by the RP operator will select the optimal solution for the technologically advanced detail. The effect of the system's operation will be the specific response to the initially posed question: "Which method is appropriate in this case?"

Bibliography

[Budzik 2011] Budzik G.: *The Use of the Rapid Protyping Method for the Manufacture and Examination of Gear Wheels*. [w:] Advanced Applications of Rapid Prototyping Technology in modern Engineering., (pod red.) Muhammad Enamul Hoque s.339-364, Croatia, 2011 INTECH - Open Acces Publisher, Croatia 2011, s. 339-365.

[Sobolak Budzik 2008] Sobolak M., Budzik G.: *Experimental method of tooth contact analysis* (*TCA*) *with Rapid Prototyping* (*RP*) *use*, Rapid Prototyping Journal, 14, 4, 2008, s. 197-200.

[Grzelak at al. 2012] Grzelka M., Marciniak L., Gapiński B., Budzik G., Trafarski A., Augustyn-Pieniążek J., Gaca M.: *Accuracy of the Element Geometry Mapping Using Non-Invasive Computer Tomograpy Method*, Journal of Automation, Mobile Robotics & Intelligent Systems –JAMRIS, Vol. 6, No 3/2012, s. 23-26.

[Budzik at al. 2013] Budzik G., Pacana J., Kozik B.: *Defining influence of load conditions on distribution and value of stresses dual-power path gear wheels applying MES*, Aircraft Engineering and AerospaceTechnology Vol. 85 No 6 2013, s. 453-459.

[Byun Lee 2005] H.S. Byun, K.H. Lee :*A decision support system for the selection of a rapid prototyping process using the modified TOPSIS method*. Int J Adv Manuf Technol (2005) 26: 1338–1347

Authors' Information

Arkadiusz Rzucidło – *Rzeszow University of Technology, Ph.D. in Departament of Computer Science. Al. Powstancow Warszawy 8, 35-959 Rzeszow, Poland. Email: arzucidl@prz.edu.pl*

Major Fields of Scientific Research: Internet Technology, Data visualizations, Decision supporting in production, Software technologies, Information systems.

Grzegorz Budzik - *Rzeszow University of Technology, Professor in Departament of Machnine Construction. Al. Powstancow Warszawy 8, 35-959 Rzeszow, Poland. Email: gbudzik@prz.edu.pl*

Major Fields of Scientific Research: Rapid Prototyping Technology, CAD-CAM Systems

Łukasz Przeszłowski - *Rzeszow University of Technology, Msc. in Departament of Machnine Construction. Al. Powstancow Warszawy 8, 35-959 Rzeszow, Poland. Email: lprzeszl@prz.edu.pl*

Major Fields of Scientific Research: Rapid Prototyping Technology, CAD-CAM Systems, CAE Systems,