

Decision Support Models and Software for the Differential Immunophenotype Diagnostics of Leukosis and Lymphomas

Natalia Novoselova¹, Igor Tom², Michael Belevtsev³,

^{1,2} *United Institute of Informatics Problems*, ³ *Belarusian Research Center for Pediatric Oncology, Hematology and Immunology*

Abstract – This article describes the software and underlined decision support models for the immunophenotype diagnostics of leukosis (leukemia) and lymphomas adjusted for the marker or human leukocyte antigen (CD-antigen) coexpressions. Using the model knowledge base, the decision inference algorithm allows computing the degree of manifestation of the disease subtypes for the input immunophenotype features. Software provides the two-stage diagnostics of the leukemia subtypes and the lymphoma diagnostics using the set of the developed rules, possibility to observe the diagnostic results and corresponding reference information. The patient data are organized according to the unified registration card, which provides the possibility to work at the different diagnostic levels: diagnostics of the extended groups of leukosis, diagnostics of the leukemia subtypes, diagnostics of the adult and child lymphomas.

Keywords – Diagnostic rules, decision support, immunophenotyping, modeling.

I. INTRODUCTION

Acute leukosis and lymphomas are among the most widespread oncological diseases. In the childhood, these types of the oncological disease take the first place with a frequency of 4.3 cases for the population of 100,000 [1]. In Belarus, the survival rate for these diseases comprises more than 80%, which is comparable with the best world practice, but the situation with the adult patients is slightly worse. The general approach to improve the survival rate is the treatment personalization, based on the enhancement of the diagnostic accuracy.

The basic diagnostic method for these types of disease is the identification of the differentiation stage and maturity of the tumor cells using the immune marker analysis, based on the specified combination of the monoclonal antibodies, which recognize the CD-antigens on the cell surface. At present, more than 30 leukemia and lymphoma subtypes are described with no less than 40 markers, which in different combinations (presence/absence) are used for the diagnostics. The standard marker combination is presented only in less than a half number of cases; the rest is characterized by the presence of the coexpressions, nonrelevant to the particular disease and the absence of some core markers. The diagnostic accuracy depends on the knowledge of all the available marker combinations and the possible coexpressions, which require the higher level of training and deep background in the

domain. Therefore, to aid in the diagnostic process we have developed the decision support models and the corresponding software for the differential diagnostics of the leukemia and lymphoma subtypes adjusted for the marker coexpressions. Using the automation of the data processing, the proposed software system helps to differentiate the disease subtypes and thereafter to raise the quality and diagnostic timeliness for the adequate treatment. The first attempts in this direction were made for the childhood leukemia on the basis of the standard combinations of marker expressions with a subsequent identification of coexpressions [1]. In the proposed software system, the decision-making is implemented on the basis of the diagnostic rules, extracted from the database and presented as the weighted marker combinations and the specialized decision inference algorithm for the identification of the degree of manifestation for each disease subtype. The treatment personalization is provided by more accurate diagnostic rules, accounting for marker coexpressions. The application of the software system for the personal training will allow raising the educational level of the specialists dealing with differential diagnostics and reducing the training costs.

II. DECISION SUPPORT MODELS FOR THE DIFFERENTIAL IMMUNOPHENOTYPE DIAGNOSTICS OF LEUKOSIS AND LYMPHOMAS

The decision support models [2] consist of the following components (Fig. 1):

- Input data module;
- Knowledge base module;
- Decision-making module;
- Output data module.

Input data module provides the following functions:

- inputting and storing the values of the immunophenotype features for the diagnostic cases;
- input data processing necessary for the implementation of the differential diagnostics. The processing includes discretization of the expression values of the markers according to the specified thresholds. The markers with the values greater than the threshold are considered to be expressed and their new value is equal to 1, otherwise the markers are latent and the value is equal to 0.

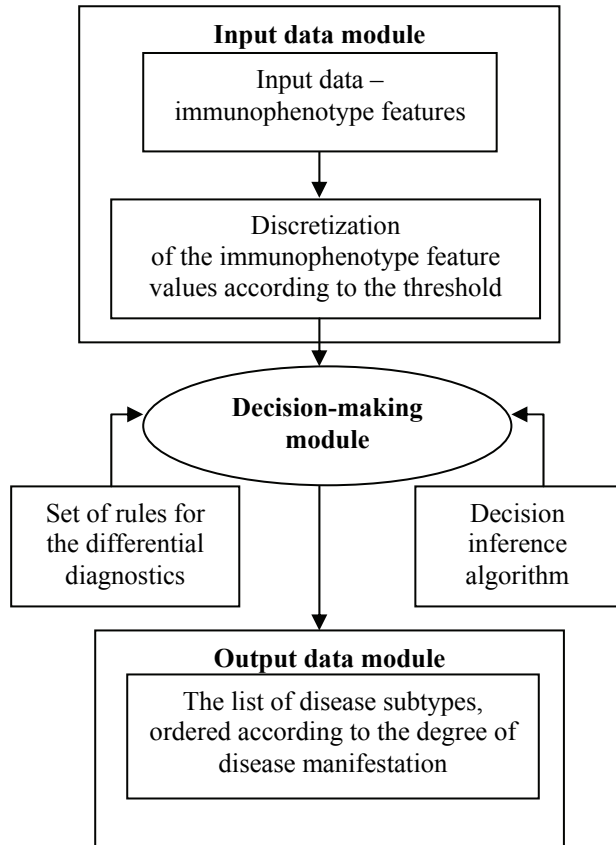


Fig. 1. The general scheme of the decision support model.

Decision-making module is the main part of the model allowing one to conduct the disease diagnostics on the basis of the available knowledge base and decision inference algorithm.

Knowledge base [3] includes the standard rules and rules developed for all the individual disease subtypes. The developed rules are based on the statistical analysis of the immunophenotype features using the cross-tabulation method.

Decision inference algorithm allows computing the degree of manifestation of the disease subtypes for the input diagnostic case using the model knowledge base.

Output data module provides the following functions:

- output data conversion necessary for the formation of the list of disease subtype ordered according to the degree of manifestation.
- output of the list of the disease subtypes with corresponding values of their manifestation.

The models for the diagnostics of leukemia and lymphomas have the identical structure with the differing knowledge base content and different organization of the decision-making module.

The model output list allows estimating the degree of manifestation for each individual disease subtype, where the subtype at the top of the list presents the most probable diagnosis.

III. KNOWLEDGE BASE AND DECISION-MAKING MODULE FOR THE LEUKOSIS DIAGNOSTICS

The general scheme of the decision-making module is presented in Fig. 2. According to Fig. 2, the two-stage diagnostics is proposed, where at the first stage the biphenotype (BAL) cases are revealed and the most probable extended leukemia group is determined (B-cell Acute Lymphoblastic Leukemia (B-ALL), T-cell Acute Lymphoblastic Leukemia (T-ALL), Acute Myeloid Leukemia (AML)).

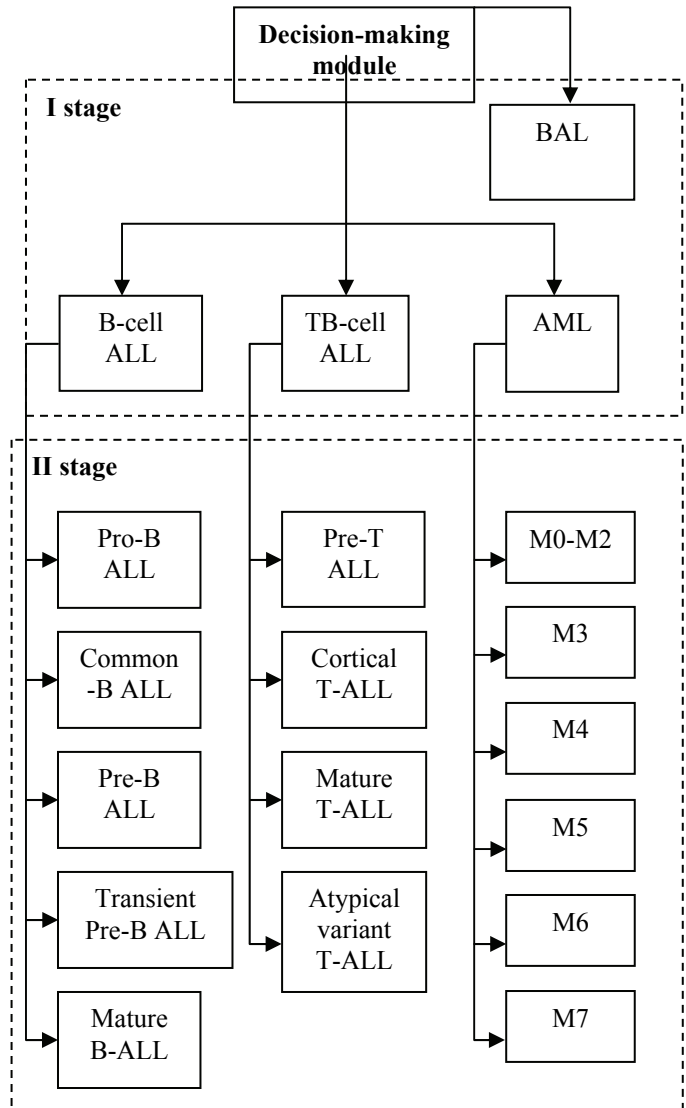


Fig. 2. Two-stage leukemia diagnostics scheme.

At the second stage, the specific disease subtype is identified for the determined general group, i. e., the in-depth analysis is pursued. The content of the model knowledge base is presented in Table I.

The set of rules, which includes the number of markers (MPO, CD19, CD3 etc.) according to WHO classification [4–5] for the biphenotype diagnostics, is the following:

Rule 1: If $((\text{MPO}=1) \text{ and } (\text{SUM3} \geq 2)) \text{ and } ((\text{CD19}=1) \text{ and } (\text{SUM1} \geq 2))$ then «Biphenotype B-ALL + AML».

Rule 2: If ((**MPO=1**) and (**SUM3**>=2)) and ((**s/cyCD3=1**) and (**SUM2**>=2)) then «Biphenotype T-ALL + AML».

Rule 3: If ((**CD19=1**) and (**SUM1**>=2)) and ((**s/cyCD3=1**) and (**SUM2**>=2)) then «Biphenotype T-ALL + B-ALL»,

where

SUM1= SUM(CD10, s/cyCD22, cyCD79a)

SUM2= SUM(CD4, CD8, CD5, CD1a, CD2, TCRα/β, TCRγ/δ)

SUM3= SUM(CD13, CD33, CD117, CD15)

The diagnostic rules for the extended groups of leukosis and the corresponding subtypes are presented as the set of weights and importance factors for each marker, which define the profile of the corresponding group or subtype. Weights are received as a result of the statistical analysis of the contingency tables [6]–[7], constructed for the two nonmetric features. In our case, the two features are immunophenotype feature and disease group or subtype. The example of the

weights for the extended leukemia groups is presented in Table II. The weight values indicate the influence of the marker for the diagnostics of the particular leukemia group.

TABLE I
THE LIST OF DIAGNOSTIC RULES FOR THE LEUKEMIA

Rules	Specification
Identification of the presence of biphenotype	Identification according to EGIL
	Identification according to modified WHO classification
Identification of the group of leukemia : B-ALL , T-ALL , AML [8–10]	
Identification of the leukemia subtype for B-ALL	
Identification of the leukemia subtype for T-ALL	
Identification of the leukemia subtype for AML	

TABLE II
MARKER WEIGHTS FOR DIFFERENT LEUKEMIA GROUPS (GROUP PROFILES)

CD-antigen	B-ALL	T-ALL	AML
CD2	0.04298	0.851064	0.262821
CD3	0.008671	0.404255	0.050633
CD5	0.028329	0.979167	0.080537
CD7	0.022663	1	0.457317
CD4p_CD8m	0.002833	0.086957	0.268293
CD4m_CD8p	0.008547	0.5	0.037037
CD4p_CD8p	0	0.434783	0.006289
CD1a	0.023881	0.543478	0.103093
tdt	0.933824	0.717949	0.105263
CD19	0.997199	0.021739	0.078313
CD20	0.384181	0	0.104167
CD22	0.97191	0	0.022556
CD10	0.91573	0.234043	0.124224
cy_CD79a	0.968858	0.357143	0.020339
s_lgm	0.017442	0	0
cy_lgm	0.151408	0.02381	0.08547
hla_dr	0.988372	0.068182	0.64557
CD13p	0.310734	0.104167	0.825301
CD33p	0.194842	0.106383	0.939759
CD117p	0.023121	0.136364	0.882353
CD14p	0.058824	0	0.188889
CD15p	0.109422	0.02439	0.590062
MPOp	0.143885	0.0175	0.772414
CD11bp	0.760486	0.139373	0.76392
CD11cp	0.189201	1.255625	1.011119
CD34	0.735955	0.404255	0.533333
CD45p_CD14m	0.715543	0.955556	0.968553

IV. KNOWLEDGE BASE AND DECISION-MAKING MODULE FOR THE LYMPHOMA DIAGNOSTICS

The general scheme of the decision-making module is presented in Fig. 3. According to Fig. 3, the diagnostics is

performed separately for children and adults, where for each group the different set of lymphoma subtypes is considered.

The set of diagnostic rules, which comprise the model knowledge base, is presented in Table III.

The same as for leukosis, the diagnostic rules are the set of

weights and importance factors for each particular marker, which define the profile of the corresponding lymphoma subtype.

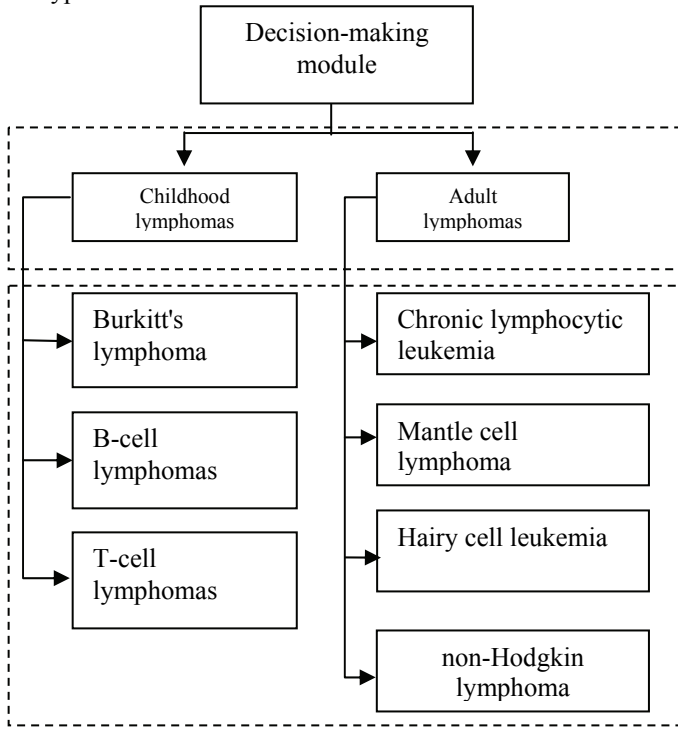


Fig. 3. Lymphoma diagnostics scheme.

TABLE III
THE LIST OF DIAGNOSTIC RULES FOR THE LYMPHOMAS

Rules	Specification
Identification of the lymphoma subtype for children	Subtypes: Burkitt's lymphoma, B-cell lymphomas (Pro-B, common-B, Pre-B), T-cell lymphomas (Pro-T, Pre-T, cortical-T).
Identification of the lymphoma subtype for adults	Subtypes: chronic lymphocytic leukemia (CLL), mantle cell lymphoma (MCL), hairy cell leukemia (HCL), Non-Hodgkin lymphoma (NHL).

The example of the weights for the adult lymphomas is presented in Table IV.

TABLE IV
MARKER WEIGHTS FOR ADULT LYMPHOMA GROUP

CD-antigen	CLL	MCL	HCL	NHL
CD19+	1.00	1.00	0.47	1.00
FMC7	0.05	0.95	1.00	0.89
CD23	1.00	0.05	0.00	0.22
CD5	1.00	1.00	0.00	0.07
CD20	0.65	1.00	1.00	0.96
CD43	0.99	0.11	0.00	0.15
CD24	1.00	0.95	0.06	0.89
CD79b	0.30	1.00	1.00	0.81
CD11c	0.00	1.00	1.00	0.67
CD25	0.00	0.00	1.00	0.00
CD103	0.00	0.00	1.00	0.00
kappa	0.37	0.39	0.42	0.67
lambda	0.26	0.63	0.58	0.31

V. DECISION INFERENCE ALGORITHM FOR THE DIAGNOSTICS OF LEUKOSIS AND LYMPHOMAS

The diagnostic decision is made using the specially developed algorithm. The model knowledge base composed of the set of rules and the binary (discretized marker values) profile of the diagnostic cases is used as the inputs to the algorithm. According to the algorithm, each disease subtype is assigned the value of the degree of manifestation PF . The degree value reflects the correspondence between the disease subtype profile and the marker profile of the patient $x = (x_1, \dots, x_N)$. To calculate the PF , the following expression is used:

$$PF = \frac{\sum C_n}{N} \quad (1)$$

where PF is the degree value for the particular disease subtype, C_n stands for coefficients of the individual markers for the particular disease subtype ($0 \leq C_n \leq 1$), $n = 1, \dots, N$, N is the number of the markers with the nonempty values. The coefficients C_n are calculated using the marker weights ω_{ij} and importance factors k_{ij} as follows:

$$C_n^i = \begin{cases} k_{ij} \cdot \omega_{ij}, & \text{если } x_j = 1 \\ k_{ij} \cdot (1 - \omega_{ij}), & \text{если } x_j = 0 \end{cases} \quad (2)$$

where C_n^i stands for marker coefficients for the i th disease subtype, ω_{ij} is weight of the j th marker for the i th disease subtype, k_{ij} is importance values for the j th marker for the i th disease subtype, x_j is the discretized value of the j th marker.

The importance of the individual specific markers for the diagnostic of the particular disease subtype is considered with the importance factor k , e.g., if $k = 2$ the value of the marker coefficient C_n is doubled.

For the differential diagnostics, all the values of the immunophenotype features, which characterize the patient, are considered. The decision is also made when not all the marker values are present. However, in the absence of the information for the most specific markers the diagnostics will be inaccurate. The diagnostic result is the list of the disease subtype with the corresponding degrees of manifestation $PF_1, PF_2, \dots, PF_M$. Finally, the degree values are divided by the maximal value to receive the result in the interval $[0, 1]$:

$$PF_m = \frac{PF_m}{\max(PF_1, PF_2, \dots, PF_M)}, \quad m = \overline{1, M} \quad (3)$$

The example of the decision-making process for the B-ALL diagnostic case with the input values in Table V is considered. The first stage of the leukemia diagnostics is performed, i. e., the recognition of the extended groups: B-ALL, T-ALL, AML. According to Table V, 17 markers are negative (expressed on less than 20% cells), 8 – positive and for 2 markers the values are absent.

To calculate the value of the degree of manifestation PF

consider the weights for the B-ALL profile in Table II. The importance factors of the markers are presented in Table VI.

TABLE V
THE MARKER PROFILE FOR THE CONSIDERED EXAMPLE

No.	Marker name	Value 1 – positive, 0 – negative, –1 – absent value	No.	Marker name	Value 1 – positive, 0 – negative, –1 – absent value
1	CD2	0	15	s_lgm	0
2	CD3	0	16	cy_lgm	0
3	CD5	0	17	hla_dr	1
4	CD7	0	18	CD34	0
5	CD4p_CD8m	0	19	CD45p_CD14m	1
6	CD4m_CD8p	0	20	CD13p	0
7	CD4p_CD8p	0	21	CD33p	0
8	CD1a	0	22	CD117p	0
9	tdt	1	23	MPOp	0
10	CD19	1	24	CD15p	0
11	CD20	1	25	CD11bp	–1
12	CD22	1	26	CD11cp	0
13	CD10	1	27	CD14p	–1
14	cy_CD79a	1			

The value of PF consists of two summands: C_1 – the sum of marker coefficients with a positive input value, C_2 – the sum of marker coefficients with a negative input value. C_1 and C_2 are calculated subject to the marker importance factors. For the considered example, $C_1 = 6.87$ and $C_2 = 14.98$.

TABLE VI
IMPORTANCE COEFFICIENTS FOR THE B-ALL PROFILE

No.	Marker name	Importance	No.	Marker name	Importance
1	CD2	1.00	15	s_lgm	1.00
2	CD3	1.00	16	cy_lgm	1.00
3	CD5	1.00	17	hla_dr	2.00
4	CD7	1.00	18	CD34	1.00
5	CD4p_CD8m	1.00	19	CD45p_CD14m	1.00
6	CD4m_CD8p	1.00	20	CD13p	1.00
7	CD4p_CD8p	1.00	21	CD33p	1.00
8	CD1a	1.00	22	CD117p	1.00
9	tdt	2.00	23	MPOp	1.00
10	CD19	2.00	24	CD15p	1.00
11	CD20	1.00	25	CD11bp	1.00
12	CD22	2.00	26	CD11cp	1.00
13	CD10	2.00	27	CD14p	1.00
14	cy_CD79a	2.00			

Hence $PF_{B-ALL} = \frac{C_1 + C_2}{25} = 0,874$. The values of

$PF_{T-ALL} = 0,493$ and $PF_{AML} = 0,494$ are calculated just the same. The diagnosis is made in accordance with the maximal value $PF^* = \max\{PF_{B-ALL}, PF_{T-ALL}, PF_{AML}\}$ and corresponds to the B-ALL leukemia group. The diagnosis predicted with the model is equal to the true diagnosis.

VI. SOFTWARE FOR THE DIAGNOSTICS OF LEUKEMIA S AND LYMPHOMAS

The proposed models and decision inference algorithm are implemented as software in Visual Studio environment [11]. The software interface for the data input and editing is created dynamically on the basis of the internal electronic presentation of the patient registration card. The process of the decision-making is performed using the electronic diagnostic rules, which are stored in the XML files [12]. The main functions, which are implemented, are the following:

- registration and storage of the results of the immune marker analysis on the basis of the XML template of the patient registration card;
- storage of the diagnostic rules for the leukemia and lymphoma subtypes on the basis of the XML templates;
- editing and checking the filling completeness of the patient registration card;
- performing the disease diagnostics according to the developed decision-making models for leukemia and lymphomas;
- adjustment of the disease diagnostics level: leukemia or lymphoma subtypes;
- delivery of the diagnostic results as an ordered list of the disease subtypes with the corresponding degrees of manifestation;
- delivery of the information according to the selected diagnosis and the list of the corresponding coexpressions;
- observing and storing the patient registration card;
- getting the reference information for the leukemia and lymphoma subtypes.

Figure 4 presents the screenshot with the initial patient data and opened control tab – B-ALL marker type.

Software allows making the adjustment of the diagnostic level to select the disease type for further decision-making process (Fig. 5). After the selection of the certain disease type, the corresponding diagnostic rules are loaded into the computer program. The results of the two-stage leukemia diagnostics are depicted in Fig. 6, where the initial patient data with known diagnosis Common B-ALL are used.

Fig. 7 presents the diagnostic results for the adult patient with hairy cell leukemia.

D:\MarkerDiagnosticEng\MarkerDiagnostic\bin\Release\Save\case2_CommonB(E).xml

T-ALL marker type AML marker type Linear-nonspecific marker type

Official information (physician) Passport data(patient) B-ALL marker type

Marker	Expression value	Threshold value	Select
CD19+	83	20	☒
CD20+	15	20	☒
CD22+	85	20	☒
CD10+	CD10 expresses the young forms of B-lymphocyte and subpopulation of the cortical lymphocytes. CD10 expresses most types of acute lymphoblastic leukemia		☒
cy_IgM+	9	10	☒
s_IgM+	2	20	☒
s_IgA+	0	20	☐
s_IgG+	0	20	☐
CD19+kappa	0	20	☐
CD19+lambda	0	20	☐
CD79a+	43	10	☒
CD21+	0	20	☐
CD23+	0	20	☐
CD24+	0	20	☐
CD25+			☐

Fig. 4. Patient input data with the marker expression values.

Parameter adjustment

Level of diagnostics

Leukosis diagnostics

Leukosis diagnostics

Lymphoma diagnostics

Performing the diagnostic of leukosis subtypes

Template of diagnostic rules

Standard diagnostic rules using the panel of monoclonal antibodies

\Rule\RegCard.SubData0

Cancel Apply

Fig. 5. Selection of the disease type for diagnostics.

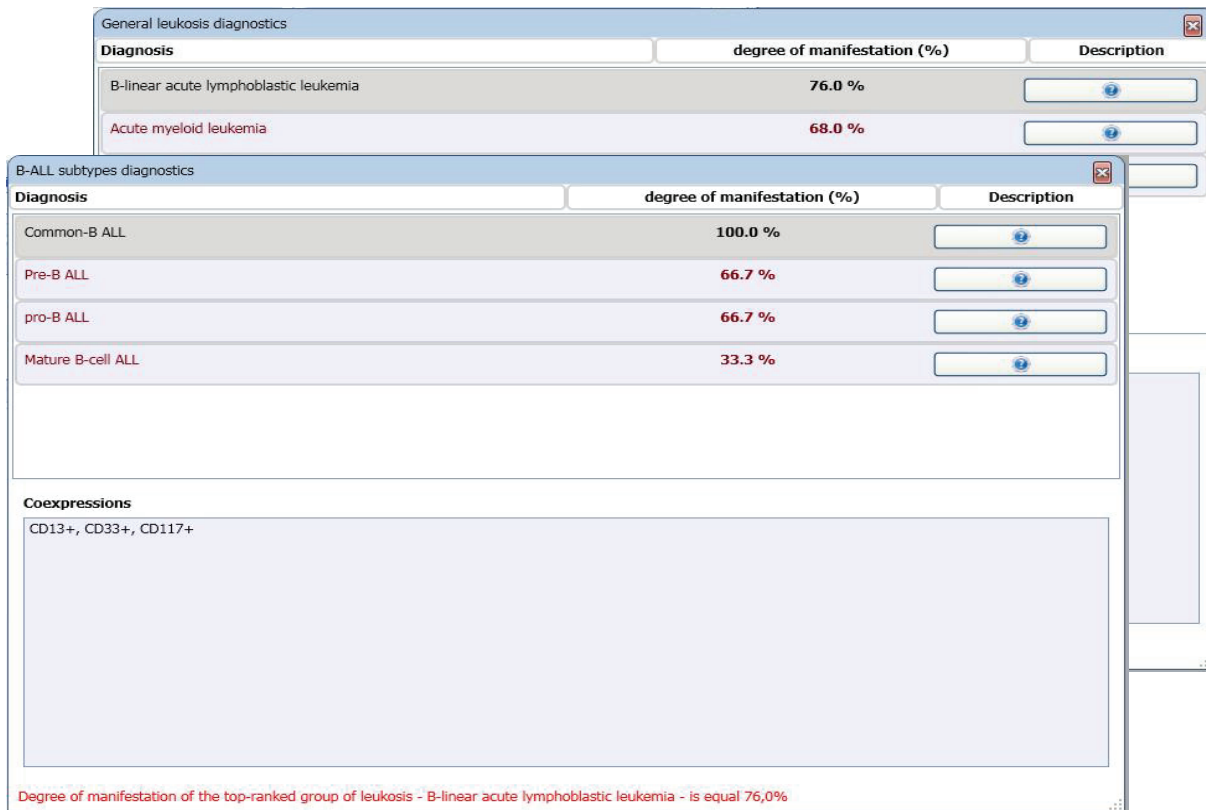


Fig. 6. Results of the two-stage leukemia diagnostics.

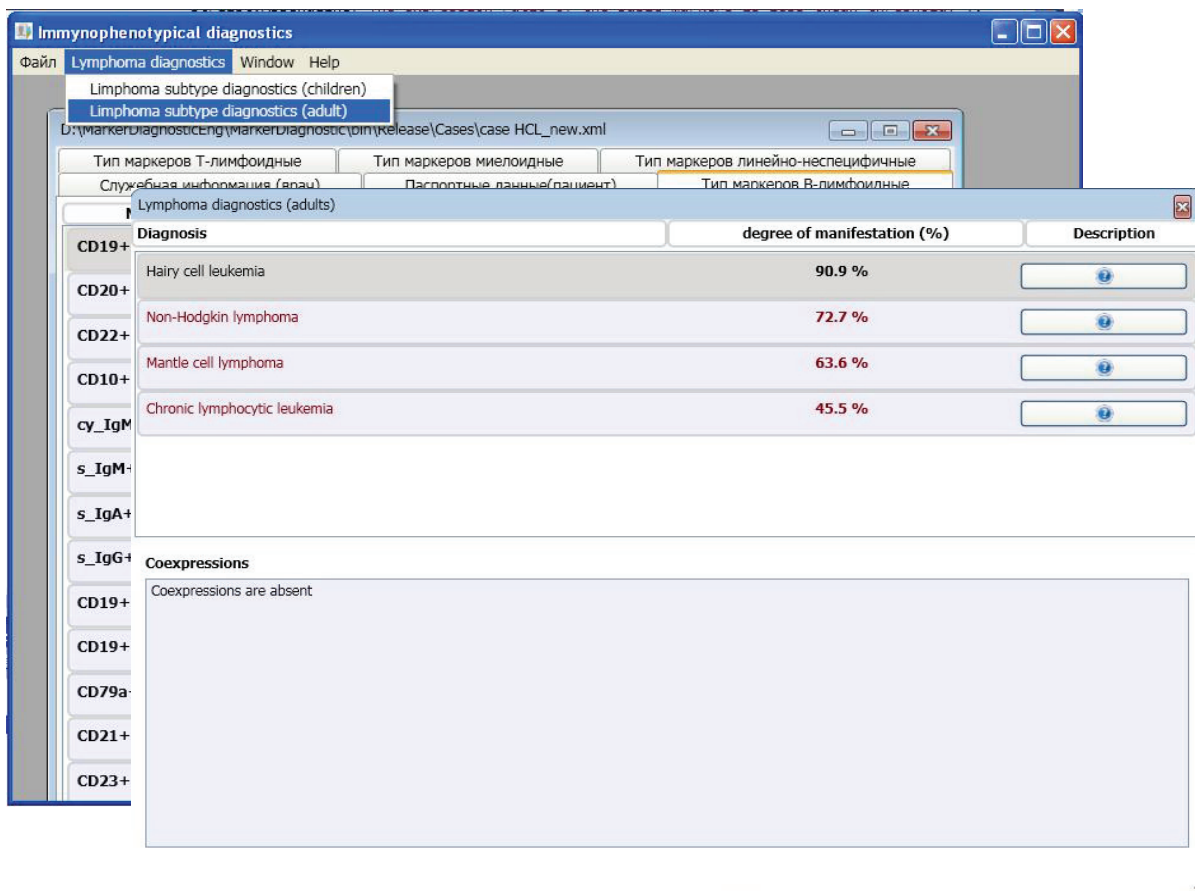


Fig. 7. Diagnostic results for the adult patient.

The constituent part of the software system is the electronic help on the basic types of leukemia and lymphoma. The help system allows dynamically browsing the information for each individual leukemia or lymphoma subtype

VII. VERIFICATION OF THE DECISION SUPPORT MODELS

Verification of the decision support model for the diagnostics of the lymphoma subtypes in children was performed using the input dataset with 79 patients, characterized by the expression values of the 31 markers. The same dataset was used for the identification of the individual marker weights for the disease subtypes, i.e., disease subtype profiles, which are used as the diagnostic rules for the disease prediction.

Diagnostic (classification) accuracy for the dataset was equal to 97.5%. Only one data case with Burkitt's lymphoma and one data case with B-cell lymphomas were incorrectly predicted. The classification results are presented in Table VII.

The classification accuracy with the standard combination of markers was equal to 96%.

TABLE VII

CLASSIFICATION RESULTS FOR THE CHILD DATASET WITH LYMPHOMAS

Lymphoma subtype		True values		
		Burkitt's lymphoma	T-cell lymphomas	B-cell lymphomas
Prediction	Burkitt's lymphoma	55	0	1
	T-cell lymphomas	0	19	0
	B-cell lymphomas	1	0	3

Verification of the decision support model for the diagnostics of the lymphoma subtypes in children was performed using the input dataset with 180 patients, characterized by the expression values of the 13 markers. Diagnostic accuracy for the dataset was equal to 97.2%. Only two data cases with NHL and 3 cases with CLL were incorrectly classified. The classification results are presented in Table VIII.

The classification accuracy with the standard combination of markers was equal to 94%.

TABLE VIII

CLASSIFICATION RESULTS FOR THE ADULT DATASET WITH LYMPHOMAS

Lymphoma subtype		True values			
		CLL	MCL	HCL	NHL
Prediction	CLL	114	0	0	3
	MCL	0	19	0	0
	HCL	0	0	17	0
	NHL	1	1	0	25

Verification of the decision support model for the diagnostics of the leukemia subtypes was performed using the input dataset with 577 patients, characterized by the expression values of the 27 markers. At the first stage of diagnostics, 7 data cases were revealed as biphenotypes, among them 3 cases were attributed to B-ALL+T-ALL biphenotype and 4 cases to B-ALL+AML biphenotype. The

description of the binary marker profiles of the detected cases with biphenotypes are depicted in Table IX. For these cases the markers, which are specific to several types of leukemia, are simultaneously expressed (e.g., for the first case the T-ALL markers – CD2, CD3, CD5, CD7 and B-ALL markers – CD19, CD20, CD22, CD79).

TABLE IX

BINARY MARKER PROFILES FOR CASES WITH BIPHENOTYPES (Bif_BT CORRESPONDS TO B-ALL+T-ALL, Bif_BM CORRESPONDS TO B-ALL+AML)

	Bif_BT	Bif_BT	Bif_BT	Bif_BM	Bif_BM	Bif_BM	Bif_BM
CD2	1	1	1	0	0	0	0
CD3	1	1	1	0	0	0	0
CD5	1	1	1	0	0	0	0
CD7	1	1	1	0	0	0	0
CD4p_CD8m	0	0	0	0	0	0	0
CD4m_CD8p	0	0	0	0	0	0	0
CD4p_CD8p	0	0	0	0	0	0	0
CD1a		0	0	0	0	0	0
tdt	1			1	1	1	1
CD19	1	1	1	1	1	1	1
CD20	1	1	1	1	0	0	0
CD22	1	1	1	1	1	1	1
CD10		1	1	1	1	1	1
cy_CD79a	1			1	1		1
s_lgm	0	0	0	0	0	0	0
cy_lgm	1			0	0	0	0
hla_dr	1	1	1	1	1	1	1
CD13p	0	0	1	1	1	1	1
CD33p	0	0	0	1	1	1	0
CD117p	0	0	0	0	0	0	1
CD14p							
CD15p		0	0	0	0	0	0
MPOp	0			1	1	1	1
CD11bp				0		0	
CD11cp		0	0				
CD34	1	1	0	0	1	0	1
CD45p_CD14m		0	1	1	1	1	1

At the first stage of diagnostics, the diagnostic accuracy was 98.8%. Four AML cases and 3 T-ALL cases were incorrectly classified. The classification results are presented in Table X. The classification accuracy with the standard combination of markers was equal to 98.2%.

TABLE X

CLASSIFICATION RESULTS FOR THE LEUKEMIA CASES

Acute leukemia group		True values		
		B-ALL	T-ALL	AML
Prediction	B-ALL	359	2	1
	T-ALL	0	45	3
	AML	0	1	163

At the second stage of diagnostics, the diagnostic accuracy was 88.3 % for B-ALL subtypes, 82.6 % for T-ALL subtypes. The most number of errors was due to the classification of the AML subtypes (classification accuracy ~70%). Such a result can be explained by the great number of common antigens, expressed on the cell surface of the myeloid and monocytic lines; therefore, the final diagnosis must take into account the results of morphological and cytochemical analysis.

VIII. CONCLUSION

We have developed and implemented as software the decision support models for the immunophenotypical diagnostics of leukosis and lymphomas. These models allow performing the diagnostics using both the standard marker combinations and the diagnostic rules constructed as a result of the analysis of the real dataset stored at the State Institution "Belarusian Research Center for Pediatric Oncology, Hematology and Immunology".

According to the verification on the real datasets of patients with leukemia and lymphomas, the use of the generated diagnostic rules gives slightly better classification accuracy in comparison with the standard combinations.

The software is assigned to support the decision-making for the differential diagnostics of leukemia and lymphomas on the basis of immune marker analysis. The system provides the possibility to register, edit and store the patient data in the XML-format.

Applying the software will allow for more accurate disease diagnostics on the basis of the personalization of the diagnostic process, refining the diagnostics using the specialized decision support models adjusted for marker coexpressions. The software can also be considered an educational tool for training the specialists of the clinical-diagnostic medical units in order to raise their professional level.

REFERENCES

- [1] Visualization technologies in medicine. Diagnostics of the acute leukosis and the tumors of the central nervous system in children / O. Aleynikova etc. Minsk: UIIP NAS Belarus, 2007, 244 p.
- [2] N. F. Matsatsinis, Y. Siskos. *Intelligent support systems for marketing decisions*, Kluwer Academic Publishers, 2002.
<http://dx.doi.org/10.1007/978-1-4615-1147-2>
- [3] S. Krishna. *Introduction to Database and Knowledge-base Systems*. World Scientific Publishing, 1992, p. 328. ISBN 981-02-0619-4.
<http://dx.doi.org/10.1142/1374>

- [4] X. F. Zhao, I. Gojo, T. York, Y. Ning and M. R. Baer, "Diagnosis of biphenotypic acute leukemia: a paradigmatic approach," *Int. J. Clin. Exp. Pathol.*, vol. 3(1), pp. 75–86, 2010.
- [5] M.C.Béné, "Biphenotypic, bilineal, ambiguous or mixed lineage: strange leukemias," *Haematol.*, vol. 94(7), pp. 891–893, 2009.
<http://dx.doi.org/10.3324/haematol.2009.007799>
- [6] S. Glanz, *Medical-biological statistics*. M: Praktika. 1999, 459 p.
- [7] A. Agresti, *An introduction to categorical data analysis*, 2nd ed., John Wiley & Sons, Inc., 2007, 371 p.
- [8] J. W. Vardiman, "The World Health Organization (WHO) classification of tumors of the hematopoietic and lymphoid tissues: an overview with emphasis on the myeloid neoplasms," *Chem. Biol. Interact.*, vol. 184(1–2), pp. 16–20, 2010. <http://dx.doi.org/10.1016/j.cbi.2009.10.009>
- [9] G. Basso, B. Buldini, L. De Zen and A. Orfao, "New methodologic approaches for immunophenotyping acute leukemias," *Haematologica*, vol. 86(7), pp. 675–692, 2001.
- [10] A. C. Seegmiller, S. H. Kroft, N. J. Karandikar and R. W. McKenna, "Characterization of Immunophenotypic Aberrancies in 200 Cases of B Acute Lymphoblastic Leukemia," *Am. J. Clin. Pathol.*, vol. 132(6), pp. 940–949, 2009. <http://dx.doi.org/10.1309/AJCP8G5RMTWUEMUU>
- [11] A. Hejlsberg, M. Torgersen, S. Wilmuth, P. Golde, *The C# Programming Language (Covering C# 4.0)*, 4th ed., Microsoft Corporation, 2011, 815 p.
- [12] E. T. Ray. *Learning XML*, 2nd ed., O'Reilly Media, 2003, 418 p.

Natalia Novoselova received the Doctoral degree in Computer Sciences from the United Institute of Informatics Problems, National Academy of Sciences of Belarus (NASB) in 2008. Since 2000 she has been a Senior Scientific Researcher at the Laboratory of Bioinformatics, United Institute of Informatics Problems NASB in Minsk, Belarus. Her research interests include data mining and statistical methods, algorithms for multivariable data analysis, especially in the field of biology and medicine. She is the author of more than 20 research publications in these areas.

Address: Laboratory of Bioinformatics, United Institute of Informatics Problems, Surganova Str. 6, Minsk, 220012, Belarus. Phone: +375-17-2842092

E-mail: novosel@newman.bas-net.by.

Igor Tom received the Doctoral degree in Computer Science from the Institute of Engineering Cybernetics, National Academy of Sciences of Belarus (NASB) in 1986. Since 1999 he has been a Chief of the Laboratory of Bioinformatics at the United Institute of Informatics Problems of NASB. His current research interests are in the fields of the development of intelligent methods of data analysis and information technologies for medical and industrial applications. He is the author and co-author of more than 180 scientific publications, including 50 full papers in journal.

Address: Laboratory of Bioinformatics, United Institute of Informatics Problems, Surganova Str. 6, Minsk, 220012, Belarus. Phone: +375-17-2842153

E-mail: tom@newman.bas-net.by.

Michael Belevtsev received the Doctoral degree in Biological Sciences in 2003. Since 2003 he has been the Head of the Laboratory of the Immunological Research at the Belarusian Research Center, for Pediatric Oncology, Hematology and Immunology. Since 2012 he has also been a Deputy Director. His research interests include innate defects of the immune system, cancer immunology. He has more than 100 publications in the fields of immunology, oncology.

Address: Pos. Lesnoe-2, Minsk district, 223040, Belarus.

E-mail: belevtsev_m@mail.ru.

Natalija Novoselova, Igors Toms, Mihaels Belevcevs. Lēmumu pieņemšanas modeļi un programmnodrošinājums diferenciālajai imūnajai fenotipiskajai leikožu un limfomu diagnostikai

Rakstā apskatīti lēmumu pieņemšanas modeļi un atbilstošais programmnodrošinājums, kas paredzēti, lai palīdzētu speciālistam veikt leikožu un limfomu diagnostiku atbilstoši imunoloģiskajai fenotipizācijai. Modeļu zināšanu bāzes iekļauj likumu kopas, kas paredzētas leikožu un limfomu diagnostikai un kas izveidotas, balstoties uz statistiskās analīzes, kas veikta diagnostisko gadījumu datubāzē, kas uzkrāta valsts uzņēmumā „Republikas bērnu onkoloģijas, hematoloģijas un imunoloģijas zinātniski praktiskais centrs”, kā arī izmantojot standarta likumus. Pamata leikožu, limfomu un to atbilstošo apakštipu slimību diagnostikas likumi ir svara koeficientu un svarīguma koeficientu kopas katram diagnostiskajam marķierim, kas nosaka atbilstošās grupas vai apakštipa profilu. Svara koeficienti iegūti, balstoties uz saistīto tabulu statistisko analīzi. Diagnostiskie lēmumi tiek pieņemti, izmantojot īpaši izstrādātu algoritmu, kas aprēķina slimības apakštipu izteiktības pakāpi, balstoties uz imunoloģiski fenotipisko rādītāju ievades vērtībām un izmantojot modeļa zināšanu bāzi. Leikožu un limfomu diagnostiskajiem modeļiem ir kopīga moduļu struktūra, bet to diagnostisko likumu saturs un lēmumu pieņemšanas moduļa organizācija atšķiras. Modeļa izejas dati tiek attēloti, kā ranžēts slimības apakštipu saraksts, kas ļauj novērtēt katra slimības apakštipa izteiktības pakāpi katram konkrētajam diagnostikas gadījumam. Saraksta augšgalā atrodas visiespējamākie apakštipi. Specializētais programmnodrošinājums realizē izstrādātos lēmumu pieņemšanas modeļus, nodrošinot

divpakāpju leikožu diagnostikas un limfomu diagnostikas veikšanu, izmantojot izstrādāto diagnostisko likumu kopu, kā arī diagnostikas rezultātu un atbilstošās uzzīņu informācijas atainošanu. Ieejas dati ir organizēti atbilstoši vienotajai pacienta reģistrācijas kartei, kas nodrošina darbu dažādos diagnostikas līmeņos: izvērtajā leikožu diagnostikas līmenī, padziļinātajā leikožu diagnostikas līmenī un bērnu un pieaugušo limfomu diagnostikas līmenī. Izstrādātās sistēmas izmantošana veicinās precīzētas slimības kartes iegūšanu, ņemot vērā pacienta individuālos ģenētiskos datus, kā arī uzlabos diagnostiku, izmantojot specializētus lēmumu pieņemšanas modeļus leikožu un limfomu imunoloģiski fenotipiskajai diagnostikai, ņemot vērā paralēlo ekspresiju. Tāpat sistēmu var izmantot par mācību līdzekli studenti, zinātnisko grādu pretendenti un praktizējošie speciālisti, apgūstot leikožu un limfomu diagnostikas pamatmetodes, kas izmanto imunoloģiski fenotipiskos rādītājus.

Наталья Новоселова, Игорь Том, Михаэль Белевцев. Модели принятия решений и программное обеспечение для дифференциальной иммунофенотипической диагностики лейкозов и лимфом

В статье рассматриваются модели принятия решений и соответствующее программное обеспечение, позволяющее помочь специалисту при диагностике лейкозов и лимфом по результатам иммунофенотипирования. Базы знаний моделей включают в себя наборы правил по диагностике лейкозов и лимфом, которые построены на основе статистического анализа базы диагностических случаев, накопленной в государственном учреждении «Республиканский научно-практический центр детской онкологии, гематологии и иммунологии», а также стандартные правила. Правила для диагностики основных групп лейкозов, лимфом и соответствующих им подтипов заболевания представляют собой наборы весовых коэффициентов и коэффициентов важности для каждого диагностического маркера, которые определяют профиль соответствующей группы или подтипа. Весовые коэффициенты были получены на основе статистического анализа таблиц сопряженности. Диагностическое решение выносится с использованием специально разработанного алгоритма, позволяющего рассчитать степени выраженности подтипов заболевания на основе входных значений иммунофенотипических показателей с использованием базы знаний модели. Диагностические модели для лейкозов и лимфом имеют общую модульную структуру с отличающимся составом диагностических правил и организацией модуля принятия решений. Выходные данные модели представляются в виде упорядоченного списка подтипов заболевания, что позволяет оценить степень выраженности каждого из подтипов заболевания для конкретного диагностического случая. Вверху списка располагается наиболее вероятный подтип. Специализированное программное обеспечение реализует разработанные модели принятия решений, обеспечивая проведение двухуровневой диагностики лейкозов и диагностику лимфом с использованием набора разработанных диагностических правил, просмотр результатов диагностики и соответствующей справочной информации. Входные данные организованы согласно унифицированной регистрационной карте пациента, обеспечивающей возможность работы на разных уровнях диагностики: уровень расширенной диагностики лейкозов, уровень углубленной диагностики лейкозов, уровень диагностики лимфом для детей и взрослых. Использование разработанной системы будет способствовать получению уточненной картины заболевания с учетом индивидуальных генетических данных пациента, совершенствованию диагностики за счет использования специализированных моделей принятия решений для иммунофенотипической диагностики лейкозов и лимфом с учетом коэкспрессий. Система также может быть использована в качестве обучающего средства для студентов, аспирантов, практикующих специалистов по основным методам диагностики лейкозов и лимфом по набору иммунофенотипических показателей.

Least Squares Support Vector Machine Based on Wavelet-Neuron

Yevgeniy Bodyanskiy¹, Olena Vynokurova², Oleksandra Kharchenko³,

¹⁻³*Kharkiv National University of Radio Electronics*

Abstract – In this paper, a simple wavelet-neuro-system that implements learning ideas based on minimization of empirical risk and oriented on information processing in on-line mode is developed. As an elementary block of such systems, we propose using wavelet-neuron that has improved approximation properties, computational simplicity, high learning rate and capability of local feature identification in data processing. The architecture and learning algorithm for least squares wavelet support machines that are characterized by high speed of operation and possibility of learning under conditions of short training set are proposed.

Keywords – Adaptive wavelet function, forecasting, least squares support vector machine, non-linear non-stationary time series, wavelet-neuron.

I. INTRODUCTION

Nowadays computational intelligence systems are widely used to solve a large class of problems associated with information processing, which is provided in different forms, from simple tables “object-property” to complex multidimensional non-stationary stochastic time series. Systems of computational intelligence, such as artificial neural networks, neuro-fuzzy systems and wavelet systems have a high processing speed, universal approximation properties, identification of local features [1]–[7].

However, the existing approaches for their training require large volumes of training data set, wherein the volume of the original training data set should be at least two orders bigger than the number of estimated parameters of such systems. Unfortunately, this situation does not always occur, especially in solving practical tasks. In many practical problems, for example, medical diagnostics, forecasting of financial indices etc., the volume of training data set is quite insufficient for constructing and training the effective system of computational intelligence.

The solving of problems in such a situation using conventional identification theory methods [8] is not effective, for which reason the method based on empirical risk minimization was proposed by V. N. Vapnik [9]–[12] and a support vector machine was designed based on this method.

Traditional support vector machine is a computational system that minimizes the empirical risk, but from viewpoint of practical implementation it is a sufficiently complex system, because it is related to the solving of nonlinear programming task at each step apart from high dimension with constraints in form of inequalities. Thus, in such a case the idea was wonderful; however, its implementation in on-line mode was not successful [2].

Therefore, at the beginning of this century, modification of this system has been proposed, called LS-SVM (least squares support vector machine) [13]. Here principal conditions were changed in such a way that it was necessary to solve a quadratic programming problem with equality constraints at each step, and it had already created the preconditions for the neural network implementation of this approach. But still, this machine was quite complicated and therefore further attempts were made to improve, firstly, speed, to simplify computing implementation and, secondly, to reduce training sample volume.

Based on fuzzy SVM [14], [15], the wavelet least squares support vector machine was proposed and compared with all previous variants – it had improved approximation properties, but, as a result, numerical implementation had become more difficult [16], [17]. The basis of LS-SVM was radial basis function network [18], where radial basis functions were replaced by multidimensional wavelet functions. For simplification of software implementation, by reducing the number of tuning parameters and increasing the speed we have proposed to use wavelet-neuron as the basic architecture [19]–[21] that has improved approximation properties and high operation speed. However, the disadvantages of wavelet-neuron are connected with its learning algorithms that do not allow tuning all parameters under a small data set.

Alternative to learning based on optimization is the learning based on memory that is based on the concept “neurons in the data points” [18]. The most typical representative of neural networks with such learning is General Regression Neural Network (GRNN); however, it solves a task of interpolation instead of approximation that essentially complicates its usage in noisy data processing.

Therefore, the development of sufficiently simple wavelet-neuro-fuzzy systems is advisable. Such systems implement learning based on empirical risk minimization and are oriented to information processing in on-line mode. Wavelet-neuron that has improved approximation and extrapolation properties can be used as a basic element of such systems.

II. WAVELET-NEURON ARCHITECTURE

Let us consider the wavelet-neuron architecture, shown in Fig. 1. As seen, wavelet-neuron is quite close to the standard n -input formal neuron, but instead of tuning synaptic weights it contains wavelet-synapses $WS_i, i = 1, 2, \dots, n$, where the tuning parameters are not only synaptic weights, but all parameters of adaptive wavelet activation functions $\varphi_{ji}(x_i(k))$ [21].

When input vector

$$x(k) = (x_1(k), x_2(k), \dots, x_n(k))^T$$

is fed to the wavelet-neuron input (here $k=1,2,\dots$ is current discrete time) its output can be written in the form

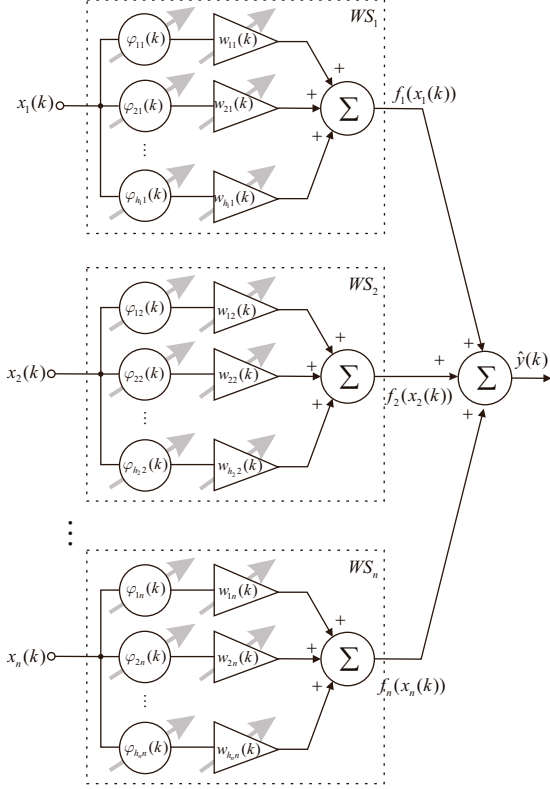


Fig. 1. Wavelet-neuron.

$$\hat{y}(k) = \sum_{i=1}^n f_i(x_i(k)) = \sum_{i=1}^n \sum_{j=1}^{h_i} w_{ji}(k) \varphi_{ji}(x_i(k)) \quad (1)$$

where $w_{ji}(k)$ is synaptic weight, $\varphi_{ji}(x_i(k))$ is wavelet activation function.

The different analytical wavelet functions can be used as activation functions, but for adaptive tuning of wavelet neuron we use an adaptive wavelet activation function which was proposed in [22], [23] and has the form

$$\varphi_{ji}(x_i(k)) = (1 - \alpha_{ji}(k) t_{ji}^2(k)) \exp(-t_{ji}^2(k)/2) \quad (2)$$

where $t_{ji}(k) = (x_i(k) - c_{ji}(k)) / \sigma_{ji}(k)$; $c_{ji}(k)$ is a center parameter of adaptive wavelet function; $\sigma_{ji}(k)$ is a width parameter of adaptive wavelet function; $\alpha_{ji}(k)$ is a shape parameter of adaptive wavelet function.

Tuning parameter α_{ji} allows changing the shape of an adaptive wavelet activation function during the training process of network, and, as a result, for $\alpha_{ji}=0$ we obtain Gaussian function, and when $\alpha_{ji}=1$ we obtain wavelet-

function “Mexican Hat”, and when $0 < \alpha_{ji} < 1$ – hybrid activation function.

Figure 2 shows the adaptive wavelet activation function with different parameters α and σ .

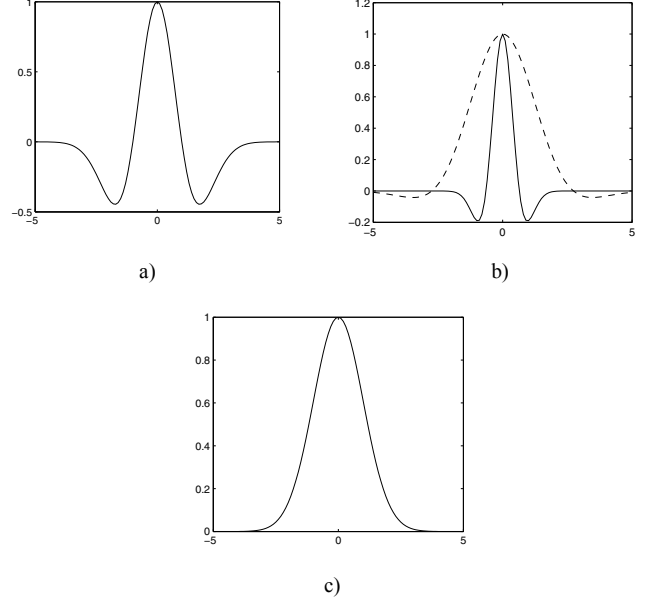


Fig. 2. Adaptive wavelet activation function: a – $\alpha = 1, \sigma = 1$; b – dashed line $\alpha = 0.3, \sigma = 1.5$; solid line $\alpha = 0.6, \sigma = 0.5$; c – $\alpha = 0, \sigma = 1$.

The learning task is to find synaptic weights $w_{ji}(k)$, centers $c_{ji}(k)$, widths $\sigma_{ji}^{-1}(k)$ and shape parameters $\alpha_{ji}(k)$ of adaptive wavelet activation function on each k -th iteration, which optimizes the learning criterion.

III. LEARNING ALGORITHM FOR ALL WAVELET-NEURON PARAMETERS

When the training data set is sufficiently large, as learning criteria we can use the conventional squared error function in the form

$$E(k) = \frac{1}{2} (y(k) - \hat{y}(k))^2 = \frac{1}{2} \left(y(k) - \sum_{i=1}^n \sum_{j=1}^{h_i} w_{ji}(k) \varphi_{ji}(x_i(k)) \right)^2 \quad (3)$$

where $y(k)$ is a reference signal.

Introducing some denominations in the form $\varphi_i(x_i(k)) = (\varphi_{1i}(x_i(k)), \dots, \varphi_{h_i i}(x_i(k)))^T$, $w_i(k) = (w_{1i}(k), \dots, w_{h_i i}(k))^T$, $c_i(k) = (c_{1i}(k), \dots, c_{h_i i}(k))^T$, $\sigma_i^{-1}(k) = (\sigma_{1i}^{-1}(k), \dots, \sigma_{h_i i}^{-1}(k))^T$, $\alpha_i(k) = (\alpha_{1i}(k), \dots, \alpha_{h_i i}(k))^T$, $t_i(k) = (t_{1i}(k), \dots, t_{h_i i}(k))^T$, we can write the learning algorithm in the form

$$\begin{cases} w_i(k+1) = w_i(k) + (e(k) J_i^w(k)) / r_i^w(k), \\ c_i(k+1) = c_i(k) + (e(k) J_i^c(k)) / r_i^c(k), \\ \sigma_i^{-1}(k+1) = \sigma_i^{-1}(k) + (e(k) J_i^\sigma(k)) / r_i^\sigma(k), \\ \alpha_i(k+1) = \alpha_i(k) + (e(k) J_i^\alpha(k)) / r_i^\alpha(k), \end{cases} \quad (4)$$

$$\begin{cases} r_i^w(k+1) = \beta r_i^w(k) + \|J_i^w(k)\|^2, \\ r_i^c(k+1) = \beta r_i^c(k) + \|J_i^c(k)\|^2, \\ r_i^\sigma(k+1) = \beta r_i^\sigma(k) + \|J_i^\sigma(k)\|^2, \\ r_i^\alpha(k+1) = \beta r_i^\alpha(k) + \|J_i^\alpha(k)\|^2 \end{cases} \quad (5)$$

where $0 \leq \beta \leq 1$ is a forgetting factor,

$$\begin{aligned} J_i^w(k) &= (J_{li}^w(k), \dots, J_{hi}^w(k))^T, \quad J_i^c(k) = (J_{li}^c(k), \dots, J_{hi}^c(k))^T, \\ J_i^\sigma(k) &= (J_{li}^\sigma(k), \dots, J_{hi}^\sigma(k))^T, \quad J_i^\alpha(k) = (J_{li}^\alpha(k), \dots, J_{hi}^\alpha(k))^T, \\ J_{ji}^w(k) &= (1 - \alpha_{ji}(k)t_{ji}^2(k)) \exp(-t_{ji}^2(k)/2), \\ J_{ji}^c(k) &= w_{ji}(k) \sigma_{ji}^{-1}(k) \left((2\alpha_{ji}(k) + 1) t_{ji}(k) - \alpha_{ji}(k) t_{ji}^3(k) \right) \times \\ &\times \exp(-t_{ji}^2(k)/2), \\ J_{ji}^\sigma(k) &= w_{ji}(k) (x_i(k) - c_{ji}(k)) \left((2\alpha_{ji}(k) + 1) t_{ji}(k) - \alpha_{ji}(k) t_{ji}^3(k) \right) \times \\ &\times \exp(-t_{ji}^2(k)/2), \quad J_{ji}^\alpha(k) = e(k) w_{ji}(k) t_{ji}^2(k) \exp(-t_{ji}^2(k)/2). \end{aligned}$$

Easy to see that for $\beta = 1$ method (4), (5) has stochastic approximation properties of adaptive identification algorithm of Goodwin-Ramage-Caines [24], and for $\beta = 0$ the learning method – the popular Widrow-Hoff learning algorithm.

As can be seen, the use of the modified quasi-Newtonian learning algorithm does not complicate numerical realization of tuning all parameters of wavelet-neuron and provides an increased convergence rate.

IV. WAVELET-NEURON LEARNING ALGORITHM BASED ON EMPIRICAL RISK MINIMIZATION

In the case when we have a short data set, the proposed learning algorithm (4), (5) cannot tune all parameters of a network. Thus, the methods based on empirical risk minimization are more effective in this situation.

Introducing the $(nh \times 1)$ – vector of adaptive wavelet activation functions

$$\begin{aligned} \varphi(x(k)) &= (\varphi_{11}(x_1(k)), \dots, \varphi_{h1}(x_1(k)), \varphi_{12}(x_2(k)), \dots, \\ &\varphi_{h2}(x_2(k)), \dots, \varphi_{hi}(x_i(k)), \dots, \varphi_{1n}(x_n(k)), \dots, \varphi_{hn}(x_n(k)))^T, \end{aligned} \quad (6)$$

and corresponding to it synaptic weights vector of wavelet-neuron

$$w = (w_{11}, \dots, w_{h1}, w_{12}, \dots, w_{h2}, \dots, w_{hi}, \dots, w_{1n}, \dots, w_{hn})^T, \quad (7)$$

we can rewrite formula (1) in the form

$$\hat{y}(k) = w^T \varphi(x(k)). \quad (8)$$

The learning of wavelet-neuron using a least squares support vector machine is connected with the optimization of criterion in the form

$$E(N) = \frac{1}{2} \|w\|^2 + \frac{\gamma}{2} \sum_{k=1}^N e^2(k) \quad (9)$$

with regard to constrains in the form of N linear equations

$$y(k) = w^T \varphi(x(k)) + e(k) \quad (10)$$

where $\gamma > 0$ is a regularization parameter.

Optimization of criterion (9) without constrains (10) leads to expression

$$w(N) = \left(\sum_{k=1}^N \varphi(x(k)) \varphi^T(x(k)) + \gamma^{-1} I \right)^{-1} \sum_{k=1}^N \varphi(x(k)) y(k) \quad (11)$$

(where $I - (nh \times nh)$ identity matrix) that coincides with least squares ridge-estimates (biased estimates).

For taking into account constraints (10), let us introduce the Lagrange function

$$\begin{aligned} L(w, e(k), \lambda(k)) &= E(k) + \sum_{k=1}^N \lambda(k) (y(k) - w^T \varphi(x(k)) - e(k)) = \\ &= \frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{k=1}^N e^2(k) + \sum_{k=1}^N \lambda(k) (y(k) - w^T \varphi(x(k)) - e(k)) \end{aligned} \quad (12)$$

(here $\lambda(k)$ – N indefinite Lagrange multipliers) and system of Karush-Kuhn-Tucker equations

$$\begin{cases} \nabla_w L(w, e(k), \lambda(k)) = w - \sum_{k=1}^N \lambda(k) \varphi(x(k)) = \bar{0}_N, \\ \frac{\partial L(w, e(k), \lambda(k))}{\partial e(k)} = y(k) - w^T \varphi(x(k)) - e(k) = 0 \end{cases} \quad (13)$$

where $\bar{0}_N - (N \times 1)$ is a vector that consists of zero elements.

Solution of system (13) can be written in the form

$$\begin{cases} w(N) = \sum_{k=1}^N \lambda(k) \varphi(x(k)), \\ \lambda(k) = y(k), \\ y(k) = w^T(N) \varphi(x(k)) + e(k) \end{cases} \quad (14)$$

or in the matrix form

$$(\gamma^{-1} I_{NN} + \Omega_{NN}) \begin{pmatrix} \lambda(1) \\ \vdots \\ \lambda(N) \end{pmatrix} = \begin{pmatrix} y(1) \\ \vdots \\ y(N) \end{pmatrix} \quad (15)$$

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

$$(\gamma^{-1}I_{NN} + \Omega_{NN})\Lambda_N = Y_N \quad (16)$$

(here $\Omega_{NN} = \{\Omega_{pq} = \varphi^T(x(p))\varphi(x(q))\}$, $p = 1, 2, \dots, N$; $q = 1, 2, \dots, N$ from it follows

$$\Lambda_N = (\gamma^{-1}I_{NN} + \Omega_{NN})^{-1}Y_N. \quad (17)$$

Then an output signal of wavelet-neuron can be written in the form

$$\hat{y}(x) = \left(\sum_{k=1}^N \lambda(k) \varphi(x(k)) \right)^T \varphi(x). \quad (18)$$

As can be seen from neuromathematical point of view (the neural network learning theory) and theory of support vector machines (empirical risk minimization), the proposed wavelet least squares support vector machine based on wavelet-neuron is more simple in the implementation, has high speed of operation and requires short volume of a training data set.

If the data are fed sequentially, the process of wavelet least squares support vector machine learning should be organized in on-line mode. When a pair $x(N+1), y(N+1)$ is fed to the input of wavelet-neuron, expression (18) can be written in the form

$$\hat{y}(x) = \left(\sum_{k=1}^N \lambda(k) \varphi(x(k)) + \lambda(N+1) \varphi(x(N+1)) \right)^T \varphi(x) \quad (19)$$

or in the matrix form

$$(\gamma^{-1}I_{N+1, N+1} + \Omega_{N+1, N+1}) \begin{pmatrix} \lambda(1) \\ \vdots \\ \lambda(N) \\ \text{-----} \\ \lambda(N+1) \end{pmatrix} = \begin{pmatrix} Y(1) \\ \vdots \\ Y(N) \\ \text{-----} \\ Y(N+1) \end{pmatrix} \quad (20)$$

or

$$\begin{pmatrix} \Omega_{NN} & | & \omega_{N+1} \\ \text{-----} & & \text{-----} \\ \omega_{N+1}^T & | & \gamma^{-1} \end{pmatrix} \begin{pmatrix} \Lambda_N \\ \text{-----} \\ \lambda(N+1) \end{pmatrix} = \begin{pmatrix} Y_N \\ \text{-----} \\ y(N+1) \end{pmatrix} \quad (21)$$

where

$$\omega_{N+1} = (\mu^T(x(1))\mu(x(N+1)), \mu^T(x(2))\mu(x(N+1)), \dots, \mu^T(x(N))\mu^T(x(N+1)))^T.$$

Using (21) we can write

$$\Lambda_{N+1} = \begin{pmatrix} \Lambda_N \\ \text{-----} \\ \lambda(N+1) \end{pmatrix} = \begin{pmatrix} \Omega_{NN} & | & \omega_{N+1} \\ \text{-----} & & \text{-----} \\ \omega_{N+1}^T & | & \gamma^{-1} \end{pmatrix}^{-1} \begin{pmatrix} Y_N \\ \text{-----} \\ y(N+1) \end{pmatrix}. \quad (22)$$

Using the Frobenius formula for a block matrix in the form [25]

$$M = \begin{pmatrix} A & | & B \\ \text{-----} & & \text{-----} \\ C & | & D \end{pmatrix}, |D| \neq 0,$$

$$M^{-1} = \begin{pmatrix} A & | & B \\ \text{-----} & & \text{-----} \\ C & | & D \end{pmatrix}^{-1} = \begin{pmatrix} K^{-1} & | & -K^{-1}BD^{-1} \\ \text{-----} & & \text{-----} \\ -D^{-1}CK^{-1} & | & D^{-1} + D^{-1}CK^{-1}BD^{-1} \end{pmatrix},$$

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

we can write

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

Further, it is easy to compute $(N+1)$ -th Lagrange multiplier using an expression in the form

$$\lambda(N+1) = -\gamma\omega_{N+1}^T K^{-1}Y_N + \gamma(1 + \gamma\omega_{N+1}^T K^{-1}\omega_{N+1})y(N+1). \quad (23)$$

Further, using the Sherman-Morrisson formula for matrix inversion we can rewrite the learning algorithm in the final form

$$\begin{cases} K^{-1} = \Omega_{NN}^{-1} + \frac{\Omega_{NN}^{-1}\omega_{N+1}\omega_{N+1}^T\Omega_{NN}^{-1}}{1 - \omega_{N+1}^T\Omega_{NN}^{-1}\omega_{N+1}}, \\ \lambda(N+1) = 1 + \gamma\omega_{N+1}^T K^{-1}(\omega_{N+1} - Y_N). \end{cases} \quad (24)$$

V. RESULT OF SIMULATION

To demonstrate the effectiveness of the proposed wavelet least squares support vector machine, the practical problem of time series forecasting, which describes the average monthly temperature in Kharkiv, Ukraine, was solved [26]. Time series consisted of 24 points and, thus, a training sample contained 16 points and 8 points were taken as a testing sample.

Values $x(k-2), x(k-1), x(k)$ were taken as prehistory for the forecasting $x(k+1)$ value. Initial value of shape parameter of the adaptive wavelet activation function was taken $\alpha = 1$. As forecasting quality criterion we used a mean squared error (MSE).

Fig. 3 shows the results of time series forecasting based on wavelet-neuron with different learning algorithms. As can be seen in Fig. 3a, the curves of the actual values (a dashed curve) and forecast ones (a solid curve) are close enough. Fig. 3b shows the results of forecasting using the wavelet-neuron and gradient learning method with a constant step, and Fig. 3c shows the results of wavelet-neuron and the proposed learning algorithm (4) and (5).

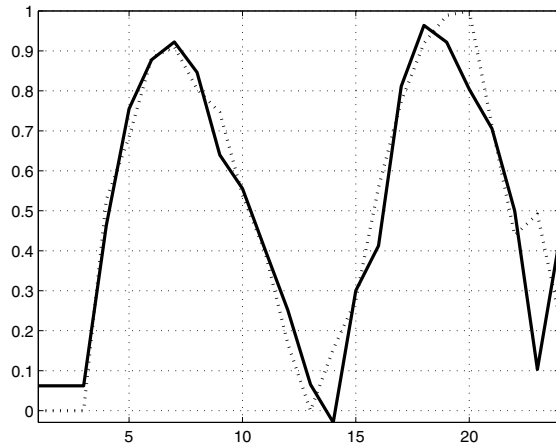


Fig. 3. (a) Result of wavelet-neuron with the proposed learning algorithm based on SVM-criterion.

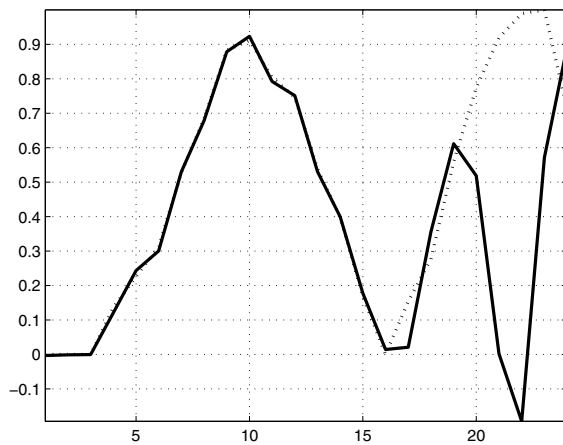


Fig. 3. (b) Results of wavelet-neuron with a gradient learning algorithm with a constant step.

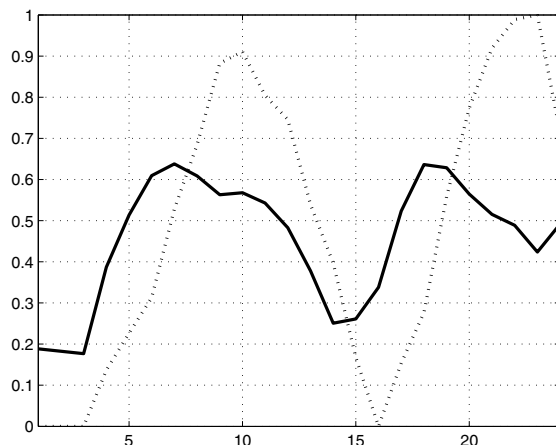


Fig. 3. (c) Results of wavelet-neuron with a learning algorithm of all-parameters (4), (5).

Table I provides the comparison of results of time series forecasting using the wavelet-neuron with different approaches to learning.

TABLE I
THE RESULT COMPARISON OF TIME SERIES FORECASTING

Neural network / Learning algorithm	Num. of input / Num. of activation function	MSEtrn	MSEchk
Wavelet-neuron / Proposed learning algorithm based on SVM-criterion	3/8 (on-line)	0.0063	0.0306
Wavelet-neuron/ Gradient learning algorithm with a constant step	3/8 (10 epoch)	0.000093	0.3186
Wavelet-neuron/ Proposed learning algorithm of all-parameters (4), (5)	3/8 (10 epoch)	0.0374	0.0572

As it can be seen, the wavelet neuron with a gradient learning algorithm shows the best result on a training set, but such a system has the worst prediction abilities. The wavelet-neuron with learning algorithm (4), (5) is not able to train all parameters of the system because of a small data set. Thus, as can be seen from experimental results, the proposed approach provides the best quality of forecasting in comparison with similar approaches due to a special learning algorithm that is able to process information in both off-line and on-line modes.

VI. CONCLUSION

The wavelet least squares support vector machine based on wavelet-neuron was introduced and investigated. The proposed wavelet least squares support machine has such advantages as computational simplicity due to the wavelet-neuron architecture, small number of tuning parameters, high speed operation thanks to the use of the second order learning algorithms and the possibility of on-line information processing.

REFERENCES

- [1] K.-L. Du and M. N. S. Swamy, *Neural Networks and Statistical Learning* London: Springer-Verlag, 2014. <http://dx.doi.org/10.1007/978-1-4471-5571-3>
- [2] S. Haykin, *Neural Networks. A Comprehensive Foundation* Upper Saddle River, N. J.: Prentice Hall, 1999.
- [3] J.-S. R. Jang, "ANFIS: Adaptive-network-based fuzzy inference systems," *IEEE Trans. on Syst. Man. and Cybern.*, vol. 23, issue 3, pp. 665–685, 1993. <http://dx.doi.org/10.1109/21.256541>
- [4] J.-S. R. Jang, C. T. Sun, and E. Mizutani, *Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*, N. J.: Prentice Hall, 1997.
- [5] O. Nelles, *Nonlinear System Identification*, Berlin: Springer, 2001. <http://dx.doi.org/10.1007/978-3-662-04323-3>
- [6] E. Uchino and T. Yamakawa, "Soft computing based signal prediction, restoration and filtering," in *Intelligent Hybrid Systems: Fuzzy Logic, Neural Networks and Genetic Algorithms*, Da Ruan Eds., Boston: Kluwer Academic Publisher, 1997, pp. 331–349.
- [7] Ye. Bodyanskiy and O. Vynokurova, "Hybrid adaptive wavelet-neuro-fuzzy system for chaotic time series identification," *Information Science*, n. 220, pp.170–179, 2013.
- [8] L. Ljung, *System Identification: Theory for the User*, N.Y.: Prentice-Hall, 1999.
- [9] C. Cortes and V. Vapnik, "Support vector networks," *Machine Learning*, n. 20, pp.273–297, 1995. <http://dx.doi.org/10.1007/BF00994018>
- [10] V.N. Vapnik and A.Ya. Chervonenkis, *Pattern Recognition Theory (Statistical Learning Problems)*, M.: Nauka, 1974. (in Russian)

- [11] V. N. Vapnik and A. Ya. Chervonenkis, *Empirical Data Dependencies Restoration*, M.: Nauka, 1979. (in Russian)
- [12] V. N. Vapnik, *The Nature of Statistical Learning Theory*, N. Y.: Springer, 1995. <http://dx.doi.org/10.1007/978-1-4757-2440-0>
- [13] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines*, Singapore: World Scientific, 2002.
- [14] S. Abe and D. Tsujinishi, "Fuzzy Least Squares Support Vector Machines for multiclass problems," *Neural Networks*, n. 16, pp. 785–792, 2003.
- [15] Ch.-F. Lin and Sh.-D. Wang, "Fuzzy Support Vector Machines," *IEEE Trans. on Neural Networks*, n. 13, pp. 646–671, 2002.
- [16] S.M. Pandhiani and A.B. Shabri, "Time Series Forecasting Using Wavelet-Least Squares Support Vector Machines and Wavelet Regression Models for Monthly Stream Flow Data," *Open Journal of Statistics*, n. 3, pp. 183–194, 2013.
- [17] D. Kumar, R. K. Tripathy and A. Acharya, "Least squares support vector machine based multiclass classification of EEG signals," *WSEAS Transactions on Signal Processing*, vol. 10, pp. 86–94, 2014.
- [18] D. Zahiriak, R. Chapman, S. K. Rogers, B. W. Suter, M. Kabrisky and V. Pyati, "Pattern recognition using radial basis function network," in *Application of Artificial Intelligence Conf.*, Dayton, OH, 1990, pp. 249–260.
- [19] T. Yamakawa, "A novel nonlinear synapse neuron model guaranteeing a global minimum – Wavelet neuron," in *Proc. 28th IEEE Int. Symp. On Multiple-Valued Logic*, Fukuoka, Japan, IEEE Comp. Soc., 1998, pp. 335–336.
- [20] Ye. Bodyanskiy, O. Vynokurova, and O. Kharchenko, "Hybrid cascade neural network based on wavelet-neuron," *Information Theories and Application*, vol. 18, no. 4, pp. 335–343, 2011.
- [21] Ye. Bodyanskiy, N. Lamonova, I. Pliss, and O. Vynokurova, "An adaptive learning algorithm for a wavelet neural network," *Expert Systems*, vol. 22, no. 5, pp. 235–240, 2005. <http://dx.doi.org/10.1111/j.1468-0394.2005.00314.x>
- [22] Ye. Bodyanskiy, I. Pliss, and O. Vynokurova, "Flexible neo-fuzzy neuron and neuro-fuzzy network for monitoring of time series properties," *Scientific J. of Riga Technical University. Information Technology and Management Science*, vol. 16, pp. 47–52, 2013.
- [23] Ye. Bodyanskiy, O. Vynokurova, E. Yegorova, "Radial-basis-fuzzy-wavelet-neural network with adaptive activation-membership function" *Int. J. on Artificial Intelligence and Machine Learning*, no. 8 (II), pp. 9–15, 2008.
- [24] G. C. Goodwin, P. J. Ramadge, and R. E. Caines, "A globally convergent adaptive predictor," *Automatica*, vol. 17, no. 1, pp. 135–140, 1981. [http://dx.doi.org/10.1016/0005-1098\(81\)90089-3](http://dx.doi.org/10.1016/0005-1098(81)90089-3)
- [25] F. R. Gantmacher, *The Theory of Matrices* AMS, Chelsea Publishing: Reprinted by American Mathematical Society, 2000.
- [26] Data Market – the open portal to thousands of datasets [Online]. Available: [http://datamarket.com/en/data/set/1loo/#ds=1loo!1n6s=2qi.2ql.2qn&display=line&title=Average+monthly+temperatures+across+the+world+\(1701-2011\)&s=8gd](http://datamarket.com/en/data/set/1loo/#ds=1loo!1n6s=2qi.2ql.2qn&display=line&title=Average+monthly+temperatures+across+the+world+(1701-2011)&s=8gd)

Yevgeniy Bodyanskiy. In 1971 he graduated with honor from Kharkiv National University of Radio Electronics. In 1980 he defended the Ph.D. Thesis. In 1984 he took an academic title of Senior Researcher. In 1990 he defended the Doctor Thesis (Dr.habil.sci.ing.). In 1994 he took an academic title of Professor. Since 1974 he has been working at Kharkiv National University of Radio Electronics. In 1974-1976 he was a Researcher; in the period of 1977-1983, he was a Senior Researcher; in 1986-1991 he was a Scientific Head of Control Systems Research Laboratory; in 1991-1992 he was a Fellow Researcher. Since 1992 he has been a Professor of the Artificial Intelligence Department KhNURE, Scientific Head of Control Systems Research Laboratory KhNURE. He has more than 600 scientific publications, including 40 inventions and 12 monographs. Research interests include hybrid systems of computational intelligence: adaptive, neuro-, wavelet-, neo-fuzzy-, real-time systems, including problems connected with control, identification, forecasting, clustering, diagnostics, fault detection in technical, economical, medical and ecological objects.
E-mail: bodya@kture.kharkov.ua.

Olena Vynokurova. In 2002 she graduated with honor from Kharkiv National University of Radio Electronics. In the period of 2002-2005, she undertook the postgraduate study in the Artificial Intelligence Department. In 2005 she defended the Ph.D. Thesis. In 2007 she took an academic title of Senior Researcher. In 2012 she defended the Doctor Thesis (Dr.habil.sci.ing.). Since 2002 she has been working at Kharkiv National University of Radio Electronics. Since 2014 she has been a Chief Researcher of the Control Systems Research Laboratory and since 2013 she has been a Professor of IT Security Department at Kharkiv National University of Radio Electronics. She has more than 120 scientific publications, including 2 monographs. Research interests include evolving hybrid systems of computational intelligence: wavelet neural networks, hybrid wavelet neuro-fuzzy systems, identification, forecasting, clustering, diagnostics, fault detection in technical, economical, medical and ecological objects.
E-mail: vinokurova@kture.kharkov.ua.

Oleksandra Kharchenko. She is a Master Student at Kharkiv National University of Radio Electronics. Her major field of research is hybrid neuro-fuzzy systems for data mining problems. She has 10 scientific publications. Research interests include evolving hybrid systems of computational intelligence: wavelet neural networks, hybrid wavelet neuro-fuzzy systems, forecasting, clustering in technical, economical, medical and ecological objects.
E-mail: kharchenko.alexandra@gmail.com.

Jevgenijs Bodjanskis, Jeļena Vinokurova, Aleksandra Harčenko. Uz wavelet neironiem balstītā minimālo kvadrātu atbalsta vektoru mašīna

Atbalsta vektoru mašīnas (SVM), kuru arhitektūra sakrīt ar RBFN un GRNN, sinoptiskie svāri tiek noteikti, risinot nelineārās programmēšanas uzdevumu, bet aktivācijas funkciju centri tieši noteikti pēc principa „neironi datu punktus”, kas pēc savas būtības ir dažādu neironu tīklu savdabīgs hibrīds, kuru apmācība bāzējas uz optimizāciju un atmiņu. Neskatoties uz vairākām SVM-tīklu priekšrocībām, no skaitļošanas viedokļa to apmācība ir pietiekami sarežģīts process, jo tas ir saistīts ar lielu dimensiju nelineāro programmēšanu. Tāpēc, kā SVM alternatīva, tika piedāvātas mazāko kvadrātu atbalsta vektoru mašīnas (LS-SVM), kuru apmācība vienkāršojas līdz lineāro vienādojumu sistēmu risināšanai, ko ir daudz vienkāršāk izskaitļot un var realizēt on-line režīmā. Klasiskais SVM wavelet analogs ir atbalsta vektoru wavelet mašīna (WSVM), kurā daudzdimensiju aktivācijas kodolu funkcijas ir aizvietotas ar adaptīvām viendimensijas wavelet funkcijām. Neskatoties uz to, ka WSVM piemīt vairāk iespēju, salīdzinot ar klasiskajām SVM, to aprēķinu apmācība ir saistīta ar pietiekami sarežģītu procedūru realizāciju, kas, loģiski, ierobežo to iespējas pielietošanai reāla laika uzdevumu risināšanai. Sakarā ar to kļūst aktuāla pietiekami vienkāršo wavelet neironu sistēmu izstrāde, kas balstās uz empīriskā riska minimizēšanu un orientējas uz informācijas apstrādi on-line režīmā. Kā šādu sistēmu bāzes elementu mēs pieņemam wavelet neironus, kam raksturīgas augstas aproksimācijas īpašības, vienkāršība, ātra apmācība un iespēja atklāt datus paslēptās saitēs.

Евгений Бодянский, Елена Винокурова, Александра Харченко. Машина опорных векторов наименьших квадратов на основе вэйвлет-нейрона

Своеобразным гибридом различных нейронных сетей, обучение которых основывается как на оптимизации, так и на памяти, являются машины опорных векторов (SVM), архитектура которых совпадает с RBFN и GRNN, синаптические веса определяются в результате решения задачи нелинейного программирования, а центры активационных функций устанавливаются по принципу «нейроны в точках данных». И хотя SVM-сети обладают целым рядом несомненных преимуществ, их обучение с вычислительной точки зрения представляется достаточно трудоемким, поскольку связано с решением задач нелинейного программирования высокой размерности. В связи с этим в качестве альтернативы SVM были предложены машины опорных векторов наименьших квадратов (LS-SVM), обучение которых сводится к решению систем линейных уравнений, что гораздо проще с вычислительной точки зрения и может быть реализовано в онлайн режиме. Вэйвлет-аналогом традиционной SVM является вэйвлет машина опорных векторов (WSVM), в которой многомерные ядерные функции активации заменены одномерными адаптивными вэйвлет-функциями. И хотя WSVM обладает большими возможностями по сравнению с традиционными SVM, их обучение с вычислительной точки зрения связано с реализацией достаточно сложных процедур, что, естественно, ограничивает их возможности для решения задач реального времени. В связи с этим представляется целесообразной разработка достаточно простых вэйвлет-нейро-систем, реализующих идеи обучения, основанного на минимизации эмпирического риска и ориентированных на обработку информации в online-режиме. В качестве базового элемента таких систем нами принят вэйвлет-нейрон, характеризующийся высокими аппроксимирующими свойствами, простотой, высокой скоростью обучения и возможностью выявлять скрытые зависимости в обрабатываемых данных.

Methods of Forecasting Based on Artificial Neural Networks

Arthur Stephenko¹, Arkady Borisov²,¹ *Ventspils University College*,² *Riga Technical University*

Abstract – This article presents an overview of artificial neural network (ANN) applications in forecasting and possible forecasting accuracy improvements. Artificial neural networks are computational models and universal approximators, which can be applied to the time series forecasting with a high accuracy. A great rise in research activities was observed in using artificial neural networks for forecasting. This paper examines multi-layer perceptrons (MLPs) – back-propagation neural network (BPNN), Elman recurrent neural network (ERNN), grey relational artificial neural network (GRANN) and hybrid systems – models that fuse artificial neural network with wavelets and autoregressive integrated moving average (ARIMA).

Keywords – ARIMA ANN, forecasting, GRANN_ARIMA, WANN.

I. INTRODUCTION

Artificial neural networks (ANNs) are a form of artificial intelligence, which attempts to mimic the function of real neurons found in the human brain [2]. ANNs are one of the most accurate and widely used forecasting models that are used in forecasting social, economic, business, engineering, foreign exchange, stock problems and others. Structure of artificial neural networks makes them valuable for a forecasting task with good accuracy.

As opposed to the traditional model-based empirical and statistical methods such as regression and Box-Jenkins approaches, which need prior knowledge about the nature of the relationships between the data, artificial neural networks are self-adaptive methods that learn from data, and only few a priori assumptions about data are needed [1].

Neural networks learn from examples and can find functional relationships among the data even if relationships are unknown or the physical meaning is the baffling [2]. Therefore, ANNs are well suited for problems, whose solutions require knowledge that is difficult to specify but for which there are enough data or observations.

Artificial neural networks can generalize [8]. After learning from the input data (a sample or pattern), ANNs can often correctly process the early unseen sample even if the sample data are noisy. Neural networks are less sensitive to error term assumptions and they can tolerate noise, chaotic components better than most other methods. Artificial neural networks are also universal function approximators. It was proved that a neural network can approximate any continuous function with any accuracy [1].

II. TYPICAL STRUCTURE OF ANN

ANN structure includes input data and artificial neurons that are known as “processing elements”, “nodes” or “units” [10]. The multilayer perceptron includes an input layer, an output layer and one or more intermediate layers called hidden layers. The size and nature of the data set affect the number of hidden layers and neurons within each layer. ANNs with one or two hidden layers perform better than neural networks with the large number of hidden layers.

A. The Propagation of Information in MLPs

The propagation of information in MLPs starts when the input data are taken into the input layer. The inputs are weighted and passed to each node in the next layer. Each processing element in a specific layer is fully or partially connected to many other processing elements using weighted connections [2].

For a time series forecasting problem, a training pattern consists of history data with a fixed number of observations. If we have N observations $y_1, y_2, \dots, y_N$ in the input data set, then using an ANN with n input nodes, we have $N-n$ training patterns than can be used for short-term forecasting – one value ahead. The first training pattern will contain $y_1, y_2, \dots, y_n$ as inputs and y_{n+1} as the output. The second training pattern will contain $y_2, y_3, \dots, y_{n+1}$ as inputs and y_{n+2} as the output. The last training pattern will contain $y_{N-n}, y_{N-n+1}, \dots, y_{N-1}$ inputs and y_N as the output. Then pattern $y_{N-n+1}, y_{N-n+2}, \dots, y_N$ will be used to get forecasting value y_{N+1} . The ANN performs the following unknown function mapping

$$y_{t+1} = f(y_t, y_{t-1}, \dots, y_{t-p}), \quad (1)$$

where y_t is the observation at time t [1].

The scalar weights along with the network architecture store the knowledge of a trained network and determine the strength of the connections between interconnected neurons [7]. If a weight value is zero, then there is no connection between two neurons and if a weight value is negative then a relationship between two neurons is prohibitive. An individual processing element receives weighted inputs from previous layers, which are summed in each node using a combination function, and a bias neuron or threshold is added or subtracted. A bias neuron is connected to every hidden or output unit and a bias neuron value is one.

Neural networks are similar to linear and non-linear least squares regression, and the bias neuron serves a similar purpose as the intercept in regression models. The bias unit is used to scale the input to a useful range $[0, 1]$ or $[-1, 1]$ to improve the convergence properties of the neural network.

The result of this combined summation is passed through a transfer function to produce the nodal output of the processing element (Fig. 1), which is weighted and passed to the processing element in the next layer [2]. The combination function and transfer function together constitute the activation function. In the majority of cases, input layer neurons do not have an activation function, as their role is to transfer the inputs to the hidden layer. The most widely used activation function for the output layer is the linear function as a non-linear activation function may introduce distortion to the predicated output. The sigmoid (logistic), exponential (hyperbolic) tangent, quadratic or linear functions are often used as the hidden layer transfer function. The relationship between the output – predicted value (y_t) – and the inputs – past observations of the time series ($y_{t-1}, \dots, y_{t-p}$) – is given by [5].

$$y_t = w_0 + \sum_{j=1}^q w_j f \left(w_{0,j} + \sum_{i=1}^p w_{i,j} y_{t-i} \right) + \varepsilon_t, \quad (2)$$

where w_j are output layer weights, $w_{i,j}$ are input layer weights, f is a transfer function, q is the number of hidden nodes, p is the number of input nodes, ε_t is a random error at time t .

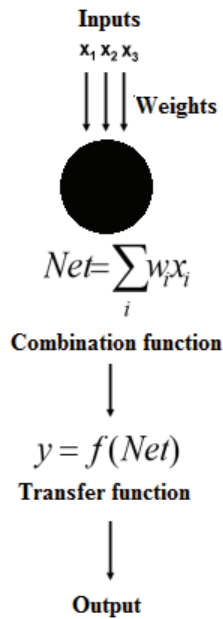


Fig. 1. The propagation of information.

The network corrects its weights and uses a learning rule until it can find a set of weights that will produce the smallest possible error between an observed value and a predicted value at time t . That process is known as “learning” or “training”. For this reason, the network training is actually an unconstrained optimization (nonlinear minimization) problem.

The neural network (2) can approximate any continuous function when the number of hidden nodes q is sufficiently large [5]. In practice, if a network structure has a small number of hidden nodes, then it works well in “out-of-sample” forecasting on data that were not used in training. There is an overfitting effect that can be found in the neural network

modeling process. An overfitted model has a good accuracy on training data, but poor accuracy on “out of the sample” data [9].

To improve the accuracy of the neural network, each data point in the input neurons needs to be normalized – rescaled within the range of $[-1, 1]$ or $[0, 1]$ and standardized to scale data and transformed to make the time series stationary. Transformation can be implemented as taking logarithmic returns of the time series, differencing the time series, etc.

B. Classes of Neural Networks

ANN learning process can be supervised and unsupervised. In supervised learning (e. g., multi-layer feed-forward neural network), the network is presented with historical data, where a training pattern contains independent variables and the corresponding (desired) outputs that are dependent variables in training data. The network then processes the inputs, the nodal output of the network is compared with the observed value of the time series, and an error is calculated. This error is used to correct the connection weights between the model inputs and outputs to reduce the error between the observed values of the time series and outputs predicted by the ANN. The input data used for learning is called the “training set”. Supervised learning is suitable for forecasting and classification tasks. In unsupervised learning (e. g., Kohonen network), there is no dependent variable specified in input (training) data. The network corrects the connection weights according to the input values. The idea of training in unsupervised networks is to cluster the input data into classes of similar features or clusters, where similar input data should generate the same output. This can be referred to as self-organization, and it is suitable for clusterization tasks.

Based on connections between processing elements, ANNs structure can be regarded as feed-forward (e. g., back-propagation) and feedback (e. g., recurrent) networks. Feed-forward network (FNN) propagates information in the forward direction only, while feedback networks propagate information in both the forward and backward directions [2].

Back-propagation neural networks (BPNNs) are a class of feed-forward neural networks with supervised learning rules. The back-propagation network is the most popular and robust multi-layer network that is used in the majority of all forecasting applications. In the learning process, back-propagation neural networks use the gradient-descent search method to correct the connection weights and reduce an error. The main problem of a standard back-propagation algorithm is its slow convergence, which is a typical problem for simple gradient descent methods [1].

Other neural networks that are also used in time series forecasting include recurrent networks, probabilistic networks and fuzzy neural networks. Although feed-forward neural networks are used in many forecasting applications, another type of neural networks – Elman recurrent neural network (ERNN) – is also used in forecasting applications with good accuracy. According to the general principle of the recurrent networks, there is a feedback connection from the outputs of some neurons in the hidden layer to neurons in the context layer that stores the delayed hidden layer outputs. The most important advantage of ERNN is a robust feature extraction ability, when the context layer stores useful information about

data points in past. Since ERNN contains the context layer, it is possible to improve forecasting accuracy by using ERNN instead of FNN [3].

III. HYBRID MODEL: ANN AND WAVELETS

The time series of the real world is often complex in nature and any single forecasting model cannot learn on different patterns equally well. Many studies in time series forecasting have found that forecasting improves in combined models and the integrated forecasting techniques outperform the individual forecasts. In hybrid models, the aim is to reduce the risk that the chosen model will be inappropriate, and combination will help obtain results that are more accurate [5]. Hybrid models can be homogeneous, such as using differently configured neural networks, or heterogeneous, such as with both linear and nonlinear models. Hybrid forecasting has been implemented using a nonlinear model, using ANN with genetic algorithm (GA) or fuzzy logic (FL), or combining linear model and the nonlinear model, using auto-regressive integrated moving average (ARIMA) model with ANN, since in reality time series data typically contain linear and nonlinear patterns [6].

In study [4], the hybrid wavelet and ANN (WANN) model were obtained by combining two methods, discrete wavelet transform and ANN model. ANN model used in this study was the multi-layer feed-forward network.

A. Wavelets

The wavelet transform is a mathematical tool that is used as a time-frequency representation of an analyzed signal. There are some important differences between wavelets and Fourier analysis that is also used as a time-frequency representation of signals. The Fourier coefficients contain only globally averaged information and the Fourier transform does not give local information. Small frequency changes in the Fourier transform will produce changes everywhere in time domain. Wavelets are local in both time and frequency domain. Wavelet transformations provide useful decomposition of the original time series and can get useful information on every decomposition level. Wavelet transforms can be very effective with nonstationary time series data [12]. Many classes of functions can be represented by wavelength in way that is more compact. For a discrete time series $x(t)$ the discrete wavelet transform (DWT) is given by

$$W_{mn} = 2^{-m/2} \sum_{t=0}^{N-1} \psi(2^{-m}t - n)x(t), \quad (3)$$

where W_{mn} is wavelet coefficient for the discrete wavelet, m is integer, which controls the scale, t is time, N is the number of time series data observations, ψ is a transforming function (mother wavelet), n is integer, which controls the time. Therefore, a time series of length N is broken into N components with zero redundancy.

The inverse DWT is given by

$$x(t) = T + \sum_{m=1}^M \sum_{n=0}^{2^{M-m-1}} W_{m,n} 2^{-m/2} \psi(2^{-m}t - n), \quad (4)$$

where, T is the time subseries mean value, M is the number of decomposition level [4].

B. WANN Model

In study [4], the Daubechies wavelet, one of the most widely used wavelet families, was chosen as the wavelet function to decompose the original time series into subtime series components, which were passed to ANN to improve the model accuracy.

To choose the number of decomposition levels, the following formula was used

$$L = \text{int}[\log(N)], \quad (5)$$

where L is the optimal (maximum) number of decomposition levels. In simple format (4) is given by

$$x(t) = A_M(t) + \sum_{m=1}^M D_m(t), \quad (6)$$

where $A_M(M)$ is approximation subseries or residual term at levels M and $D_m(t)$ ($m = 1, 2, \dots, M$) is detailed subseries which can detect small features of interpretational value in the data. The optimal number of decomposition levels for original time series data in this study was three. Original time series was decomposed into three level components ($D1$, $D2$ and $D3$) and approximation component ($A3$). Each of decomposition level component series has a determined role in the original time series and has different effects on the original time series. In this study, the effectiveness of wavelet components was determined using the coefficient of determination (R^2) between each decomposition level component subtime series and original data. The wavelet components $D2$ and $D3$ showed significantly higher R^2 compared to the $D1$ and according to the R^2 analyses, the effective wavelet components were selected as the dominant wavelet components. The combination DW was calculated by

$$DW = A3 + D2 + D3. \quad (7)$$

Fig. 2 shows the structure of the WANN model.

The hybrid model showed a great improvement in time series modeling and produced better forecasts than ANN model alone, as well as GARCH and ARIMA models alone. The study concluded that the forecasting abilities of WANN model were improved when the wavelet transformation technique was adopted for data pre-processing.

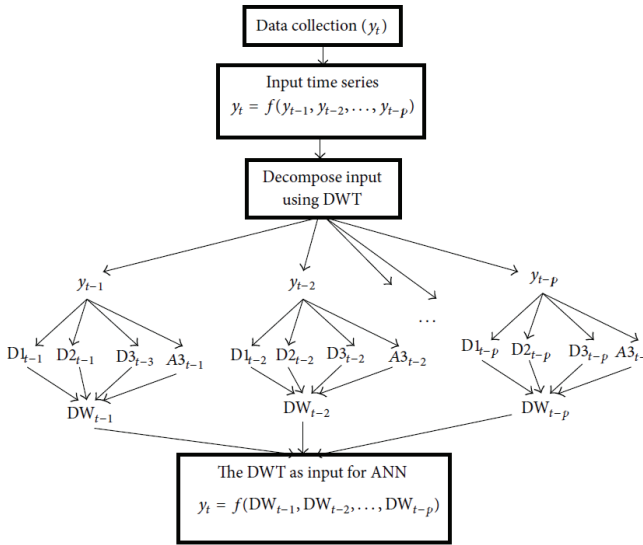


Fig. 2. The structure of the WANN model.

The decomposed periodic components obtained from DWT technique were found to be most effective in giving accurate forecast when they were used as inputs in ANN model. The accurate forecasting results showed that WANN model provides a good and potentially very useful new method for time series forecasting [4].

IV. HYBRID MODEL: ANN AND ARIMA

ANN can model both linear and nonlinear structures in time series; however, they cannot capture both structures equally well. More hybrid forecasting models have been proposed in the last years, using auto-regressive integrated moving average and artificial neural networks. These hybrid models showed good forecasting accuracy [3].

A. Auto-regressive Integrated Moving Average

One of the most widely used linear time series models is the autoregressive integrated moving average (ARIMA) model. In the *ARIMA* (p, d, q) model, the future value of a time series is assumed to be a linear function of several past observations and random errors. In the *ARIMA* (p, d, q) model order p is non-negative integer that refers to the order of the autoregressive function, d is non-negative integer - order of differencing, order q is non-negative integer, which refers to the order of moving average. An auto-regressive integrated moving average process has three different parts – an autoregressive (AR) function that describes how each time series value is a function of the previous p observations, moving average (MA) function describes how each time series value is a function of previous q errors, and an integrated (I) part that describes how to make the data series stationary by differencing d times. The ARIMA model cannot capture nonlinear patterns [11].

Before the ARIMA model can be used for forecasting, check for stationarity is carried out. A stationary time series does not depend on time and, therefore, is characterized by

statistical characteristics such as the mean, variance and autocorrelation structure. When in the observed time series there are trend and heteroscedasticity, then a time series is not stationary. Differencing and power transformation are then applied to the data to remove the trend and to stabilize the variance before an ARIMA model can be used.

If a time series is generated from an ARIMA process, it should have some theoretical autocorrelation properties. By comparing the empirical autocorrelation patterns with the theoretical patterns, it is often possible to identify one or several potential ARIMA models for the given time series. The autocorrelation function (ACF) and the partial autocorrelation function (PACF) of the sample data is often used as the basic tools to identify the order (p, d, q values) of the ARIMA model [5].

B. ANN and ARIMA

In study [5] a time series was considered to be a nonlinear function of several past observations and random errors

$$y_t = f(z_{t-1}, z_{t-2}, \dots, z_{t-p}, e_{t-1}, e_{t-2}, \dots, e_{t-p}), \quad (8)$$

where f is a nonlinear function determined by the neural network, e_t is the residual at time t . Past observation z_t is given by

$$z_t = (1 - B)^d (y_t - \mu), \quad (9)$$

where B is the backward shift (lag) operator, y_t is the original time series value at time t , μ is an ARIMA process generated time series mean value. The residuals are given by

$$e_t = y_t - z_t. \quad (10)$$

Next, a neural network was used to model the nonlinear and linear relationships existing in residuals and original data

$$z_t = w_0 + \sum_{j=1}^q w_j f \left(w_{0,j} + \sum_{i=1}^p w_{i,j} z_{t-i} + \sum_{i=p+1}^{p+q} w_{i,j} e_{t+p-i} \right) + \varepsilon_t, \quad (11)$$

where z_t is a predicted value of the time series by a neural network, f is a transfer function, w_j are output layer weights, $w_{i,j}$ are input layer weights, q is the number of hidden nodes, p is the number of input nodes, and ε_t is a random error at time t . In such hybrids, while the neural network model deals with nonlinearity, the auto-regressive integrated moving average model deals with the non-stationary linear component.

Similar approach was used in study [3]. Here seasonal ARIMA (SARIMA) models were used to analyze the linear part of the time series. Linear part of time series, $\hat{L}_t$, was obtained from the SARIMA model. ANN model used in this study was an Elman recurrent neural network. The ERNN model was developed to fit the residuals obtained from the SARIMA model by

$$e_t = y_t - \hat{L}_t. \quad (12)$$

With p inputs the ERNN model can be written

$$e_t = f(e_{t-1}, e_{t-2}, \dots, e_{t-p}) + \varepsilon_t, \quad (13)$$

where f is a nonlinear function determined by the ERNN. The estimation of e_t in (13) will give the forecasting of nonlinear component of time series, $\hat{N}_t$. Therefore, forecasting values of the time series were obtained by adding the estimates of linear and nonlinear components of the time series

$$\hat{y}_t = \hat{L}_t + \hat{N}_t. \quad (14)$$

It was observed that the following model gave a better result than other methods, such as Zhang's hybrid model of feed-forward neural network and ARIMA and Kajitani self-exciting threshold auto-regressive (SETAR) model.

V. HYBRID MODEL: GRANN AND ARIMA

In study [6] GRANN_ARIMA hybrid model (Fig. 3) was used. The grey relational analysis (GRA) was integrated with ANN to remove the redundancy inputs.

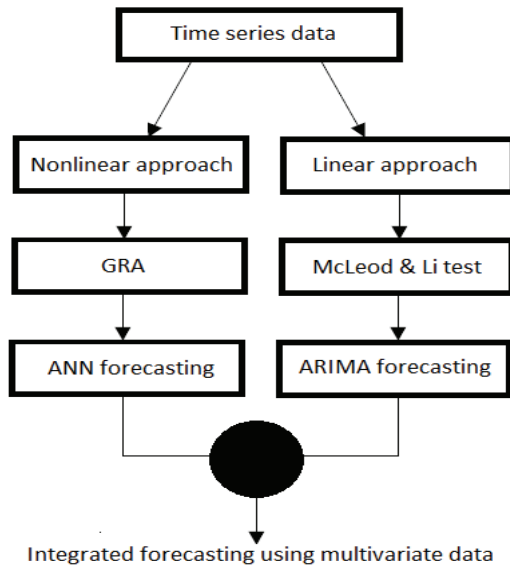


Fig. 3. GRANN_ARIMA model.

GRANN_ARIMA model integrates a nonlinear grey relational artificial neural network (GRANN) and a linear ARIMA model, combining the multivariate time series data and the grey relational analysis to select the appropriate inputs.

A. Grey Relational Analysis

Grey relational analysis is an analysis method that can be used to evaluate the degree of correlation for different data sequences. The degree of correlation between a data sequence (x) and the reference sequence (y) is expressed by a scalar value in interval $[0, 1]$. If the degree of correlation is near 1, it

indicates the high correlation between x and y . There are three main steps in the grey relational analysis. The first step is data pre-processing. Data pre-processing is normally required due to the fact that the range and unit of one data sequence may differ from others. Therefore, data must be normalized, scaled and polarized initially into a comparable sequence before proceeding to other steps. In this study the following equation for data pre-processing was used:

$$x^*(k) = \frac{x_i^0(k) - \min x_i^0(k)}{\max x_i^0(k) - \min x_i^0(k)}, \quad (15)$$

where $i = 1, \dots, m$; $k = 1, \dots, n$, m is the number of experimental data items, n is the number of parameters, $x_i^0(k)$ is the original sequences, $x_i^*(k)$ is the sequences after data pre-processing, $\min x_i^0(k)$ and $\max x_i^0(k)$ are the smallest and the largest value of $x_i^0(k)$, respectively. The range of data is corrected to be in range $[0, 1]$.

The second step is to calculate the grey relational coefficient by using

$$\xi_i(k) = \frac{\Delta \min + \zeta \Delta \max}{\Delta_{0,i}(k) + \zeta \Delta \max}, \quad (16)$$

where $\xi_i(k)$ is a grey relational coefficient at any data point (k), ζ is known as an identification coefficient within interval $[0, 1]$, and normally $\zeta = 0.5$ is used. $\Delta_{0,i}$ are deviation sequences of the reference sequence and comparability sequence

$$\Delta_{0,j} = \|x_0^*(k) - x_j^*(k)\|, \quad (17)$$

where $x_0^*(k)$ is the reference sequence and $x_i^*(k)$ is the comparative sequence. $\Delta \min$ in (16) is given by

$$\Delta \min = \min_{\forall j \in i \forall k} \|x_0^*(k) - x_j^*(k)\|, \quad (18)$$

and $\Delta \max$ in (16) is given by

$$\Delta \max = \max_{\forall j \in i \forall k} \|x_0^*(k) - x_j^*(k)\|. \quad (19)$$

The grey relational grade is the average value of the grey relational coefficients and is defined as

$$\gamma_i = \frac{1}{n} \sum_{k=1}^n \xi_i(k), \quad (20)$$

where n is the number of the reference sequence, $x_0^*(k)$. The grey relational grade γ_i represents the level of correlation between the reference sequence and the comparability sequence.

B. GRANN_ARIMA model

In the GRANN_ARIMA model, ARIMA is used as a linear model, L_t and GRANN is used as a nonlinear model, N_t

$$y_t = L_t + N_t. \quad (21)$$

Here y_t is a value of the original time series. As $\hat{N}_t$ is the forecast value of the GRANN model at time t , then the residuals e_t are obtained by

$$e_t = y_t - \hat{N}_t. \quad (22)$$

The residuals now represent the linear part of the data, and ARIMA can be used to model the residual. Residual modeling by ARIMA can be represented similarly as in (13), but here f is a linear function modeled by the ARIMA model. Therefore, the hybridized forecast model can be written as in (14), but here linear part of model $\hat{L}_t$ is obtained from (13). Unlike model in (8), a nonlinear model is implemented first rather than followed by the linear model.

To validate the performance of the proposed model, Kuala Lumpur Stock Exchange (KLSE) daily close price was used as a time series data. Grey relational analysis was used as a feature selection tool, and four factors were found as the most influential factors relating to KLSE close price: syarian index, trading/service index, composite index and industrial index. A three-layer feed-forward neural network with a single output unit, nine hidden units and four input units with the learning rate 0.5 and momentum 0.9 was used to model nonlinear part of forecast. ARIMA (0,1,3) model was used to model the residuals of close price – linear part of forecast. The network structure and learning parameters were determined by trial and error. The forecasting accuracy was compared with several models, and these include individual models (ARIMA, multiple regression, grey relational artificial neural network), several hybrid models (MARMA, MR ANN, ARIMA ANN), and the artificial neural network trained using the Levenberg Marquardt algorithm. The experiments showed that the GRANN_ARIMA model outperformed other models with MAPE error of 0.16 % and 99.84 % forecasting accuracy. The empirical results obtained showed that the GRANN_ARIMA model could be a good alternative for time series forecasting due to its promising forecasting accuracy.

VI. CONCLUSION

In this paper, an overview of artificial neural network applications in forecasting and possible forecasting accuracy improvements was presented. Artificial neural networks are computational models and universal approximators that can be applied to the time series forecasting with a high accuracy. Back-propagation neural networks are a class of feed-forward neural networks and are most popular and robust multi-layer networks used in the majority of all forecasting applications. Although feed-forward neural networks are used in many forecasting applications of ANNs, other type of neural networks – Elman recurrent neural network (ERNN) – is also used in forecasting applications with good accuracy. In the case of comparison with other types of multi-layer network, the most important advantage of ERNN is robust feature extraction ability. Since ERNN contains the context layer, it is possible to improve forecasting accuracy by using ERNN

instead of FNN. The time series of the real world is often complex in nature and any single forecasting model cannot learn on different patterns equally well. Forecasting improves in combined models and the integrated forecasting techniques outperform the individual forecasts. Hybrid models can be homogeneous, such as using differently configured neural networks, or heterogeneous, such as with both linear and nonlinear models, since in reality time series data typically contain both linear and nonlinear patterns. Hybrid models WANN, ARIMA ANN and GRANN_ARIMA showed good forecasting accuracy and outperformed other forecasting models.

REFERENCES

- [1] G. Zhang, B. E. Patuwo and M. Y. Hu, "Forecasting with artificial neural networks: the state of the art," *International Journal of Forecasting*, vol. 14, no. 1, pp. 35–62, March 1998. [Online]. Available: ResearchGate, <http://www.researchgate.net/>. [Accessed July 15, 2014]. [http://dx.doi.org/10.1016/S0169-2070\(97\)00044-7](http://dx.doi.org/10.1016/S0169-2070(97)00044-7)
- [2] M. A. Shahrin, M. B. Jaksia and H. R. Maier, "State of the art of artificial neural networks in geotechnical engineering," *EJGE Special Volume Bouquet 08*. [Online]. Available: EJGE, <http://www.ejge.com/>. [Accessed July 18, 2014].
- [3] C. H. Aladag, E. Egrioglu and C. Kadilar, "Forecasting nonlinear time series with a hybrid methodology," *Applied Mathematics Letters*, vol. 22, no. 9, pp. 1467–1470, September 2009. [Online]. Available: ScienceDirect, <http://www.sciencedirect.com/>. [Accessed July 20, 2014]. <http://dx.doi.org/10.1016/j.aml.2009.02.006>
- [4] A. Shabri and R. Samsudin, "Daily crude oil price forecasting using hybridizing wavelet and artificial neural network model," *Mathematical Problems in Engineering*, vol. 2014, article ID 201402, July 2014. [Online]. Available: Hindawi, <http://www.hindawi.com/>. [Accessed July 22, 2014]. <http://dx.doi.org/10.1155/2014/201402>
- [5] M. Khashei and M. Bijari, "An artificial neural network (p, d, q) model for timeseries forecasting," *Expert Systems with Applications*, vol. 37, no. 1, pp. 479–489, January 2010. [Online]. Available: Science Direct, <http://www.sciencedirect.com/>. [Accessed July 26, 2014]. <http://dx.doi.org/10.1016/j.eswa.2009.05.044>
- [6] R. Sallehudin, S. M. H. Shamsuddin, S. Z. M. Hashim and A. Abraham, "Hybrid Grey Relational Artificial Neural Network and Auto Regressive Integrated Moving Average Model for Forecasting Time-Series Data," *Applied Artificial Intelligence: An International Journal*, vol. 23, no. 5, pp. 443–486, April 2009. [Online]. Available: ResearchGate, <https://www.researchgate.net/>. [Accessed July 30, 2014].
- [7] M. Adya and F. Collopy, "How effective are Neural Networks at Forecasting and Prediction? A Review and Evaluation," *Journal of Forecasting*, vol. 17, no. 5–6, pp. 481–495, Sep.–Nov. 1998. [Online]. Available: Weatherhead School of Management, <http://weatherhead.case.edu/>. [Accessed July 21, 2014]. [http://dx.doi.org/10.1002/\(SICI\)1099-131X\(199809\)17:5/6<481::AID-FOR709>3.0.CO;2-Q](http://dx.doi.org/10.1002/(SICI)1099-131X(199809)17:5/6<481::AID-FOR709>3.0.CO;2-Q)
- [8] S. T. A. Niaki and S. Hoseinzade, "Forecasting S&P index using artificial neural networks and design of experiments," *Journal of Industrial Engineering International*, vol. 9, no. 1, Dec. 2013. [Online]. Available: Springer, <http://link.springer.com/>. [Accessed July 24, 2014].
- [9] N. Q. Hung, M. S. Babel, S. Weesakul and N. K. Tripathi, "An artificial neural network model for rainfall forecasting in Bangkok, Thailand," *Hydrology and Earth System Sciences*, vol. 13, no. 8, pp. 1413–1425, 2009. [Online]. Available: Hydrology and Earth System Sciences, <http://www.hydrol-earth-syst-sci.net/>. [Accessed July 25, 2014]. <http://dx.doi.org/10.5194/hess-13-1413-2009>
- [10] I. Kasstra and M. Boyd, "Designing a neural network for forecasting financial and economic time series," *Neurocomputing*, vol. 10, no. 3, pp. 215–236, Apr. 1996. [Online]. Available: Geocities, <http://www.geocities.ws/>. [Accessed July 27, 2014]. [http://dx.doi.org/10.1016/0925-2312\(95\)00039-9](http://dx.doi.org/10.1016/0925-2312(95)00039-9)
- [11] G. P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," *Neurocomputing*, vol. 50, pp. 159–175, Jan. 2003. [Online]. Available: Geocities, <http://www.geocities.ws/>. [Accessed July 28, 2014]. [http://dx.doi.org/10.1016/S0925-2312\(01\)00702-0](http://dx.doi.org/10.1016/S0925-2312(01)00702-0)

- [12] L. F. Ortega, "A Neuro-wavelet Method for the Forecasting of Financial Time Series," *Proceedings of the World Congress on Engineering and Computer Science*, vol. 1, San Francisco, USA, Oct. 24–26, 2012. [Online]. Available: International Association of Engineers, <http://www.iaeng.org/>. [Accessed July 29, 2014].

Arthur Stepchenko is the first-year Doctoral Student majoring in Information Technology at Riga Technical University (RTU). He received his Mg.sc.comp. degree in Computer Science from Ventspils University College (VUC) in 2013.

He is a Research Assistant at the Department of Space Technology of the Engineering Research Institute "Ventspils International Radio Astronomy Centre", Ventspils University College (VIRAC).

His research interests include remote sensing, digital image processing, machine learning, data mining and programming.

E-mail: arturs1312@gmail.com.

Arkady Borisov received his Doctoral Degree in Technical Cybernetics from Riga Polytechnic Institute in 1970 and Dr.habil.sc.comp. degree in Technical Cybernetics from Taganrog State Radio Engineering University in 1986.

He is a Professor of Computer Science at the Faculty of Computer Science and Information Technology, Riga Technical University. His research interests include fuzzy sets, fuzzy logic, computational intelligence and bioinformatics. He has 235 publications in the field.

He is a member of IFSA European Fuzzy System Working Group, Russian Fuzzy System and Soft Computing Association, honorary member of the Scientific Board, member of the Scientific Advisory Board of the Fuzzy Initiative Nordrhein-Westfalen (Dortmund, Germany).

Address: 1 Kalku Str., Riga, LV-1658; phone: +371 67089530.

E-mail: arkadijs.borisovs@cs.rtu.lv.

Artūrs Stepčenko, Arkādijs Borisovs. Prognozēšanas metodes, kas balstītas uz mākslīgajiem neironu tīkliem

Mākslīgie neironu tīkli (MNT) ir elastīga skaitļošanas struktūra un universāls aproksimators, kurus var pielietot plaša spektra laika rindu (ekonomikā, biznesā, inženierzinātnēs, valūtu apmaiņā un citās sfērās) prognozēšanai ar augstu precizitāti. Atšķirībā no tradicionālām empiriskām un statistiskām metodēm, kā, piemēram, regresijas analīzes vai Boksa-Dženkinsa pieejām, kur ir nepieciešamas plašas zināšanas par problēmas datu īpašībām, mākslīgie neironu tīkli ir pašmācības metode, kas prasa zināt tikai nelielu informācijas daudzumu par konkrēto problēmu. MNT spēj vispārināt – tie mācās no paraugiem un spēj atrast sakarības starp datiem pat tad, ja šīs sakarības nav zināmas, vai to fiziskā nozīme ir grūti izprotama, kā arī ja ieejas dati satur troksni un haotiskas komponentes. Tāpat MNT var aproksimēt jebkuru nepārtrauktu funkciju ar jebkādu vēlamu precizitāti. Visplašāk laika rindu prognozēšanai tiek lietoti daudzslāņu kļūdu atgriezeniskās izplatīšanas neironu tīkli, kas pieder vienvirziena tīklu klasei, taču ir arī cits neironu tīklu veids – Elmana rekurentais neironu tīkls. Tas ir uzrādījis labus prognozēšanas rezultātus. Daudzslāņu mākslīgais neironu tīkls sastāv no ieejas slāņa, viena vai vairākiem slēptiem slāņiem un izejas slāņa. Uz treniņa datiem apmācīto modeli ar noteiktiem svāriem var pielietot jau nezināmas laika rindas vērtības prognozēšanai testa datos. Tā kā reālās problēmas ir sarežģītas pēc dabas, tad neviens atsevišķs prognozēšanas modelis nespēj mācīties no dažādiem datu paraugiem vienādi labi. Prognozēšanas precizitāte uzlabojas, ja modeļi tiek kombinēti. Prognozēšanas hibrīdmodeļu mērķis ir samazināt risku, kas var rasties, izmantojot nepiemērotu modeli, un paaugstināt prognozēšanas precizitāti. Hibrīdmodeļi var būt homogēni, piemēram, apvienojot dažādu konfigurāciju neironu tīklus, vai heterogēni, apvienojot lineāru un nelineāru modeļi, tādā reālas laika rindas parasti satur gan lineāras, gan nelineāras komponentes. Labu prognozēšanas precizitāti ir sasnieguši modeļi, kas apvieno veivletu analīzi (dekompozīciju) ar MNT, un modeļi, kas apvieno autoregresijas integrēto mainīgo vidējo (ARIMV) ar MNT. Veivletu dekompozīcija tiek izmantota, lai iegūtu laika rindas laika-frekvenču reprezentāciju. Oriģināla laika rinda ar veivletu dekompozīcijas palīdzību var tikt sadalīta vairākās laika apakšrindu komponentēs, un jauna laika rinda, kas sastāv no šīm komponentēm, tiek nosūtīta MNT prognozēšanai. Kā citu hibrīdu veidu var izmantot MNT ar ARIMV. Tā kā ARIMV ir paredzēts lineāru procesu prognozēšanai, tad to var izmantot lineāru prognozes komponentu iegūšanai, bet MNT var izmantot nelineāro prognozes komponentu iegūšanai, un vēlāk šīs komponentes tiek saskaitītas un dod galīgo prognozi.

Артур Степченко, Аркадий Борисов. Методы прогнозирования, основанные на искусственных нейронных сетях

Искусственные нейронные сети (ИНС) являются гибкой вычислительной структурой и универсальным аппроксиматором, они могут быть использованы в прогнозировании временных рядов широкого спектра (в экономике, бизнесе, инженерной науке, обмене валют и других сферах) с высокой точностью. В отличие от традиционных эмпирических и статистических методов, как например, регрессионного анализа или подходов Бокса-Дженкинса, где необходимы широкие знания о свойствах данных проблемы, ИНС являются методом самообучения, которым необходим только небольшой объем данных о проблеме. ИНС способны к обобщению – они обучаются по образцам, и могут найти взаимосвязь между данными, даже если эта взаимосвязь неизвестна или её физический смысл трудно понять, или даже если входные данные содержат шум и хаотические компоненты. ИНС могут аппроксимировать любую непрерывную функцию с любой желаемой точностью. Самый распространённый вид ИНС для прогнозирования временных рядов это многослойная сеть обратного распространения ошибки, которая принадлежит к классу сетей прямого распространения, но и другой вид нейронных сетей – рекуррентная сеть Элмана - показала хорошие результаты прогнозирования. Модель, обученную на тренировочных данных с определёнными весами, можно использовать для прогнозирования неизвестной величины временного ряда в тестовых данных. Так как реальные проблемы по природе являются сложными, то ни одна отдельная модель не может обучаться на разных выборках данных одинаково хорошо. Точность прогнозирования увеличивается, если модели комбинируются. Целью гибридных моделей прогнозирования является уменьшение риска, который может возникнуть при выборе неправильной модели, и повышение точности прогнозирования. Гибридные модели могут быть однородные, например, объединяя ИНС разных конфигураций, или неоднородные, объединяя линейную и нелинейную модель, так как реальные временные ряды содержат и линейную и нелинейную компоненту. Хорошей точности прогнозирования достигли модели, которые объединяют вейвлет-анализ (декомпозицию) с ИНС, и модели, которые объединяют интегрированную модель авторегрессии скользящего среднего (ИМАРСС) с ИНС. Вейвлет-декомпозиция используется, чтобы получить частотно-временное представление временного ряда. Оригинальный временной ряд с помощью вейвлет-декомпозиции может быть разложен на несколько компонент, и новый временной ряд, который состоит из этих компонент, передаётся в ИНС для прогнозирования. Другим видом гибрида может быть ИНС вместе с ИМАРСС. Так как ИМАРСС предназначен для прогнозирования линейных процессов, то с его помощью можно получить линейную компоненту прогноза, а ИНС может использоваться для прогнозирования нелинейной компоненты прогноза, потом эти компоненты складываются и дают окончательный результат.

Towards the Model-driven Software Configuration Management Process

Arturs Bartusevics¹, Leonids Novickis²,^{1,2} *Riga Technical University*

Abstract – Software configuration management is one of the most important disciplines within the software development project, which helps control the software evolution process and allows including into the end product only tested and validated changes. To achieve this, software management completes certain tasks. Concrete tools are used for technical implementation of tasks, such as version control systems, servers of continuous integration, compilers, etc. There are situations when some tools or scripts are implemented to support small particular parts of complex software configuration management process without a general picture of total process. Maintenance of such solutions is complex and more expensive than it could be. The study describes a model-driven configuration management approach, which foresees the development of an abstract environment model for the configuration management process that later is transformed to lower abstraction level models and tools are indicated to support the technical process. This approach allows implementing all tools and solutions according to the planned configuration management process and developing reusable solutions.

Keywords – Configuration management model, model-driven approach, software configuration management.

I. INTRODUCTION

Software Configuration Management is a discipline that controls the evolution of software development process. To achieve this goal, software configuration management deals with the tasks such as configuration item identification, version control, configuration status accounting and configuration audit [1], [3], [4]. Software configuration management is described in ISO quality standards, where it is noted that the results of the process are needed for other sub-processes in software development [3]. General recommendations for configuration management process described in the quality standards are taken as the basis for many information technology companies worldwide. As it is noted in the studies [1] and [4], there are mainly two ways to achieve the standards. The first one is focused on the development of methodology and describes particular activities to be taken to implement configuration management. The second one is focused on the development of tools that most commonly deal with only one configuration management task.

One of the main problems of software configuration management implementation is the fact that implementing strategy of initial phase of software development process is not well thought out. However, a well-thought-out and realized configuration management strategy is the first condition which suggests that configuration management

process has been implemented [1], [3], [4]. Due to the fact that configuration management implementation strategy is not well thought out and there are no specific activities, they start solving real tasks in the project. For example, there is a source code and it is submitted to version control, but there are no baseline strategies for development and maintenance. It is also not thought out how to organize parallel development and maintenance of parallel branches, but in case of project development such a need may arise. This is a situation that develops solutions to specific situations that arise in practice, but the solution does not necessarily correspond to the process of a general concept. Such solutions are difficult to maintain because they are eventually complemented with multiple corrections and adjustments, often resulting in dependency on another task solutions. It is the so-called “master factor” situation, when only one or a few people know how the particular software configuration management solution works, and it takes them a lot of time to find solution maintenance [1].

The second problem is related to the reuse of solutions for configuration management. If configuration management strategy has not been developed and thought out, every project of information technology company needs to develop configuration management solutions from scratch, because previous project solutions were purposed for dealing with too specific tasks and it is difficult to find solutions to concrete tasks there. Sometimes it is easier to implement a solution from scratch than to choose the existing one as basis.

The problems mentioned above additionally stimulate creating many tools that are focused on specific configuration management solution and tool manufacturers promise to solve all potential problems with a particular tool. However, both configuration management experts [1], [4] and quality standards [3] emphasize the need to choose a configuration management plan, which defines specific tasks, and only then to choose the tools that implement these tasks.

Configuration management experts [1], [4], propose at the initial stage to create a model of configuration management process, then to determine activities which are necessary for process implementation, and only then to choose the appropriate tools. Thus, the article first analyzes the existing solutions based on models and their disadvantages. Then it is offered a new model-driven configuration management solution that suggests specific activities for configuration management process plan and its implementation. The final part of the article defines further research directions as well as expected benefits of the designed solution.

II. RELATED RESEARCH

Initially, the model-oriented solutions appeared in the studies, where only one of the main configuration management tasks was solved, for example, configuration item identification [10], [2], [13], [6].

In article [10], the authors propose making developed product modeling. At first, it is emphasized that by defining potential candidates of configuration items, not only a source code should be taken into account, but also the documentation that describes the product from different points of view. Modeling approach is based on a gradual system of tree-like structure development.

While researching version of controlled solutions, one may encounter model creation [2], [5]. Study [2] describes an algorithm, which consists of meta-model controlled items. The meta-model describes the syntax of the model, which could be exposed to version control. The algorithm is intended for version control systems that control models, not a source code, which is specific to projects that use a model-driven development approach. Thanks to meta-model, it is possible not only to manage the change of models, but also to make the merge of different model versions and to see merge conflicts, like it can be seen in source code file change combination. A similar solution is offered in the following source [5]. There is mentioned the fact that the version control and parallel development may not be limited by source code file control. The projects, which use a model-driven approach, need to control the model versions, while the vast majority of existing version control systems fails to maintain versions of the models. Therefore, the authors of study [5] propose an approach that requires the development of an abstract model for version control system.

Model-driven approach is used for building and installation processes in the framework of configuration management [13], [6]. The following article [13] is not about specific tools, but about an abstract model, which describes the building and installation process. The solution offers a model that describes the activities necessary to implement in the building and installation process. However, the following article [6] emphasizes the importance of a correct choice of configuration items for qualitative product building that should be included in the building process. Article [6] proposes a model-based methodology, which provides a set of configuration items to be included in the building process. Methodology uses heuristics and analyzes all possible configuration item sets that could potentially be included in the building process.

There are model-based solutions that are focused not on one, but numerous configuration management tasks [12], [7], [8]. In order to have a better idea about solution maturity and disadvantages in the context of a model-driven approach, general principles of model-driven approaches [11], [9] used in software development have been studied. Model-driven solutions should have the following features:

- The existence of a meta-model. The solution must have the source from where one can create a configuration management process model.
- Different levels of abstraction models. Model-driven solution should have an abstract model that describes a configuration management process, independently from computing and tools that will be used in the process of implementation. However, there should be other types of models, where you can see signs of computing, platform specificity and tool role in the process.
- Relationship between different types of models and transformation rules, which allows the model with one abstraction level to transform into the model of another abstraction level. Transformation rules should be formalized in order to be subjected to computer processing.
- Tool support. In order not to leave the solution at a theoretical level, we need tools that are able to create models from the existing meta-models and to make necessary transformation.
- Relationship with a problematic environment. This relationship is needed for a viable solution. Otherwise, the created and implemented model is unable to respond to changes in the problematic environment and it will be impossible to quickly determine weaknesses in the model. Over time, this can lead to specific tool configuration or code corrections, implementation process and the process model will lose its relevance.

Within the following solution [12], the authors have developed configuration management and model-driven development unification concept, meta-model, which allows creating an abstract model of product configuration, Eclipse Modeling Framework tools, which allow obtaining a specific configuration model from an abstract model platform, as well as instructions for expanding the created methodology and tools. Although the solution corresponds to the main principles of a model-driven concept, it is focused on projects where development occurs based on model-driven approaches. Meta-model is only a configuration item for task identification.

Study [8] presents a methodology, which considers a configuration management process as a whole process. Configuration management principles for solution creation are taken from ITIL standard (Information Technology Infrastructure Library) and later abstract models are created, from which one can create an abstract configuration management process, and later this model can be transformed into a platform interdependent model. The approach uses the main principles of model-driven development. The meta-model based models offer the necessary abstraction that improves the process of configuration management, transparency and allows the user, if necessary, to implement a model of a particular technology, for example, to make model transformation. A system prototype is created that implements model-driven configuration management. The author of article

[8] provides extensive research further, in order to arrive at a particular solution, which is also recognized in the conclusions. The solution is focused only on one technology (JAVA). Thus, it is difficult to assess the possibility of practical application, and an opportunity to realize the relationship with the problem is also not analyzed.

Solution [7] offers an approach to various configuration management tools for mutual integration. In order to maintain a complete configuration management process, a number of tools are required: version control system, problem management system, building servers, continuous integration servers and many other tools. In practice, all the above-mentioned tools work separately from each other. In order to facilitate a configuration management process, an approach has been offered to integrate all these tools together. However, in order to integrate various configuration management tools, it is necessary to define a general concept of each integrated tool [7]. The article offers configuration management task ontology. This ontology is used as a configuration management model that shows how various configuration management tools are integrated. Ontology is mainly based on the version control, which is one of the key concepts of configuration management. Ontology provides information about configuration management of sub-reciprocal links, which are expressed in concepts, links, tasks, agents, output and input data sets [7]. There is no guidance how ontology can be used in specific project configuration management. In terms of relationship between problems, it is not clear how to make changes in ontology, which ontology editors it is recommended to use and how to determine the moment when these changes have to be made.

Solution [14] intends to apply the fuzzy logic theory to create a multi-criteria decision-making system. Decision-making main objective is to determine the optimal set of configuration items that will be subjected to a configuration management process.

III. GENERAL APPROACH TO MODEL-DRIVEN SOFTWARE CONFIGURATION MANAGEMENT

This article proposes a new model-driven configuration management implementation approach, which unlike other model-driven solutions is focused on a gradual process rather than solving a specific task. Another difference is that the solution does not impose any specific technology or tool, but suggests using the existing solutions by increasing the reuse.

During the development of configuration management model-driven approaches, the main ideas of MDA [9], [11] were taken into account. A solution should contain abstraction models with different levels and clearly defined rules how to move from one abstraction level model to another.

The highest level of abstraction MDA is CIM (computation independent model). While researching configuration management definitions, it was searched for something, which could build a configuration management model with the highest level of abstraction.

Summing up definitions [1], [3], [4], the authors managed to come to a conclusion that configuration management in its highest level of abstraction answers the following question: "How to transfer software performed changes from one instance to another at the right moment?" A key highlight of this subject is placed on the changes that were made to the developed product and its ability to transfer changes from one instance to another, because usually several instances are used in a software development project [1].

Thus, creating a configuration management model with the highest level of abstraction, the following aspects should be taken into account: the existing instances, and the way, software development changes will go starting from development and finishing with exploitation, as well as which versions are formed for the developed product. This type of model, developed in the framework of the solution, is called an environment model and will be hereinafter denoted as an EM.

Each model must have the source or meta-model; therefore, in the developed concept the element "Meta-model of EM" has been introduced. At the initial stage, a configuration manager, who is planning the process, is proposed to make an EM using elements from a meta-model.

Expert system knowledge base represents rules, which are able to determine general configuration management risks depending on the state of EM. Such rules hereinafter will be marked as "Rules of Risks, Compilation". The expert system within the developed concept is indicated as "Expert System". If a configuration manager creates an environment model that does not meet basic principles of configuration management [1], [4], an expert system then reports about potential risks of the model. As soon as a configuration management process EM has been obtained and an expert system has identified risks, it is necessary to form a model with a lower level of abstraction. The essence of a model has been defined in the following article [7], which says that regardless of the planned configuration management process, in the end it supports a variety of tools which carry out certain activities to implement source code management, compilation, building activity etc., performs the basic tasks of configuration management. Environment Model shows the further way of software changes. Taking into account the following opinion [7], it would be logical at this stage to understand what kind of action should be taken to implement an EM described way in accordance with principles of configuration management [1], [4]. The following information [11] should also be taken into account: a CIM model is followed by a platform-independent model (PIM), and actions at this stage cannot be attracted to specific technologies and platforms. Thus, the concept has introduced the element "Platform Independent Action Model" that is marked as PIAM. This model has also a specific source or meta-model "Meta-model of PAM".

A new element "E→P" appears in an expert system or the transformation rules from an Environment Model to a Platform Independent Action Model. Thus, obtaining an

Environment Model, a configuration manager uses “E→P” transformation rules to get a PIAM model. PIAM meta-model defines global activities, which deal with configuration management tasks.

Once a configuration management PIAM model has been created, a configuration manager chooses a solution from a “Solutions database” to each action of PIAM. Thus, it becomes known how PIAM model specific items will be implemented in each task. This model is called “Platform Specific Action Model”, and is marked as a PSAM. Fig. 1 shows a general scheme of model-driven configuration management. Arrows with numbers represent steps of approach.

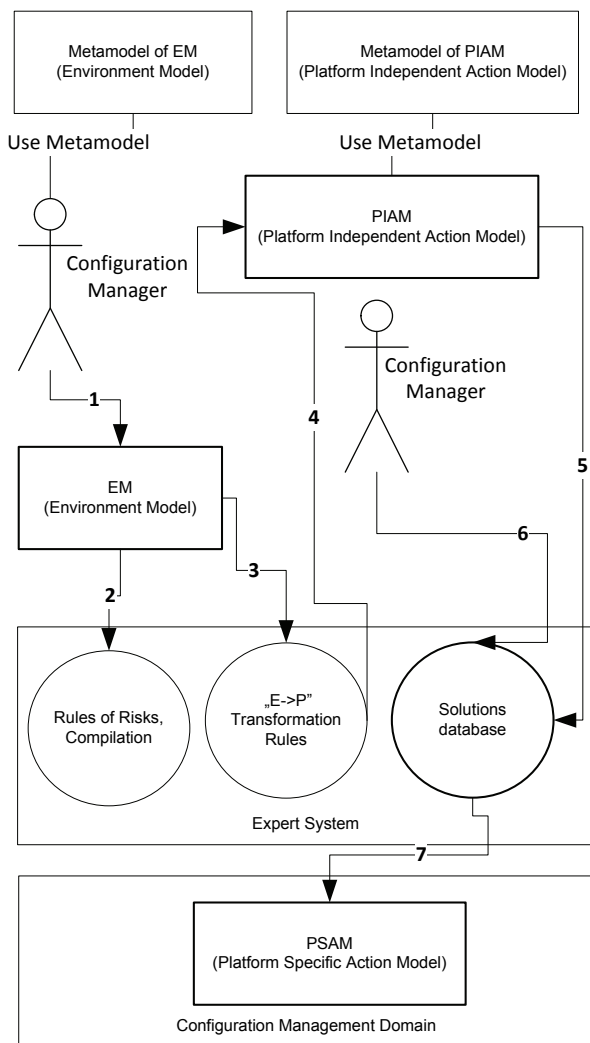


Fig. 1. Conception of model-based configuration management.

A detailed description of elements of the model-driven approach for the implementation of configuration management process is provided below:

EM (Environment Model) – Displays the flow of configuration items in the context of different instances. For example, the development and maintenance of development project displays how changes come from the development environment to different test environments, and finally, to the

operation or production environment. It also shows how new versions of configuration items are created and in what order these versions are transferred from one instance (environment) to another.

Metamodel of EM – Source for an Environment Model. It contains all the components of which EM may consist. During the EM modeling course, the user can take components only from a meta-model.

PIAM – shows which actions are needed to ensure the implementation of the flow of EM model configuration elements through the environments. This model has only abstract actions. For example, there is an element “PREPARE_BASELINE”, but there is no information about platform, technology, manufacturers etc.

Metamodel of PIAM – Source of PIAM model. It contains all the components of which PIAM may consist.

PSAM – PIAM extension model, which eliminates abstraction from technologies, platforms, manufacturers, and etc. All actions have specifically selected solutions from a solution database, so that the model contains certain information necessary for activity implementation.

Configuration Management Domain – an environment, where a configuration management system operates: tools, interaction, project factors that affect the configuration management process etc.

Rules of Risks, Compilation – Rules, which determine problem areas in an EM, depending on the problem characteristics and model component condition. These rules also contain different types of recommendations to prevent one problem or another, as well as how to react to changes in a problematic environment (changing one or more parameters).

E→P – rules that determine how to get a PIAM from an EM.

Solutions database – a database of solutions of actions from a PIAM.

Expert System – an expert system that combines configuration management solution database and transformation rules from an EM to a PIAM, as well as “Rules of Risks, Compilation”.

IV. PRACTICAL IMPLEMENTATION OF A MODEL-DRIVEN APPROACH OF CONFIGURATION MANAGEMENT

EM consists of the following elements:

a) **Environment** – instance that contains configuration elements that make up the software product,

b) **ConfigurationItemFlow** – a flow, through which configuration items are transferred from one environment to another,

c) **Event** – consists of one or more configuration flows. Event may contain several flows, because sometimes, before transferring configuration to some environment, at first it is transferred to the appropriate environment copy to make sure that transferring will occur correctly. For example, in order not to jeopardize the test environment, at first configuration may be transferred to the test environment copy and then to the real

test environment. In this case, one event will be responsible for transferring changes from the development environment to the test environment, and this event includes two configuration flows: one flow transfers the changes to the copy of the test environment, and in case of success, the second flow is triggered, which transfers the same configuration to the real test environment.

Once an EM has been created and all the environments have been identified, all the events and flows can create a PIAM, which at the highest level includes the element “Continuous Integration Server” – a centralized location, from where all the events in configuration transfer between environments will be managed. Continuous Integration Server will contain all the events of EM with all relevant configuration flows. Then “E→P” transformation rules will be used, which are based on configuration management practices, standards or frameworks [1], [3], [4], and for each configuration flow activities from the following action set will be determined:

DEVELOPMENT – a process of development is included in a PIAM model, as configuration management controls what kind of development regulations should exist that control principles of configuration item development and changes [1], [3]. By simulating these activities, it is possible to determine, for example, certain technology and principles of configuration item development of a certain project, procedure that requires changes will be agreed, tools for performing change management, regulation documents, instructions, etc. It should always be kept in mind that if the development process is not controlled and managed, further steps in the configuration management process will be difficult to predict [1].

COMMIT_CHANGES – an activity that determines preservation of change development in a version control system. Activity is controlled by a version control task. All developed product configuration items should be subjected to version control, in order to be able to always see when, who and why made changes [1], [3], [4]. This activity is modeling a version control of a certain project and specific rules to be followed by developers, saving changes in a version control system. Moreover, tools that automatically monitor compliance with required rules and principles can be recorded into activity attribute.

PREPARE_BASELINE – this activity, depending on a version control system, controls the procedure and technical support of environmental baseline preparation. Configuration management controls how to define a baseline to any changes, against which the changes will be made [1]. Environment model assumes that each environment has its own baseline – an approved condition of product source code that corresponds to environment configuration. Thus, before any configuration transmission from one environment to another, it is necessary at first to rebuild a source code of the relevant environment. In practice, this activity provides the transfer of changes between two branches of version control system, which is also called the merge. This activity describes the process of configuration

transfer between two repository branches and provides a description of tools that support it technically.

COMPILE_BUILD – an activity that builds a product from a corresponding source code. Compilation and the build should result in the creation of an executable item.

INSTALL_BUILD – an activity that installs the built product into a certain environment (instance).

PRODUCT_DELIVERY – an activity that prepares and ships the built product to a customer. This activity is necessary in order to be able to install the product into the environment, which is supported by a customer, and a developer team does not have access, for example, to a production environment. Activity is modeling a procedure, which is necessary to follow while preparing the product for shipment. For example, together with the built product it is necessary to send the description of the version and instructions how to install the product into instance; these requirements control configuration management and quality standards or frameworks [1], [3], [4].

ENV_UPDATE_NOTIFICATION – an activity that sends a signal to the development team that environment is updated by new version of product. At this moment, the development team should record the fact, and accordingly supplement configuration item information in a version control system, as well as install the same version of the product into all environments, which is a test copy of the customer’s environment. This process needs an obligatory renewal of the appropriate environmental baseline. This action controls the procedure that is mandatory to comply by receiving information from a customer about any environment restoring.

Each PIAM configuration flow for each activity contains the following attributes, which will not be completed, because at this stage a specific solution to any activity has not been chosen yet:

Platform – a platform specific name, which is necessary for activity implementation. This attribute is necessary because one activity may have multiple solutions across multiple platforms, for example, a script which transmits changes between two repository branches can be implemented in both Linux and Windows platforms.

SolutionName – a unique solution name to differentiate and manage it.

NeededTools – tools needed for technical implementation of the solution.

LocationsOfSolutions – this attribute is intended to store any information about the existing solutions for a certain action. There, for example, the script names, location of special programs, instruction etc. can be stored.

Description – additional description for a particular action in the context of the project or solution that could provide additional information to a configuration manager.

Once a PIAM is ready, a configuration manager chooses a solution to each configuration management activity from “Solution Database”. If an appropriate solution is missing, then it should be developed by putting information into a solution database and only then the solution should be

functioning. As soon as all the activity attributes are filled, thanks to a solution database, a PIAM becomes a PSAM, which is an extended version of platform-independent action model. In order to begin process model functioning in the problem environment, it remains to realize the selected solutions for each activity. Fig. 2 shows a visual example of EM, PIAM and PSAM.

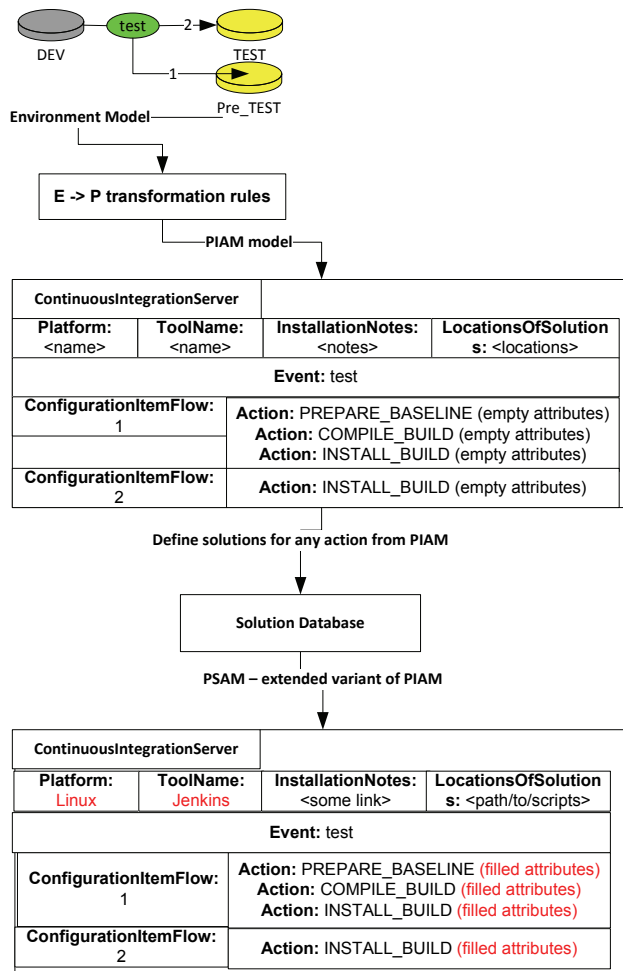


Fig. 2. Example of configuration management models.

In the example, shown in Fig. 2, an environment model is implemented with three environments: DEV, TEST and Pre_TEST. The task is to transfer configuration from a DEV to TEST environment, and it is the responsibility of the Event “test”. ConfigurationItemFlow “1” restores a source code in a version control system that meets a TEST environment, compiles the restored source code, and installs it into a Pre_TEST environment. If the process has been successful, the ConfigurationItemFlow “2” is implemented, which installs into a TEST environment the same installation package, which comes from the first flow COMPILE_BUILD action. Thus, the Event “test” has met the main goal and has transferred configuration from a DEV to TEST instance. To start PSAM functioning in a problem environment, it remains to implement the selected solution activities

PREPARE_BASELINE, COMPILE_BUILD and INSTALL_BUILD.

V. RESULTS, FURTHER STUDIES AND EXPECTED BENEFITS

Based on studies of configuration management solutions and solution improvement tendency, a new model-driven approach has been created for implementation of configuration management. Within the framework of the research, meta-models for EM and PIAM models for a new approach have been developed.

Further studies should develop the prototype of proposed approaches to automate the modeling and transformation process. The results of experiments should determine the weaknesses of the current solutions, as well as develop criteria, which will help to determine whether the solution meets quality standards.

The main expected benefit of the proposed solution is the opportunity to see common configuration management processes and to track compliance with the technical implementation. Thanks to a PSAM that provides information on what tools are needed and how they exchange information, there is an opportunity to automate the process. Thanks to the model, there is an opportunity to configure tools in compliance with specific process goals, rather than to correct certain errors or solve only one of the tasks. This allows spending resources for tool configuration more efficiently, preventing not a specific error, but its cause, as well as it minimizes the risk of the error occurrence in the future. Another benefit could be the process transparency. If a new person joins the project, he has a chance to get acquainted with a configuration management process in a whole and understand how and for what purposes the tools operate. Process transparency will allow if necessary making changes in tool configuration more effectively. The main goal of the solution is to improve the quality of configuration management process, its transparency, as well as increase the efficiency of tool configuration and reduce the maintenance costs, so that there will be more opportunities to eliminate the causes of error, but not the error itself.

REFERENCES

- [1] R. Aiello, *Configuration Management Best Practices: Practical Methods that Work in the Real World*, 1st ed., Addison-Wesley, 2010.
- [2] K. Altmanninger, “Models in conflict – towards a semantically enhanced version control system for models,” *Lecture Notes in Computer Science, LNCS*, 2008; vol. 5002. pp. 293–304. http://dx.doi.org/10.1007/978-3-540-69073-3_31
- [3] R. Bamford, J. Deibler, *Configuration Management and ISO 9001: SSQC*, 1999.
- [4] A. Berczuk, *Software Configuration Management Patterns: Effective TeamWork, Practical Integration*, 1st ed., Addison-Wesley, 2003.
- [5] T. Buchmann, A. Dotor, B. Westföchtel, “Model-Driven Development of Software Configuration Management Systems,” *4th International Conference on Software and Data Technologies, ICSOFT*, 2009, pp. 309–316.
- [6] O. Bushehrian, “Automatic object deployment for software performance enhancement,” *The Institution of Engineering and Technology*, vol. 5, Iss. 4, pp. 375–384, 2011.
- [7] R. Calhau, R. Falbo, “A Configuration Management Task Ontology for Semantic Integration,” *Proceedings of the 27th Annual ACM*

- Symposium on Applied Computing, ACM, New York, NY, USA, 2012, pp. 348–353. <http://dx.doi.org/10.1145/2245276.2245344>
- [8] H. Giese, A. Seibel, T. Vogel, A Model-Driven Configuration Management System for Advanced IT Service Management, 2009. [Online]. Available: http://www.hpi.unipotsdam.de/giese/gforge/publications/pdf/GSV-MRT09_paper_7.pdf [Accessed: Sept. 15, 2014].
- [9] O. Nikiforova, N. Pavlova, K. Gusarova, O. Gorbiks, J. Vorotilovs, A. Zaharovs, D. Umanovskis, J. Sejans, "Development of the Tool for Transformation of The Two-Hemisphere Model to The UML Class Diagram: Technical Solutions and Lessons Learned," *Proceedings of the 5th International Scientific Conference „Applied Information and Communication Technologies”*, 2012, Jelgava, Latvia, pp. 11–19.
- [10] Object-Oriented Software Engineering Using UML, Patterns and JAVA "Software Configuration Management", 2002. [Online]. Available: http://www.bilkent.edu.tr/~bakporay/cs_413/Bruegge_L28_ConfigurationManagement_ch12lect1.ppt [Accessed: Sept. 28, 2014].
- [11] J. Osis, E. Asnina, *Model-Driven Domain Analysis and Software Development: Architectures and Functions*: IGI Global, Hershey – New York, 2011, 514 p. <http://dx.doi.org/10.4018/978-1-61692-874-2>
- [12] W. Pindhofer, "Model Driven Configuration Management," Master thesis, Wien University, Wien, 2009.
- [13] P. N. Sindhuja, G. Surajit, "Software Deployment: Concepts and Technologies," *ICFAI Journal of Systems Management*, 2008.
- [14] J. Wang, Lin Yung-I, "A fuzzy multicriteria group decision making approach to select configuration items for software development," *Mathematics WEB, Fuzzy Sets and Systems*, 2002, pp. 343–363.
- [15] Y. Udovichenko, Upravljenje projektami ili kessonajna boleznj projektov, 2011. [Online]. Available:

<http://experience.openquality.ru/software-configuration-management> [Accessed: Aug. 10, 2014].

Arturs Bartusevičs currently is a Doctoral Student at Riga Technical University, the Faculty of Computer Science and Information Technology, the Institute of Applied Computer Systems. He obtained BSc (2008) and MSc (2011) degrees in Computer Science and Information Technology from Riga Technical University. His research areas are software configuration management, release building and management process and its optimization. He works at Ltd. Tieto Latvia as a Software Configuration Manager. E-mail: arturik16@inbox.lv.

Leonids Novickis, a Head of the Division of Applied Systems Software. He obtained *Dr. sc. ing.* degree in 1980 and *Dr. habil. sc. ing.* degree in 1990 from the Latvian Academy of Sciences. He is the author of 180 publications. Since 1994 he has been regularly involved in different EU-funded projects: AMCAI (INCO COPERNICUS, 1995-1997) – WP leader; DAMAC-HP (INCO2, 1998-2000), BALTPORTS-IT (FP5, 2001-2003), eLOGMAR-M (FP6, 2004-2006) – scientific coordinator; IST4Balt (FP6, 2004-2007), UNITE (FP6, 2006-2008) and BONITA (INTERREG, 2008-2012) – RTU coordinator; LOGIS, LOGIS-Mobile and SocSimNet (Leonardo da Vinci) – partner. He was an independent expert of IST and Research for SMEs in FP6 and FP7. He is a corresponding member of the Latvian Academy of Sciences and an elected expert of the Latvian Council of Science. His research fields include Web-based applied software system development, business process modeling, e-learning and e-logistics. E-mail: lnovickis@gmail.com.

Artūrs Bartusevičs, Leonīds Novickis. Modeļu vadāmais programmatūras konfigurācijas pārvaldības process

Programmatūras konfigurācijas pārvaldība ir viena no svarīgākajām programmatūras izstrādes disciplīnām, kas kontrolē produkta evolūciju un nodrošina to, lai produkta gala versijā būtu tikai kvalitatīvas un notestētas komponentes. Lai to nodrošinātu, konfigurācijas pārvaldība veic vairākus uzdevumus: konfigurācijas vienumu identifikāciju, versiju kontroli, būvējumu un instalāciju pārvaldību, konfigurācijas vienumu uzskaiti. Šo uzdevumu risināšanai tiek izmantoti konkrēti rīki, kas implementē uzdevumus tehniski. Bieži ir situācijas, kad konkrēts rīks ir implementēts, lai risinātu konkrētu mazu problēmu vai uzdevumu, taču pilnīga konfigurācijas pārvaldības procesa pārskats nav. Tādā veidā implementētu risinājumu ir ļoti grūti uzturēt, tas ātri apaug ar labojumiem un pielāgojumiem, un to vairs nav iespējams pielietot citā projektā bez papildu modifikācijām. Līdz ar to sanāk patērēt neracionāli daudz līdzekļu konfigurācijas pārvaldības procesu atbalstam. Raksts piedāvā jaunu modeļvadāmu pieeju konfigurācijas pārvaldības ieviešanai. Jaunās pieejas ietvaros tiek piedāvāts izveidot abstraktu vides modeli un vēlāk to transformēt cita veida modeļos, pakāpeniski ieviešot risinājumu tehniskās detaļas. Jaunā metodoloģija ļauj ieviest konfigurācijas pārvaldības risinājumus atbilstoši plānam, atkārtoti izmantojot jau esošos risinājumus. Atšķirībā no citām līdzīgām metodoloģijām, šī pieeja neuzspiež kādu noteiktu rīku. Tas ļauj reorganizēt esošos risinājumus, palielinot iespēju izmantot tos atkārtoti. Raksta sākumā ir sniegta literatūras analīze, kuras ietvaros esošos modeļvadāmos risinājumus parāda konfigurācijas pārvaldībā. Tiek analizēti esošo risinājumu trūkumi. Turpinājumā jaunā pieeja tiek aprakstīta kopumā, dotas modeļu definīcijas. Raksta nobeigumā tiek dots modeļu vizuālais piemērs, aprakstīti turpmākie pētījumi un jaunā risinājuma sagaidāmie ieguvumi.

Артур Бартусевич, Леонид Новицкий. Процесс управления конфигурациями программного обеспечения, основанный на моделях

Управление конфигурациями программных продуктов является одной из самых важных дисциплин, которая контролирует эволюцию программных продуктов и обеспечивает вхождение только качественных и протестированных составляющих в конечную версию программного продукта. Для того, чтобы обеспечить данное условие, необходимо выполнить ряд задач: идентификация конфигурационных элементов, контроль версиями, управление сборками и инсталляциями продукта, учёт конфигурационных элементов. Для решения подобных задач используется ряд программ и различных самодельных приложений. Возникают ситуации, когда технические решения покрывают лишь узкий круг задач и проблем, в то время как полная картина процесса управления неизвестна. Такие решения являются очень специфическими для конкретных проектов, решения очень тяжело поддерживать, так как они обрастают исправлениями и обновлениями. Подобные решения нельзя использовать повторно в других проектах. Это приводит к тому, что на управление конфигурациями тратится нерационально много средств. Статья предлагает новый подход для внедрения процесса управления конфигурациями, основанного на моделях. Сначала предлагается составить абстрактную модель среды, постепенно трансформируя её в другие модели, поэтапно уточняя технические детали реализации. Новый подход не требует внедрения конкретных программ или технологий и позволяет реорганизовать имеющиеся решения для увеличения их повторного использования. В начале статьи анализируются другие подходы, основанные на моделях, их недостатки. Позже описывается новый подход в целом, даются определения моделям. В конце работы приводится визуальный пример моделей, ожидаемые преимущества, а также делаются выводы.

Modeling-based Risk Analysis of a Small Enterprise

Kristine Bendza¹, Jelena Pecerska^{2, 1} *SIA "JUMS 1", ² Riga Technical University*

Abstract – The paper discusses risk management and analysis of a small enterprise. A particular enterprise-oriented analysis approach is developed and performed using such research methods as descriptive modeling, business process modeling, ABC analysis, risk identification and classification, qualitative and quantitative risk analysis. The customer service level is used as a performance indicator of quality. The paper introduces the statistical approach to the analysis of the dependence of quality from an inventory management strategy under uncertainty of demand in different scenarios. The analysis of simulation results shows how a certain approach can be applied to reveal sensitive spots and gaps in the business process.

Keywords – Business process modeling, risk analysis, risk management, small business.

I. INTRODUCTION

The pressure of globalization and rapid technological advancement of the recent years have made risk management and analysis a focus for most companies. Despite the wide variety of issue-related information sources, there is no unified approach to risk analysis; furthermore, most of the sources tend to consider only large enterprises. The aim of the study is to demonstrate a particular enterprise-oriented modeling-based risk analysis approach of a small enterprise.

Research on the significance of small enterprises revealed that small business has relevant benefits for both the entrepreneur and the country, having a significant impact on the development of Latvian entrepreneurship and its successful recovery from the 2008/2009 economic crisis, being accountable for over 90 % of all economically active enterprises in Latvia in the year 2013, encouraging regional business and employing over 50 % of all residents throughout the last 5 years, also being accountable for over 40 % of the overall gross value added in the country and, therefore, playing an important role in the gross domestic product generation and development of the Latvian economy [1].

In order to choose the most applicable methods, tools and techniques of risk analysis for a particular small enterprise, a local map shop, an in-depth study of risk analysis, is carried out based on [2]–[24], starting with the very basics of risk definitions, classification, risk management frameworks and other important subject related issues.

This is followed by the development of the business process model for the particular business, using modeling language ArchiMate [25]. The developed model is used as the foundation for understanding company's business in a clear illustrative way and underlying the course of the following risk analysis.

This approach reflects the novelty of the study through combining business process modeling with a risk analysis, as well as combining different risk analysis methods and techniques to develop an individual risk analysis approach.

The practical risk analysis of the enterprise is carried out in three major steps. Step one is described as the preliminary analysis, devoted to inventory arrangement and ABC analysis, defining the most important inventory group in terms of use by value. Step two performs the method of qualitative risk analysis combining such risk analysis techniques as the semi-structured interview and cause-and-effect analysis for risk identification and classification, and combination of risk assessment table and risk matrix for risk evaluation and record of prevention and mitigation measures. Step three executes the quantitative risk analysis for the time period of 34 weeks, by developing a mathematical simulation model using Microsoft Office Excel, combining such techniques as the Monte Carlo simulation and scenario analysis.

II. PRELIMINARY ANALYSIS

The study of the particular enterprise business process showed that the main process was the sales management and the main object in this process was the inventory [26]. In order to carry out further inventory analysis, a certain arrangement of the shop's inventory was performed, distinguishing seven groups of items based on their type. Each item type was given a code name consisting of a Latin alphabet letter. This arrangement was further used for the ABC analysis to define the most important item groups in terms of use by value which would require most attention of management and analysis [27]. The use in terms of value was calculated for each of the previously defined product groups, as in (1). Equations (1) and (2) were adapted from [27].

$$U = Q \cdot P, \quad (1)$$

where U is the use in terms of value of a certain item, €; Q is a number of item units sold in a particular time period; P is a unit cost, €.

Percent of total use in terms of value was calculated, as in (2):

$$U_{T\%} = \frac{U}{U_T} \cdot 100\%, \quad (2)$$

where $U_{T\%}$ is a percent of total use in terms of value for an item group, %; U is the use in terms of value of an item group, €; U_T is a total use in terms of value of all item groups together, €.

A typical percentage distribution of use by value for categories was as follows: A category items – 80 %, B – 15 %, C – 5 %. This is also referred to as a standard Pareto analysis or “rule of 80/20”, which suggests that 20 % of inventory items need 80 % of the attention, while the remaining 80 % of

items need only 20 % of the attention [28]. It is important to note that this percentage can also vary, depending on the actual items [29], as illustrated in Fig. 1.

Items of *A* category were further analyzed in a quantitative risk analysis as they contributed most in terms of use by value.

III. QUALITATIVE ANALYSIS

Major risks of the retailer under the study are mainly associated with the inventory or will have an impact on inventory management if they occur.

Qualitative analysis starts with risk identification through a semi-structured interview, which involves certain aspects of a scenario analysis. A list of risks is provided in [26] considering most realistic potential risks and risk sources to the particular business. The list was constructed on the basis of the study on the subject related literature sources [2]–[24] and an insight in the company's business process. Results of the semi-structured interview were comprehensively illustrated through a cause-and-effect analysis or the so-called Ishikawa diagram, which can be found in [26]. Fragment of the diagram is given in Fig. 2.

The negative effect resulting in occurrence of certain risks is defined as the failure to meet predefined business objectives and satisfy customers. Example of a risk cause is customer associated risks such as inaccurate demand forecast, return of a purchase and other risks.

The obtained information is then summarized in a risk assessment table, using risk matrix to rank the risks and see which risks acquire most attention, according to their likelihood of occurrence and the impact on the main business objective should they occur.

Assigned likelihood and impact criteria have to be stipulated a prior assignment.

In the table suggestions are made on how the identified risks could be prevented or mitigated should they occur or what measures have already been taken by the management. Both the adapted risk matrix and table can be found in [26].

This analysis brought management's attention to the extreme risks in the business that had not been identified and considered prior a qualitative analysis; an example of such risk factor is suspension of globe supply.

It is important to understand that each business has both common and very specific risks that have to be considered during the analysis, as well as classification and division of

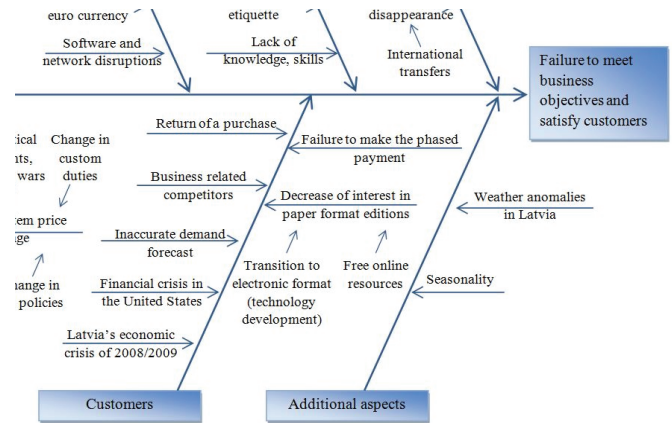


Fig. 2. Fragment of Ishikawa diagram.

Risks and the design of the risk assessment table are highly subjective as there is no unified approach.

IV. QUANTITATIVE ANALYSIS

This analysis provides a more numerical data-based approach to the risk previously described by a qualitative analysis, as concisely presented in Table I.

Inventory data obtained from the company presents certain quantitative information gaps that put restrictions on the further options of the quantitative analysis. For instance, there is no information concerning the dynamics of the demand, which excludes the use of discrete-event simulation software tools [30]. This aspect determines the criteria for choosing the indicator under the study of the analysis having it not to be directly connected with costs. Such an indicator is a service level.

As the available data are quite limited, generalized and insufficient to perform an accurate and in-depth quantitative analysis, the purpose of this study is to demonstrate a practical use of such quantitative risk analysis techniques as the Monte Carlo simulation and scenario analysis and see what general trends can be observed in the change of service level under different demand variation scenarios using Microsoft Office Excel 2007 (MS Excel), thus analyzing the risk of failing to satisfy the customer.

The analysis was performed for the time period of 34 weeks, considering only the items of *A* category as they had been previously categorized as the most important in terms of use by value [26].

TABLE I

ANALOGY BETWEEN THE ANALYZED ASPECTS OF QUALITATIVE AND QUANTITATIVE ANALYSIS

Risk analysis method	Analyzed risk	Effect on the main goal
Qualitative analysis	Difficult to predict demand accurately	Increasing number of unsatisfied customers
Quantitative analysis	Demand variation	Service level fluctuations

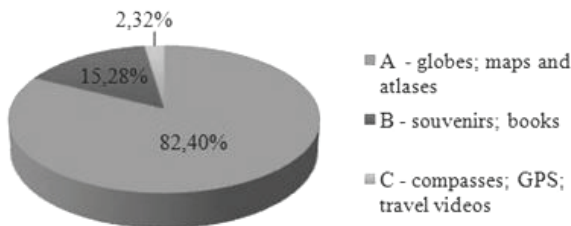


Fig. 1. Map shop's inventory categories and percentage of their total use in terms of value.

The first step of the quantitative risk analysis is to determine the type of theoretical distribution that best fits the character of the demand. An assumption is made that the demand is best described by truncated normal distribution, that is known to represent the properties of a service process and allows analyzing a system with several random variables, and also excludes negative values [30].

The next assumption is the choice of a supply tactic, when all that is known regarding the supply is the total value of the 34-week period. As there is no information on whether, if any, actual tactic is used by the company; it is assumed that a weekly supply is equal to the mathematical average, as in (3):

$$\bar{S} = \frac{S_T}{n}, \quad (3)$$

where $\bar{S}$ is an average weekly supply, €; S_T is a total supply per time period of 34 weeks, €; n is a number of weeks.

When the basic assumptions are made, it can be continued with Monte Carlo simulation for random number generation to model demand – the continuous random variable of our system. To start this simulation using MS Excel, we must define the initial parameters, which are as follows: average weekly demand; standard deviation of the demand according to the chosen scenario; a and b (see Fig. 3) as additional constraints for random number generation.

The average weekly demand is calculated in the same manner, as in (3), except the values in the equation are for demand.

Equation (4) is adapted from [31] and used to assign standard deviation for demand.

$$\sigma_D = x \cdot \bar{D}, \quad (4)$$

where σ_D is a standard deviation for demand, €; $\bar{D}$ is an average weekly demand, €; x is a decimal fraction of the demand variation percent according to a chosen variation (a chosen scenario).

MS Excel does not operate with truncated normal or distribution, only the normal distribution [32], which allows negative values, not valid in the case of demand. As illustrated in Fig. 3, adapted from [33], continuous random variables (demand, in our case) are described using a cumulative distribution function and the generation of random variables uses the inverse-transform technique $Y: F_Y(y)$.

In simple terms, we start with the value on the vertical axis x_i , representing the random numbers generated by MS Excel (greater than or equal to 0 and less than 1) to find the value on the horizontal axis y_i , representing the demand value for each week (D_i), as opposed to the traditional approach. However, this technique also allows for negative values, not valid for demand. Additional constraints a and b are required to respectively define the lower and upper limits on the horizontal axis. a is defined as 0 and b is defined as 3 standard deviations ($3\sigma_D$), thus eliminating the area on the left-hand side of the vertical axis, to provide solely positive random variable generation.

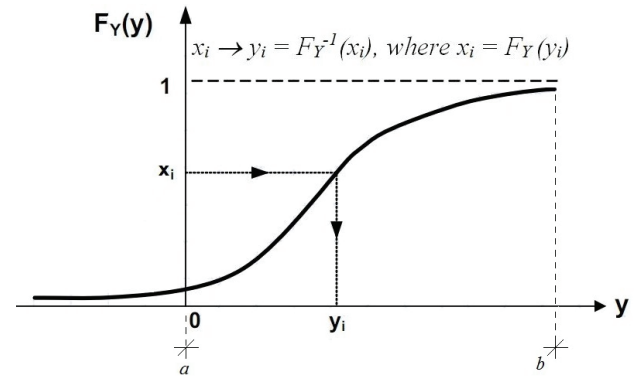


Fig. 3. Graphical representation of the inverse-transform technique.

When the previously mentioned steps are performed, the Monte Carlo simulation or, in other words, random demand value generation, can take place using a combination of such MS Excel functions as *NORMINV* and *RAND* [32].

After the demand is generated for each week of the 34-week period, we continue with the weekly residue calculation, taking into account that no negative values must appear. Equation (5), adapted from [34], describes the calculation of residue. The same equation is used to calculate loss, except in the case of loss MS Excel function allows for negative values.

$$R_i = I_i + S_i - D_i, \quad (5)$$

where R_i is a residue at the end of the week, €; I_i are items in stock at the beginning of the week, €; S_i is the average weekly supply, €; D_i is a demand of the week, €; i is a serial number of a particular week, ($i = \overline{1,34}$).

Table II presents one section (1 experiment) of the MS Excel simulation table. In total, 10 scenarios were carried out with demand variations of 20 %, 25 %, 30 %, 40 %, 45 % – 50%. Fewer values of variation simply proved to have no effect on the service level when performing the MS Excel simulation.

For each scenario, 100 replications (100 experiments) were conducted, each time generating a different demand value for each week, accordingly also changing in terms of stock,

TABLE II
SECTION OF A DEMAND VARIATION SIMULATION TABLE REPRESENTING ONE
EXPERIMENT UNDER THE SCENARIO OF 50% DEMAND VARIATION

Weeks	Stock (EUR)	Supply (EUR)	Demand (EUR)	Residue (EUR)	Loss (EUR)
1	10305.55	1085.18	2963.97	8426.76	0.00
2	8426.76	1085.18	1475.53	8036.41	0.00
...	...	...	...	...	...
19	222.34	1085.18	1916.62	0.00	-609.10
20	0.00	1085.18	1169.59	0.00	-84.41
...	...	...	...	...	...
34	917.02	1085.18	1639.61	362.59	0.00
Total demand (EUR)					49010.56
Total loss (EUR)					-2171.50
Service level SL_i					0.96

residue and loss values. The following approach was used – first, 20 experiments were conducted and then each of the 20 experiments was replicated 5 times.

A total loss and demand were calculated as a sum of each week indicators (accordingly, $\sum_{i=1}^n L_i$ and $\sum_{i=1}^n D_i$) for each experiment, allowing for further calculation of service level of each 34-week period, described by (6), adapted from [34].

$$SL_i = \frac{L_T + D_T}{D_T}, \quad (6)$$

where SL_i is a service level of a particular experiment expressed as a decimal number, 1 being 100 % service level; L_T is a total loss of the 34-week period of a particular experiment, €; D_T is a total demand of the 34-week period of a particular experiment, €.

After obtaining values of service levels of each experiment, the average service level for each scenario is calculated and used for the final analysis.

Standard deviation for a service level is a measure of the variability of service levels of each scenario and is calculated using MS Excel function, based on (7), adapted from [32].

$$\sigma_{SL} = \sqrt{\frac{\sum_{i=1}^n (SL_i - \overline{SL})^2}{(n-1)}}, \quad (7)$$

where σ_{SL} is a standard deviation for service level of a particular scenario; SL_i is a service level of a particular experiment; $\overline{SL}$ is an average service level of the particular scenario; n is a number of experiments.

Before continuing with further calculations regarding a service level, it is assumed that the obtained service level will fall into the defined confidence interval with a probability of 95%, which is a standard assumption [31].

For terminating systems, when the initial conditions are known, interval estimation of the average service level of each scenario is described in (8), adapted from [31].

The estimation of (8) starts with obtaining the probability coefficient t from the ‘t-distribution table’ found in various issue associated books, such as [30: 303] or using MS Excel function *TINV* [32]. In order to use such a table, two values have to be known: $\alpha/2$, which is half of the probability to make a mistake and design an interval that does not include the defined average service level and $(n-1)$, which are the degrees of freedom. According to $(n-1) = 19$ and $\alpha/2 = 0.025$ the probability coefficient t equals 2.093 [30].

$$\mu_{SL} \in \overline{SL} \pm \frac{\sigma_{SL} \cdot t_{n-1}^{\alpha/2}}{\sqrt{n}}, \quad (8)$$

where μ_{SL} is a mean value of service level; ϵ describes boundaries of the interval, where ‘+’ is used for the upper limit calculation of the confidence interval and ‘-’ for the lower limit; t is a probability coefficient; n is a number of experiments; α is a probability to make a mistake.

Results of all 100 experiments for each demand variation scenario are then summarized. A short summary of the data obtained from the MS Excel simulation is also presented in Table III.

Because of the limited data, the obtained numerical values are not considered to be accurate and are only analyzed in terms of the pattern they produce.

However, the large initial stock of 10305.55 € at the beginning of the 1st week of the analyzed period did not allow drawing precise numerical conclusions. All the more, analysis of the data showed that the average demand of the period was almost eight times less than the initial stock and the total residue of the period approximately three times less than the initial stock, indicating that the management was currently trying to minimize the initial stock rather than adapt a particular order quantity tactic. This could be a consequence of the 2008/2009 financial crisis, during which, the company suffered significant financial loss, so the following years, including the analyzed period, could have been dedicated to minimize the inventory pile-up of the crisis years. This is also the reason behind the minor service level fluctuations beginning only at the demand variation of 40 %, which is not genuine in the real life. Moreover, the service levels of 100 % in case of 20 % and 30 % demand variation are deceptive, the data of an item group are used, not exact data on each particular item and its demand or dissatisfied demand.

Comparison of scenarios of supply once per week and once per two weeks presented almost no difference in service level fluctuations, for the previously described reason.

Nevertheless, a general pattern is observed and illustrated in Fig. 4, displaying that, as the demand variation increases, service level starts to decrease. In the case of this particular analysis, values of the service level have little significance as they are predicted to be unrealistic, what is important, is the trend line, illustrating the previously described interrelation of demand variation and service level.

TABLE III
SUMMARY OF THE OBTAINED SERVICE LEVELS AND DEMAND VARIATIONS

Demand variation	Average service level	Lower limit for service level	Upper limit for service level
50 %	0.957	0.946	0.968
49 %	0.977	0.971	0.983
48 %	0.973	0.966	0.980
47 %	0.973	0.964	0.982
46 %	0.975	0.968	0.983
45 %	0.991	0.987	0.995
40 %	0.996	0.994	0.997
30 %	1.000	1.000	1.000
25 %	1.000	1.000	1.000
20 %	1.000	1.000	1.000

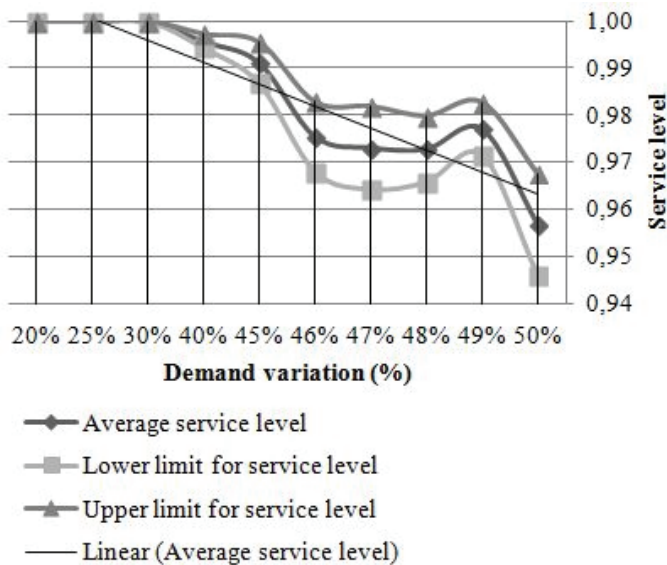


Fig. 4. General pattern of demand variation effect on a service level.

The trend line pattern agrees with the logic that as the demand variation increases (the accuracy of predicting such demand decreases), the service level decreases (more unsatisfied customers emerge). This in turn allows concluding that the developed simulation model within the Ms Excel is correct and can serve to produce more precise conclusions in terms of value when more precise data, such as satisfied and unsatisfied demand for each item, are available.

V. CONCLUSION

Small enterprises in Latvia are accountable for the majority of all economically active enterprises in the year 2013, thus encouraging development of Latvian entrepreneurship and employing over half of all residents of the country, also playing an important role in the gross domestic product generation and development of the Latvian economy. These facts substantiate the significance of risk analysis of a small enterprise in particular.

There is no unified definition of risk, nor a unified risk management and analysis framework or classification approach, making risk analysis highly subjective in terms of design and the specific business.

Development of a business process model, using any modeling language that best fits the need of the business under analysis, is a significant milestone for the risk analysis, facilitating not only the comprehension of the particular business but also giving an insight to elements and processes presenting potential sources for risk occurrence.

ABC analysis serves as a useful preliminary analysis tool to identify the most important item groups in terms of use by value that will require most attention.

Qualitative analysis demonstrates the practical development and use of a combination of such techniques as the semi-structured interview and cause-and-effect analysis to identify and classify risks as well as divide them into groups. Risk assessment table in combination with the risk matrix evaluates

the previously identified risks and makes suggestions on how these risks could be prevented or mitigated, should they occur.

Quantitative analysis demonstrates the practical use and development of a combination of such risk analysis techniques as the Monte Carlo simulation and scenario analysis using the software tool Microsoft Office Excel, and reveals that as the demand variation increases, service level starts to decrease. This in turn allows concluding that the developed simulation model is correct and can serve to produce more precise conclusions in terms of value when more precise data are available. In the case of this particular analysis, values of the service level have little significance as they are predicted to be unrealistic, because of the limited and generalized data.

The developed and performed modeling-based risk analysis of a particular small enterprise allows identifying gaps for business improvement.

For more precise mathematical simulation, a research on both satisfied and unsatisfied demand of the best-selling items is suggested to reveal critical demand variation in terms of non-tolerable service level.

REFERENCES

- [1] European Commission for Enterprise and Industry, "SME Performance Review: Database for the Annual Report," 2014. [Online]. Available: http://ec.europa.eu/enterprise/policies/sme/facts-figures-analysis/performance-review/index_en.htm. [Accessed: March, 2014].
- [2] A. M. Marchetti, *Enterprise Risk management Best Practices: From assessment to Ongoing Compliance*. Hoboken, New Jersey, USA: John Wiley & Sons, 2011, pp. 7–43.
<http://dx.doi.org/10.1002/9781118386699>
- [3] C. Verbano and K. Venturini, "Managing Risks in SMEs: A Literature Review and Research Agenda," *Journal of Technology Management and Innovation*, vol. 8, no. 3, 2013, pp. 187–189.
<http://dx.doi.org/10.4067/S0718-27242013000400017>
- [4] M. Abkowitz, D. Shimko and F. Fabozzi, *Approaches to Enterprise Risk Management*, Huntingdon, United Kingdom: Bloomsbury Information Ltd., 2010, pp. 3–66.
- [5] B.M. Ayyub, *Risk Analysis in Engineering and Economics*. USA: Chapman & Hall/CRC Press, 2003, 570 p.
<http://dx.doi.org/10.1201/9780203497692>
- [6] C.A. Wilhelmsen and L.T. Ostrom, *Risk Assessment: Tools, Techniques and Their Applications*. Somerset, NJ, USA: John Wiley & Sons, 2012, 392 p.
- [7] J. J. Hampton, *Fundamentals of Enterprise Risk Management: How Top Companies Assess Risk, Manage Exposure and Seize Opportunity*. Saranac Lake, New York, USA: AMACOM Books, 2009, pp. 4–18.
- [8] T. Aven, *Foundations of Risk Analysis*. Hoboken, New Jersey, USA: John Wiley & Sons, 2012, 211 p.
<http://dx.doi.org/10.1002/9781119945482>
- [9] P. Hopkin, *Fundamentals of Risk Management: Understanding, evaluating and implementing effective risk management*. United Kingdom: Kogan Page Ltd., 2010, pp. 11–61.
- [10] J.P. Chavas, *Risk Analysis in Theory and Practice*. Saint Louis, Missouri, USA: Elsevier Academic Press, 2004, pp. 5–11.
<http://dx.doi.org/10.1016/B978-012170621-0.50001-8>
- [11] S. Segal, *Corporate Value of Enterprise Risk Management: The Next Step in Business Management*. Hoboken, New Jersey, USA: John Wiley & Sons, 2011, 404 p.
- [12] J. Mun, *Modeling risk: Applying Monte Carlo Simulation, Strategic Real Options, Stochastic Forecasting and Portfolio Optimization*, Hoboken, New Jersey, USA: John Wiley & Sons, 2010, 986 p.
- [13] *Risk management – Vocabulary*, International Standard ISO Guide 73:2009, 2009, 15 p.
- [14] *A Practical Guide to Risk Assessment*. Price water house Coopers, 2008, pp. 5–32. [Online]. Available: http://www.pwc.com/en_US/us/issues/enterprise-risk-management/assets/risk_assessment_guide.pdf. [Accessed: May, 2014].

- [15] "3rd Annual Survey on Supply Chain Resilience," *Business Continuity Institute*, 2011, pp. 6–11.
- [16] B. Bemeleit, J. Schumacher and C. Hans, "Methods of Risk Assessment and their Sustainability in Logistic Environment," *Symposium on Risk Management and Cyber-Informatics: RMCI'05 in the Context of the 9th Multi-conference on Systemics, Cybernetics and Informatics*, SCI 2005, 2005, pp. 425–431.
- [17] W.G., Shenkir and P.L. Walker, *Enterprise Risk Management: Tools and Techniques for Effective Implementation*. Montvale, New Jersey, USA: Institute of Management Accountants, 2007, pp. 4–14.
- [18] A. Szabo, "Risk Management: An Integrated Approach to Risk Management and Assessment," *Annals of the University of Oradea: Economic Science Series*, vol. 21, no. 2, 2012, pp. 776–781.
- [19] *Understanding Supply Chain Risk: A Self – Assessment Workbook*. Bedford, United Kingdom: Cranfield University, Cranfield School of Management, Centre for Logistics and Supply Chain Management, 2003, pp. 16–21.
- [20] *Risk management – Principles and guidelines*, International Standard ISO 31000:2009, 2009, 24 p.
- [21] *Risk management – Risk assessment techniques*, International Standard ISO/IEC 31010:2009, 2009, 176 p.
- [22] Y. Yacov, *Risk Modeling, Assessment and Management*. Hoboken, New Jersey, USA: John Wiley & Sons, 2005, 837 p.
- [23] A.J. Anderson, *Business Risk Management: Models and Analysis*. Oxford, United Kingdom: John Wiley & Sons, 2013, 367 p. <http://dx.doi.org/10.1002/9781118749388>
- [24] D. Ritter, "Home page – certified project management professional D. Ritter," Oct. 2008. [Online]. Available: <http://certifiedpmp.wordpress.com/2008/09/17/risk-analysis-techniques/> [Accessed: Apr., 2014].
- [25] Website of the independent modeling language for enterprise architecture 'ArchiMate', Dec. 2012. [Online]. Available: <http://archi.cetis.ac.uk>. [Accessed: Apr., 2014].
- [26] K. Bendza, "Modeling-based risk analysis of a small enterprise," M.S. thesis, Riga Technical University, BALTECH Study Centre, Riga, Latvia, 2014.
- [27] D. Walters, *Logistics: An Introduction to Supply Chain Management*. United Kingdom: Palgrave Macmillan, 2003, pp. 274–275.
- [28] R. Koch, *The 80/20 Principle: The Secret of Achieving More with Less*. New York, USA: The Doubleday Publishing Group, 2008, pp. 3–6.
- [29] G. Ghiani, G. Laporte and R. Musmanno, *Introduction to Logistics Systems Planning and Control*. United Kingdom: John Wiley & Sons, 2004, pp. 148–149.
- [30] S. Robinson, *Simulation: The Practice of Model Development and Use*. Chichester, United Kingdom: John Wiley & Sons, 2004, 316 p.
- [31] J. Shao, *Mathematical Statistics, 2nd Edition*. New York, USA: Springer, 2003, 591 p. <http://dx.doi.org/10.1007/b97553>
- [32] L. W. Winston, *Microsoft Office Excel 2007: Data Analysis and Business Modeling*. Washington, USA: Microsoft Press, 2007, 601 p.
- [33] J. Banks, J. S. Carson and B. L. Nelson, *Discrete-Event System Simulation*, 4th ed., USA: Prentice-Hall, 2004, 608 p.
- [34] T. M. Choi, *Handbook of EOQ Inventory Problems: Stochastic and Deterministic Models and Applications*, New York, USA: Springer, 2014, 281 p.



Kristine Bendza obtained the degree of *Mg .sc. ing.* in Business Management in the field of Industrial Engineering and Management from BALTECH Study Centre, Riga Technical University in 2014 and the degree of *B. sc. ing.* in Civil Engineering as well as the Qualification of Civil Engineer from Riga Technical University in 2012. Currently she works as A Sales Manager for a local retailer.
E-mail: kristine.bendza@gmail.com.

Jelena Pečerska is a Doctor of Information Technology in the field of System Analysis, Modeling and Development. She has been working for Riga Technical University since 1979. At the moment, she is an Assistant Professor with the Department of Modeling and Simulation of Riga Technical University. Her professional interests include methodology of discrete-event simulation, combined simulation, supply chain modeling, practical applications of discrete-event simulation and discrete-event simulation in education. She is a member of the Latvian Simulation Society.
E-mail: jelena@itl.rtu.lv.

Kristīne Bendža, Jeļena Pečerska. Modelēšanā bāzētā mazā uzņēmuma risku analīze

Globalizācijas un straujās tehnoloģiju attīstības pieaugošā spiediena ietekmē pēdējo gadu laikā risku pārvaldība un analīze ir nonākusi daudzu uzņēmumu uzmanības centrā, strauji ieņemot svarīgu vietu vadītāju skatījumā. Neskatoties uz informācijas avotu daudzveidību, vienota pieeja risku analīzei nepastāv, pamatojoties uz dažādajiem uzņēmējdarbības veidiem, ekonomiskajiem, politiskajiem un citiem aspektiem, turklāt lielākā daļa literatūras šajā kontekstā apskata tikai lielos uzņēmumus. Šis raksts ir veltīts maza uzņēmuma risku analīzei. Šī raksta mērķis ir prezentēt īpaši pētāmajam uzņēmumam pielāgotu analīzes pieeju. Rakstā ir aplūkotas analīzes metodes konkrētam uzņēmumam. Praktiskā uzņēmuma risku analīze ietver trīs galvenos posmus. Pirmajā analīzes posmā tiek paveikta uzņēmuma krājumu ABC analīze, un krājumi tiek sagrupēti pēc to svarīguma. Uz šīs analīzes pamata ir piedāvāta kvalitatīva un kvantitatīva pieeja pētāmā uzņēmuma risku analīzei. Otrajā posmā veikta riska faktoru kvalitatīvā analīze, izmantojot metožu kombināciju – daļēji strukturēto intervju un cēloņu un sekas analīzi. Rezultātā riska faktori tiek identificēti un klasificēti. Tāpat šajā posmā risku novērtēšanai un pasākumu izstrādei, risku sekas novērtēšanai un mazināšanai tiek izmantota riska novērtējuma tabula kombinācijā ar riska matricu. Trešajā posmā paveikta uzņēmuma risku kvantitatīvā analīze. Kvantitatīvā analīze apstiprina apkalpošanas kvalitātes līmeņa atkarību no pieprasījuma mainīguma pakāpes, un rezultātā uzņēmuma riska faktoru pētījumu veikšanai tiek piedāvāts matemātiskais modelis. Analīze balstās uz Monte Karlo modelēšanas rezultātiem, modelis tika izstrādāts, izmantojot *Microsoft Office Excel* programmrīku. Statistiskajā modelēšanā bāzētā mazā uzņēmuma risku analīze sniedz ieskatu par izstrādāto un pielietoto analīzes pieeju uzņēmējdarbības jutīgo punktu un nepilnību identificēšanai.

Кристина Бенджа, Елена Печерская. Анализ рисков малого предприятия на основе моделирования

Повышенное давление глобализации и быстрого технического прогресса в последние годы выдвинули управление и анализ рисков, играющих всё более важную роль в повестке дня руководителей, в центр внимания большинства предприятий. Несмотря на разнообразие источников информации, единого подхода к анализу рисков не существует, поскольку подход зависит от конкретного бизнеса, экономических, политических и других аспектов. Кроме того, в большей части публикаций рассматриваются только крупные предприятия. Данная статья рассматривает анализ рисков малого предприятия. Целью статьи является рассмотрение подхода к анализу рисков малого предприятия. В статье рассмотрены методы анализа рисков для конкретного предприятия. Практический анализ рисков предприятия включает в себя три основных этапа: на начальном – предварительном – этапе выполнен ABC анализ складских запасов предприятия, и запасы сгруппированы по степени их важности. На базе предварительного анализа предложен качественный и количественный подход к анализу рисков изучаемого предприятия. На втором этапе выполнен качественный анализ факторов риска с использованием комбинации методов полуструктурированного интервью и причинно-следственного анализа. В результате факторы риска идентифицированы и классифицированы. На этом же этапе использована таблица оценки рисков в сочетании с матрицей рисков для оценки рисков и формулирования мероприятий по предотвращению рисков и смягчению их последствий. На третьем этапе выполнен количественный анализ рисков предприятия. Количественный анализ позволил подтвердить зависимость уровня качества обслуживания от степени изменчивости спроса и предложить математическую модель для проведения исследований факторов риска предприятия. Анализ основан на результатах моделирования с применением метода Монте Карло в программе *Microsoft Office Excel*. Анализ рисков малого предприятия на основе статистического моделирования даёт представление о разработанном и реализованном подходе к идентификации чувствительных точек и недостатков бизнес-процесса.

Building Ontology from Relational Database

Andrejs Kaulins¹, Arkady Borisov^{2, 1, 2} *Riga Technical University*

Abstract – This paper proposes an approach of building OWL2 ontology from data in a relational database. Compared with existing methods, the approach can acquire ontology from a relational database automatically by using a group of learning rules. Approach is independent from database implementation; it is designed based on standards, used to manage database systems. In this paper, we describe object mapping from a relational database and OWL2 ontology including classes, data properties, object properties, instances, axioms and annotations. Axioms are important part of OWL2 specification, and the approach is also suitable for the axiom building process.

Keywords – Ontology, ontology building, relational database relational model.

I. INTRODUCTION

Many approaches are developed to extract knowledge from existing relational databases. Tools are implemented using OWL language. However ontology models are not complete, still a lot of things are out of scope. This paper aims to solve the problem of building OWL ontology using brand new notation for describing ontologies. This notation OWL2 is recommended by W3C consortium, and it is extension of OWL [9]. It includes well-defined semantics for ontology reusability, reasoning and contains many profiles, thus enabling one to use OWL2 for describing knowledge in specific domains.

Relational database consists of data, data implemented as some relational model. Also operations and constraints are developed in the database to implement business rules. It is not an easy task to extract all these things and build ontology on them.

We provide a new approach, based on OWL2 language, and relation database is not normalized; this approach works with the database “as it is”.

The paper is organized as follows. Section I analyzes current related studies; Section II gives an overview of standards used in relational database management systems. Section III gives some preliminary definitions used in the mapping algorithm between objects in the relational database and OWL2 ontology. In Section IV we describe relational database sample schema. Section V defines approach learning rules, and section VI draws conclusions and gives an insight into the future research.

II. RELATED WORK

Much attention has been devoted to the issue of building an ontology from a relational database. Ontology is used to support the sharing and reuse of formally represented knowledge among AI systems [2]. To solve the problem, researchers use different approaches. One of them is using concept hierarchy [4]. Concept hierarchy is a type of

background knowledge, which expresses the structure of concept from a low-level concept to a more general concept. Advantages of this approach – it uses information entropy as a metric for the correlation between any two features. As a result, this approach allows extracting knowledge selectively and merging it to the existing ontology. Weakness of this approach: it does not allow building axioms used in reasoning. Ontology is built using OWL recommendation without formally defined meaning. For defining reasoning predicates some semantic web rule language should be used. Another approach [5], [10], [11] can acquire ontology from relational database automatically by using a group of learning rules. It is also based on OWL recommendation. The method [6] designed to build a conceptual model from a relational model. It uses data analysis in tables to build generalization abstraction. These conceptual models can be used in ontology building. This approach also lacks a reasoning process, and could not be used in decision support systems. Our proposed approach is aimed at building an ontology model with strong defined semantics and knowledge extracted from a relational database. OWL2 new recommendation owns these features. It could use other ontologies as well. To build ontology, we used learning rules, even comments and remarks in the source database were mapped in ontology as annotations and axioms.

III. DATABASE MANAGEMENT STANDARDS

There are more than 50 implementations of relational database management systems. Most popular is DB/2 from IBM, SQL server from Microsoft, Oracle Database from Oracle, MySQL and Sybase. To enable one to share information between databases, the structured query language (SQL) is used. To enable the portability of SQL applications across conforming implementations, standards are used. The latest standard is SQL99 (the so-called SQL/3) [8], SQL/4 is now undergoing the development phase.

IV. PRELIMINARY DEFINITIONS

The underlying model of relational database is the relational model [1], in which each relation R_i corresponds to a table. The columns of a relation are called attributes denoted as A . Each attribute has type and range of values of this type. Let us define some predicates, which acquire the table primary key, table foreign key, attribute name, type and range of values:

1. $pkey(R_i)$ – returns the name of primary key; the key can be composed of one or many table columns;
2. $|pkey(R_i)|$ – a number of attributes participated in key composition;
3. $pkeyAttr(pkey(R_i))$ – returns attributes contained in primary key, when $|pkey(R_i)| > 1$;

4. $\text{fkey}(R_i)$ – returns the name of foreign key; key values are values from the primary key in another table. Through a foreign key a referential constraint is implemented in a relational model;
5. $\text{fkeyAttr}(\text{fkey}(R_i))$ – returns attributes contained in the foreign key, when $|\text{fkey}(R_i)| > 1$;
6. $\text{attr}(R_i)$ – returns attributes contained in a specific relation R_i , output is an array of all attributes, contained in table $A_{11}, \dots, A_{1n}$, where n – is an attribute count;
7. $\text{attrType}(R_i, A_i)$ – returns the attribute type of attribute A_i , contained in table R_i ;
8. $\text{attrDom}(R_i, A_i)$ – returns a value range of values of attribute A_i , contained in table R_i ;
9. $\text{attrName}(R_i, A_i)$ – returns the name of attribute A_i contained in table R_i ;
10. $\text{tuple}(R_i, T_1)$ – returns values of record 1 (tuple 1) of table R_i . Number of rows in table R_i can change, but for reference let us assume m is the current maximum number of tuples;
11. $\text{attrValue}(R_i, A_i, T_1)$ – returns the value of attribute A_i in table R_i of tuple 1 - T_1 ;
12. $\text{tableName}(R_i)$ – returns the name of table R_i .

Relational database contains not only tables, but also constraints and triggers, which help encapsulate business rules into the database [3]. Constraints can define ranges on attribute values, limit value duplication in attributes (unique constraint), or define reference to other tables (referential constraint).

Relationships between relations are important. Two relations can have one-to-one, one-to-many or many-to-many relationship types. The data duplication degree in a database depends on the database normalization type [12]. In practice, it is not always possible to normalize a data model. These factors are taken into account designing the proposed approach.

V. RELATIONAL DATABASE SCHEMA

Let us consider a relational database schema example to demonstrate mapping for tables, constraints and data to ontology (Fig. 1). This sample schema is taken from oracle predefined examples and could be installed with oracle database installation [7].

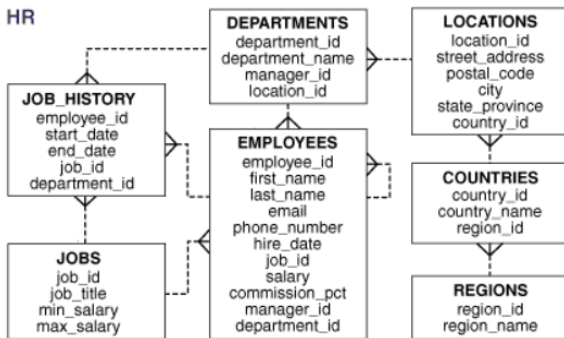


Fig. 1. A relational database schema example.

VI. PROPOSED APPROACH AND LEARNING RULES

An approach for building ontology in OWL2 consists of 6 major steps:

Step I. Relation (table) analysis. At this step using predicates defined in Section IV, we find out the name of tables, names of all columns, comments on table and columns. Using a database dictionary we find out all constraints, defined in the table, including a primary key, foreign key(s). Table I provides details of database schema structure and table constraints.

Based on this information, we could build OWL2 class as a knowledge concept in ontology.

Rules used at this step:

Rule 1: For relations $R_1, R_2, \dots, R_k$ in database. Let us suppose that $P_1 = \text{pkey}(R_1)$, $P_2 = \text{pkey}(R_2)$, ..., $P_k = \text{pkey}(R_k)$. If between relations $R_1, R_2, \dots, R_k$ there is equivalence, i.e., $R_1(P_1) = R_2(P_2) = \dots = R_k(P_k)$, then the information spread across $R_1, R_2, \dots, R_k$ should be integrated into an ontological class.

Rule 2: An ontological class can be created based on relation R_i , if the following conditions can be satisfied:

$|\text{pkey}(R_i)| = 1$ or $|\text{pkey}(R_i)| > 1$ and there is A_i , where $A_i \in \text{pkeyAttr}(\text{pkey}(R_i))$ and $A_i \notin \text{fkeyAttr}(\text{fkey}(R_i))$.

In ontology OWL2, a class entity for table HR.EMPLOYEES can be declared as:

Ontology(<http://www.my.example.com/example>

Declaration(Class(a:Employees))

Declaration(Class(a:Departments))

Declaration(Class(a:Jobs))

Declaration(Class(a:Locations))

Declaration(Class(a:Regions))

Declaration(Class(a:Countries))

Declaration(Class(a:Job_History))

Step II. In the selected table we analyze columns and extract all constraints with predicates $\text{attr}(R_i)$, $\text{attrType}(R_i, A_j)$, $\text{attrDom}(R_i, A_j)$, $\text{attrName}(R_i, A_j)$, $\text{pkey}(R_i)$, $\text{pkeyAttr}(\text{pkey}(R_i))$, $\text{fkey}(R_i)$, $\text{fkeyAttr}(\text{fkey}(R_i))$

Rules used at this step:

Rule 3: For relations R_i and R_j , if there is a dependency relationship, i.e., $R_i(A_i) \subseteq R_j(A_j)$ and $A_i \notin \text{pkeyAttr}(\text{pkey}(R_i))$ are satisfied, then object property P_i can be created on A_i . On ontology OWL2 these can be declared as follows (for relation "HR"."EMPLOYEES"):

ObjectPropertyAssertion(a:worksInDepartment
a:Employees a:Departments)

ObjectPropertyAssertion(a:haveJobt a:Employees a:Jobs)

ObjectPropertyAssertion(a:haveManager a:Employees
a:Employees)

The third object property describes a relation of table to itself, i.e., managers are also employees, but each employee has a manager.

TABLE I
RELATIONAL DATABASE SCHEMA DETAILS

No.	Relation	Constraints	Primary Key	Foreign Key
1	"HR"."EMPLOYEES" ("EMPLOYEE_ID" NUMBER(6,0), "FIRST_NAME" VARCHAR2(20 BYTE), "LAST_NAME" VARCHAR2(25 BYTE), "EMAIL" VARCHAR2(25 BYTE), "PHONE_NUMBER" VARCHAR2(20 BYTE), "HIRE_DATE" DATE, "JOB_ID" VARCHAR2(10 BYTE), "SALARY" NUMBER(8,2), "COMMISSION_PCT" NUMBER(2,2), "MANAGER_ID" NUMBER(6,0), "DEPARTMENT_ID" NUMBER(4,0))	1. "LAST_NAME" IS NOT NULL 2. "EMAIL" IS NOT NULL 3. "HIRE_DATE" IS NOT NULL 4. "JOB_ID" IS NOT NULL 5.CONSTRAINT "EMP_SALARY_MIN" CHECK (salary > 0) 6.CONSTRAINT "EMP_EMAIL_UK" UNIQUE ("EMAIL")	"EMPLOYEE_ID"	1. CONSTRAINT "EMP_DEPT_FK" FOREIGN KEY ("DEPARTMENT_ID") REFERENCES "HR"."DEPARTMENTS" ("DEPARTMENT_ID") 2. CONSTRAINT "EMP_JOB_FK" FOREIGN KEY ("JOB_ID") REFERENCES HR"."JOBS" ("JOB_ID") 3. CONSTRAINT "EMP_MANAGER_FK" FOREIGN KEY ("MANAGER_ID") REFERENCES HR"."EMPLOYEES" ("EMPLOYEE_ID")
2	"HR"."DEPARTMENTS"("DEPARTMENT_ID" NUMBER(4,0), "DEPARTMENT_NAME" VARCHAR2(30 BYTE), "MANAGER_ID" NUMBER(6,0), "LOCATION_ID" NUMBER(4,0))	"DEPARTMENT_NAM E" IS NOT NULL	"DEPARTMENT_ID"	1."DEPT_LOC_FK" FOREIGN KEY ("LOCATION_ID") REFERENCES "HR"."LOCATIONS" ("LOCATION_ID") 2."DEPT_MGR_FK" FOREIGN KEY ("MANAGER_ID") REFERENCES "HR"."EMPLOYEES" ("EMPLOYEE_ID")
3	"HR"."JOBS"("JOB_ID" VARCHAR2(10 BYTE), "JOB_TITLE" VARCHAR2(35 BYTE), "MIN_SALARY" NUMBER(6,0), "MAX_SALARY" NUMBER(6,0))	"JOB_TITLE" IS NOT NULL	"JOB_ID"	-
4	"HR"."LOCATIONS"("LOCATION_ID" NUMBER(4,0), "STREET_ADDRESS" VARCHAR2(40 BYTE), "POSTAL_CODE" VARCHAR2(12 BYTE), "CITY" VARCHAR2(30 BYTE), "STATE_PROVINCE" VARCHAR2(25 BYTE), "COUNTRY_ID" CHAR(2 BYTE))	"CITY" IS NOT NULL	"LOCATION_ID"	CONSTRAINT "LOC_C_ID_FK" FOREIGN KEY ("COUNTRY_ID") REFERENCES "HR"."COUNTRIES" ("COUNTRY_ID")
5	"HR"."REGIONS"("REGION_ID" NUMBER "REGION_NAME" VARCHAR2(25 BYTE))	"REGION_ID" IS NOT NULL	"REGION_ID"	-
6	"HR"."COUNTRIES"("COUNTRY_ID" CHAR(2 BYTE), "COUNTRY_NAME" VARCHAR2(40 BYTE), "REGION_ID" NUMBER)	"COUNTRY_ID" IS NOT NULL	"COUNTRY_ID"	CONSTRAINT "COUNTR_REG_FK" FOREIGN KEY ("REGION_ID") REFERENCES "HR"."REGIONS" ("REGION_ID")
7	"HR"."JOB_HISTORY"("EMPLOYEE_ID" NUMBER(6,0), "START_DATE" DATE, "END_DATE" DATE, "JOB_ID" VARCHAR2(10 BYTE), "DEPARTMENT_ID" NUMBER(4,0))	1. "EMPLOYEE_ID" IS NOT NULL 2. "START_DATE" IS NOT NULL 3. "END_DATE" IS NOT NULL 4. "JOB_ID" IS NOT NULL	PRIMARY KEY ("EMPLOYEE_ID", "START_DATE")	1. CONSTRAINT "JHIST_JOB_FK" FOREIGN KEY ("JOB_ID") REFERENCES "HR"."JOBS" ("JOB_ID") 2. CONSTRAINT "JHIST_EMP_FK" FOREIGN KEY ("EMPLOYEE_ID") REFERENCES "HR"."EMPLOYEES" ("EMPLOYEE_ID") 3. CONSTRAINT "JHIST_DEPT_FK" FOREIGN KEY ("DEPARTMENT_ID") REFERENCES "HR"."DEPARTMENTS" ("DEPARTMENT_ID")

	EMPLOYEE_ID	FIRST_NAME	LAST_NAME	EMAIL	PHONE_NUMBER	HIRE_DATE	JOB_ID	SALARY	COMMISSION_PCT	MANAGER_ID	DEPARTMENT_ID
1	198	Donald	OConnell	DOCONNEL	650.507.9833	2007.06.21	SH_CLERK	2600	(null)	124	50
2	199	Douglas	Grant	DGRANT	650.507.9844	2008.01.13	SH_CLERK	2600	(null)	124	50
3	200	Jennifer	Whalen	JWHALEN	515.123.4444	2003.09.17	AD_ASST	4400	(null)	101	10
4	201	Michael	Hartstein	MHARTSTE	515.123.5555	2004.02.17	MK_MAN	13000	(null)	100	20
5	202	Pat	Fay	PFAY	603.123.6666	2005.08.17	MK_REP	6000	(null)	201	20
6	203	Susan	Mavris	SMAVRIS	515.123.7777	2002.06.07	HR_REP	6500	(null)	101	40
7	204	Hermann	Baer	HB&ER	515.123.8888	2002.06.07	PR_REP	10000	(null)	101	70
8	205	Shelley	Higgins	SHIGGINS	515.123.8080	2002.06.07	AC_MGR	12008	(null)	101	110
9	206	William	Gietz	WGIEZT	515.123.8181	2002.06.07	AC_ACCOUNT	8300	(null)	205	110
10	100	Steven	King	SKING	515.123.4567	2003.06.17	AD PRES	24000	(null)	(null)	90
11	101	Neena	Kochhar	NKOCHHAR	515.123.4568	2005.09.21	AD VP	17000	(null)	100	90
12	102	Lex	De Haan	LDEHAAN	515.123.4569	2001.01.13	AD VP	17000	(null)	100	90
13	103	Alexander	Hunold	AHUNOLD	590.423.4567	2006.01.03	IT_PROG	9000	(null)	102	60

Fig. 2. Table "HR"."EMPLOYEES" data example.

Rule 4: For relations R_i and R_j , two ontological object properties “has-part” and “is-part-of” can be created, if the two conditions are satisfied:

$$|pkey(R_i)| > 1 \text{ and}$$

$$fkeyAttr(fkey(R_i)) \subset pkeyAttr(pkey(R_j))$$

Rule 5: For relations R_i , R_j and R_k , if $A_i = pkeyAttr(pkey(R_i))$, $A_j = pkeyAttr(pkey(R_j))$, $A_i \cup A_j = fkeyAttr(fkey(R_k))$ and $A_i \cap A_j = \emptyset$, then two object properties P_i , P_j can be created based on the semantics of R_k .

Rule 6: For relations $R_1, R_2, \dots, R_i$ and R_k in the database, if $A_1 = pkeyAttr(pkey(R_1))$, $A_2 = pkeyAttr(pkey(R_2))$, $\dots$, $A_i = pkeyAttr(pkey(R_i))$, $A_1 \cup A_2 \cup \dots \cup A_i = fkeyAttr(fkey(R_k))$ and $A_1 \cap A_2 \cap \dots \cap A_i = \emptyset$, then object properties $P_1, P_2, \dots, P_i$ can be created. Rule 6 extends Rule 5 for n-ary relationship provided by R_k .

This rule introduces mapping datatype properties (basic oracle database and OWL2 datatypes are presented in Table II) to ontology, presented in OWL2:

Rule 7: For an ontological class C_i and datatype properties set of C_i denoted as $DP(C_i)$, if C_i is corresponding to relations $R_1, R_2, \dots, R_i$ in database, then for every attribute in $R_1, R_2, \dots, R_i$, if it cannot be used to create an object property by using Rule 3, then it can be used to create datatype property of C_i . The domain and range of each property are extracted using predicates for relational database $attrDom(R_i, A_j)$ and $attrValue(R_i, A_j, T_k)$.

For table "HR"."EMPLOYEES" we could define datatypes as follows:

```
DataPropertyRange( a:hasFirstName xsd:string )
DataPropertyRange( a:hasLastName xsd:string )
DataPropertyRange( a:hasEmail xsd:string )
DataPropertyRange( a:hasPhoneNumber xsd:string )
DataPropertyRange( a:hasHireDate xsd:dateTime )
DataPropertyRange( a:hasSalary xsd:decimal )
DataPropertyRange( a:hasCommision% xsd:decimal )
```

Rule 8: For relations R_i and R_j , supposed that $P_i = pkeyAttr(pkey(R_i))$, $P_j = pkeyAttr(pkey(R_j))$, if $P_i(R_i) \subseteq P_j(R_j)$ is satisfied, then class/property corresponding to R_i is subclass/subproperty of the class/property corresponding to R_j .

Table "HR"."EMPLOYEES" contains employees and also managers, so in the ontology this fact could be expressed as follows:

```
Ontology( http://www.my.example.com/example
SubClassOf( a:Managers a:Employees ) )
```

Rules for learning cardinality

Rule 9: For relation R_i and $A_i \in attr(R_i)$, if $A_i = pkeyAttr(pkey(R_i))$ or $A_i = fkeyAttr(pkey(R_i))$, then the minCardinality and maxCardinality of the property P_i corresponding to A_i is 1.

In our example, at each location there may be only one warehouse, in the ontology model this can be expressed as:

```
ObjectExactCardinality( 1 a:hasWarehouse a:Location )
```

Rule 10: For relation R_i , and $A_i \in attr(R_i)$, if A_i is declared as IS NOT NULL, the maxCardinality of the property P_i corresponding to A_i is 1.

Rule 11: For relation R_i , and $A_i \in attr(R_i)$, if A_i is declared as UNIQUE, the maxCardinality of the property P_i corresponding to A_i is 1

TABLE II

BASIC ORACLE DATABASE DATATYPE MAPPING TO OWL2 ONTOLOGY

Data type	RDBMS data types	OWL2
number	number	xsd:int
number	number(precision, scale)	xsd:decimal
number	binary_float	xsd:float
number	binary_double	xsd:double
text	char, nchar, varchar, varchar2, nvarchar2, long, clob, nclob	xsd:string
date time	date	xsd:dateTime
date time	timestamp, timestamp with time zone	xsd:time
binary	blob, dfile	xsd:base64Binary

Step III. The third step is the extraction of rows (tuples) from the relational database and ontology addition with individuals.

Rule 12: Every tuple T_i in the relational database corresponds to an individual of class in the ontology model.

In table "HR"."EMPLOYEES", each record can be represented as an individual of class a:Employees:

```
DataPropertyAssertion(a:hasFirstName a:Employees
"Douglas" ^ xsd:string )
DataPropertyAssertion(a: a:hasLastName a:Employees
"Grant" ^ xsd:string )
DataPropertyAssertion(a: hasEmail a:Employees "DGRA"
^ xsd:string )
DataPropertyAssertion(a: hasPhoneNumber a:Employees
"6505079844" ^ xsd:int )
DataPropertyAssertion(a: hasHireDate a:Employees
"2008.01.13" ^ xsd:dateTime )
DataPropertyAssertion(a: hasSalary a:Employees "2600"
^ xsd:int )
DataPropertyAssertion(a: hasCommision% a:Employees
"2" ^ xsd:int )
```

Mappings between objects in the oracle database and OWL2 ontology is provided in Table III.

Step IV. Next we analyze table values and refine some relationship between OWL entities. That can be done using predicate $attrValue(R_i, A_j, T_k)$.

Step V. Finally we can add to ontology some information from the relational database, which will not participate in a reasoning process, i.e., commentaries, remarks, etc.:

```
AnnotationAssertion ( rdfs:comment a:hasFirstName "First
name of the employee. A not null column.")
```

Step VI. Next step is ontology validation, for validation process our approach uses domain specific rules to build axioms and derive new facts in domain.

TABLE III
RDBMS OBJECT AND OWL2 ENTITY MAPPING

No.	RDBMS object	OWL2 model object
1	Table	OWL2 Class
2	Table Column	Functional property
3	Table Row	OWL2 individual
4	Column metadata: Data Type Not Null Null	OWL2 restriction: All values from restriction Cardinality() restriction maxCardinality() restriction
5	Constraint: Not Null Unique Foreign Key Check	OWL2 property: ObjectExactCardinality(1 OPE CE) InverseFunctionalProperty objectProperty hasValue

After all these steps are done, we can continue knowledge extraction from the relational database, switching to a next table or can stop the building process.

VII. CONCLUSION AND FUTURE RESEARCH

In this paper, we proposed a new approach for building ontology, using new OWL2 language for describing ontologies. Approach uses mapping rules to define different entities in ontology – classes, functional properties, object data properties, class cardinality, restrictions, individuals and annotations. To demonstrate OWL2 semantics and language notation, we used a plain example, implemented in the oracle database. Approach is not restricted to specific database implementation, because it is based on international standards, used to manage data – SQL99.

In future, we plan to implement this approach on Java and to build a system prototype for an expert advising system in health treatment domain.

REFERENCES

- [1] E. F. Codd, "A Relational Model of Data for Large Shared Data Banks", *Communications of ACM*, vol. 13, no. 6, June 1970, pp. 377–387.
- [2] T. R. Gruber, "A translation approach to portable ontology specifications", technical report KSL 92–71, 1993.
- [3] C. J. Date, *An introduction to Database Systems*, 8th ed., Addison-Wesley, 2004, p. 1024.

Andrejs Kauliņš, Arkādijs Borisovs. Ontoloģiju izveide no relāciju datubāzes

Ontoloģijas izveides process prasa daudz laika un dažādu speciālistu iesaistīšanos. Relāciju datubāzes satur informāciju, ko var izmantot, lai veidotu ontoloģijas. Šajā rakstā ir aprakstīta metode, kas ļauj automātiski uzbūvēt ontoloģiju. Būvēšanas procesā no relāciju datu bāzes izmanto kartēšanas noteikumus. Šie noteikumi nosaka, kādā veidā objekti no relāciju datu bāzes ir sasaistīti ar ontoloģijas elementiem. Metodei nav ierobežojumu relāciju datubāzei, jo tā ir balstīta uz pieņemtiem datu glabāšanas standartiem SQL99. Lai aprakstītu ontoloģiju, tika izmantota OWL2 valoda. Kartēšanas noteikumi ļauj veidot klases un to īpašības, datu īpašības un eksemplārus ontoloģijā. Kartēšanas noteikumu izveidošana ir sarežģīts uzdevums. Tomēr tas ir atrisināts vienkāršiem datu modeļiem. Reālos dzīves gadījumos, relāciju datu bāzē tiek izmantotas sarežģītas datu struktūras. Tās ietver patvaļīgus objektus, kurus definē pats lietotājs. Rakstā ir apskatīti kartēšanas noteikumi, kuri darbojas ar dažādu relāciju datu bāzu datu modeļiem. Aprakstīta atbalsta sistēma, kura nodrošina ontoloģijas konstruēšanu ar kartēšanas noteikumu palīdzību. Galvenais šīs sistēmas mērķis ir atklāt saikni starp objektiem relāciju datu bāzē un ontoloģijas elementiem. Nākotnē ir plānots realizēt aprakstīto metodi, izmantojot JAVA platformu un uzbūvējot sistēmu, balstītu uz ontoloģiju veselības aprūpes jomā.

Андрей Каулиньш, Аркадий Борисов. Построение онтологии по реляционной базе данных

Процесс построения онтологии требует много времени и участия широкого круга специалистов. Реляционные базы данных содержат информацию, которая может быть использована для построения онтологий. В данной работе рассмотрен метод построения онтологии из реляционной базы данных, использующий правила отображения. Метод позволяет построить начальную версию онтологии в автоматическом режиме. Описанный метод не имеет ограничений к реляционной базе данных, поскольку использует стандарты описания данных SQL99. Правила отображения определяют соответствия между объектами реляционной базы данных и элементами онтологии. Для описания онтологии использован язык OWL2. Правила соответствия позволяют строить классы, свойства классов и данных, и экземпляры онтологий. Построение правил отображения является трудной задачей, которая решена для простой модели базы данных. Во многих случаях реляционные базы данных используются для хранения сложных конструкций данных. К таким относятся произвольные объекты, определённые пользователем, коллекции данных. В статье систематизированы правила отображения для различных моделей реляционных баз данных. Рассмотрена система поддержки отображений между реляционными базами данных и онтологиями. Основной задачей такой системы является выявление объектов в реляционной базе данных и построение связей, позволяющих построить онтологию. В будущем планируется реализовать описанный метод на JAVA платформе и использовать систему в области здравоохранения.

- [4] H. A. Santoso, S. C. Haw, Z. N. Abdul-Mehdi, "Ontology extraction from relational database: Concept hierarchy as background knowledge," *Knowledge-Based Systems*, vol. 24, issue 3, Apr. 2011, pp. 457–464, 2010. <http://dx.doi.org/10.1016/j.knosys.2010.11.003>
- [5] M. Li, X.Y. Du, S. Wang, "Learning ontology from relational database", *4th International Conference on Machine Learning and Cybernetics*, 2005, pp. 3410–3415.
- [6] N. Lammari, I. Comyn-Wattiau, J. Akoka, "Extracting generalization hierarchies from relational database: A reverse engineering approach", *Data & Knowledge Engineering* vol. 63, issue 2, 2007, pp. 568–589. <http://dx.doi.org/10.1016/j.datak.2007.04.002>
- [7] R. Leyderman, *Oracle Database Sample Schemas*, 11g Release 2 (11.2), E10831-02, Oracle, 2010.
- [8] ISO/IEC SQL/Framework Standard, ISO/IEC 9075-1:1999, ISO, 1999.
- [9] B. Motik, P. F. Patel-Schneider, B. Parsia, OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax. Available: <http://www.w3.org/TR/owl2-syntax/>, [Accessed Oct. 12, 2014].
- [10] M. Reda Chbihi Louhdi, H. Behja and S. Ouatiq El Alaoui, "Transformation rules for building OWL Ontologies from relational databases", *Second International Conference on Advanced Information Technologies and Applications*, ICAITA-2013, 2013, pp. 271–283.
- [11] I. Astrova, N. Korda and A. Kalja, "Rule-Based Transformation of SQL Relational Databases to OWL Ontologies", *Proceedings of the 2nd International Conference on Metadata & Semantics Research*, 2007, pp. 415–424.
- [12] H. V. Jagadish, T. Nadeau, S. S. Lightstone, T. J. Teorey, *Database Modeling and Design*, 5th ed., Morgan Kaufmann, 2011.

Andrejs Kaulins received the B. Sc. and M. Sc. degrees in 1993 and 2002 from Riga Technical University and the University of Latvia, respectively. Since 2013 he has been studying at Riga Technical University to obtain a Doctoral Degree in Computer Science. Currently major field of study is complex IT system design based on ontologies. E-mail: andrejs.kaulins@rtu.lv ; phone +37126737122.

Arkady Borisov received the Doctoral degree in Technical Cybernetics from Riga Polytechnic Institute in 1970, and *Dr. habil. sc. comp.* degree from Technical Cybernetics from Taganrog State Radio Engineering University in 1986.

He is a Professor of Computer Science with the Faculty of Computer Science and Information Technology, Riga Technical University. His research interests include fuzzy sets, fuzzy logic, computational intelligence and bioinformatics. He has more than 235 publications in the field.

He is a member of IFSA European Fuzzy System Working Group, Russian Fuzzy System and Soft Computing Association, member of the Scientific Advisory Board of the Fuzzy Initiative Nordrhein. E-mail: arkadijs.borisovs@cs.rtu.lv.

Application of Fuzzy Logic for Risk Assessment

Andrejs Radionovs¹, Oleg Uzhga-Rebrov²,^{1,2} *Rezekne Higher Education Institution*

Abstract – Risk assessment is an important task in many areas of human activity: economic, technical, ecological etc. Preliminary data adequacy in risk assessments is carried out on the basis of statistical methods and experts' evaluation on potential losses and probabilities of the event. But in many cases, risk assessment must be carried out under the conditions of lack of initial information or uncertainty of information. For that reason, special risk assessment approaches (methods) are necessary. One of them is the usage of fuzzy logic approach. In this paper, fuzzy logic approach is used to manage this uncertainty in information concerning accidental releases of toxic chemicals at chemical plants. This approach can be used by plant risk advisers in Latvia to make right decisions in the situations where chemical releases can harm not only the environment but also human health.

Keywords – Fuzzy logic, fuzzy representation of knowledge, fuzzy rules, fuzzy set theory, knowledge-based systems.

I. INTRODUCTION

Growing concern connected with environmental changes for the worse and rise in potential risks, which happened due to many kinds of human activities, and usage of new technologies are the reason for a rapid increase in the assessment and management of environmental risks. There are many studies devoted to the analysis of ecological risks, for example, [1] and [2]. In this research, a considerable emphasis is placed on the risk assessment of human lives in the case of chemical release. In this point risk assessment, in principle, can be divided into two parts according to the type of the released substance: liquid and gaseous. The assessment of environmental risks and people's health on the territory of the plant and out of it, because of the accident at the chemical plant and release of gaseous substance will be considered in this article.

Assessment of environmental risk involves identifying the events that imply hazards and assessing the magnitude of their consequences and frequency [3]. A lot of data are necessary for a complete ecological risk assessment. A question then arises: how to obtain the necessary initial data for ecological risk assessment? If sufficient statistical material is available, the task in many cases can be achieved successfully. There are many studies, describing the analysis of ecological risks thanks to estimating the statistical data, e. g., [4]. Real life example can be the assessment of cost of car insurance policy. There is enough statistical information about possibilities of accidents and the cost of cars. Estimated losses caused by car crashes can be easily calculated as average losses related to the previous cases. Based on the obtained evaluations, the risk can be easily estimated and the cost of insurance policy can be determined. Unfortunately, having enough statistical material, while giving evaluation of ecological risks, is mostly an

exception not a rule. Exceptional situations and their rare origin do not allow using an effective statistic system for getting a necessary evaluation. To overcome a lack of objective information, expert estimation is often used. Experts-specialists based on their professional knowledge, experience and, sometimes, intuition can provide the necessary data.

Many methods are developed for obtaining and using uncertain probabilistic evaluations such as interval probabilities, second order probabilities etc. The shortcomings of these methods are: their complexity and bad interpretability of the uncertain results. In 1965, L. Zadeh in his work [5] proposed a principally new conceptual basis for dealing with imprecise information – a fuzzy set theory. The theory has widely been developed during the past years. Nowadays, fuzziness is used practically in all fields of scientific and practical activities, including risk assessment.

II. FUZZY LOGIC AND FUZZY SETS

These days there are a lot of textbooks and reference books available on the theory of fuzzy sets and fuzzy logic. Work [6] can be used as an example, but a more detailed and complete description of the theory can be found in [7]. Fuzzy sets theory presents a number of mathematical principles for knowledge representation based on degrees of membership.

The definition “fuzzy logic” has appeared recently. That was in the sixties with the 1965 proposal of fuzzy set theory by Lotfi A. Zadeh, when he published his work “Fuzzy Sets” in order to provide a model for inexact and not precise concepts and subjective judgments. Fuzzy sets deals with degrees of membership to a certain class and degrees of truth. Fuzzy logic uses fuzzy rules that should receive desired results from input linguistic data or variables. The linguistic variable – a variable whose values are sentences in a natural language. For example, a linguistic variable “temperature”, which can contain values: cold, hot etc.



Fig. 1. (a) Boolean logic.

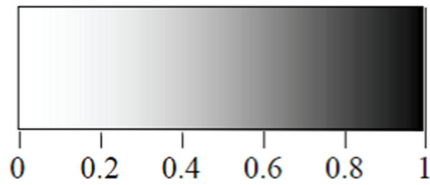


Fig. 1. (b) Multi-valued logic.

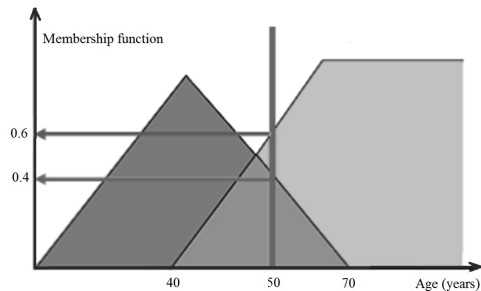


Fig. 2. Fuzzy age-classification scheme.

The advantage of fuzzy methods is their reflection in the human mind and in its ability to store and process information that is uncertain, imprecise, and resistant to classification [8]. Fuzzy logic is an alternative to the classical logic where every proposition must be either “1” (true) or “0” (false). Instead, fuzzy logic asserts that things can be simultaneously “true” and “false”, with a certain membership degree [9]. For example, in ordinary Boolean logic a 50 year-old person can be old or young depending on the scale. But in fuzzy logic (see Fig. 2) a man can be simultaneously young with degree of membership of 0.4 and old with 0.6. Fuzzy set theory provides a way to use imprecise and uncertain information generated by the system and human judgments in a precise way. When the environmental data available do not provide a proper statistical treatment, fuzzy approaches can solve this problem, since they work well for addressing poorly characterized parameters and linguistic variables [10].

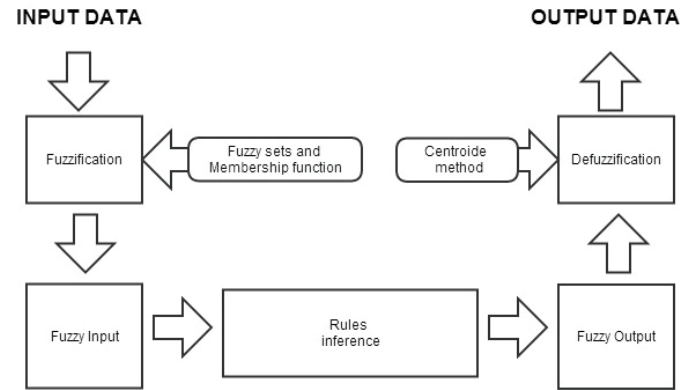


Fig. 3. Representation of the fuzzy methodology.

Fuzzy logic also can merge different kinds of parameters (e. g., environmental, health), quantitative and qualitative values.

III. FUZZY PROCESS

As it can be seen in Fig. 3, a suggested fuzzy model consists of different steps. First of all, the parameters of input and output must be defined and then converted from the original values into the linguistic categories by creating fuzzy sets for each of them. This process is called fuzzification. Secondly, a set of rules must be established. These rules will allow going from the input to the output. In this paper, the Mamdani's methodology will be used. Next, the obtained value must be defuzzified in order to give an understandable risk assessment. This process is called defuzzification. Finally, an output is obtained which is directly related to a certain category of risk. All these steps are carried out using the fuzzy toolbox that is present in Matlab and they are explained further.

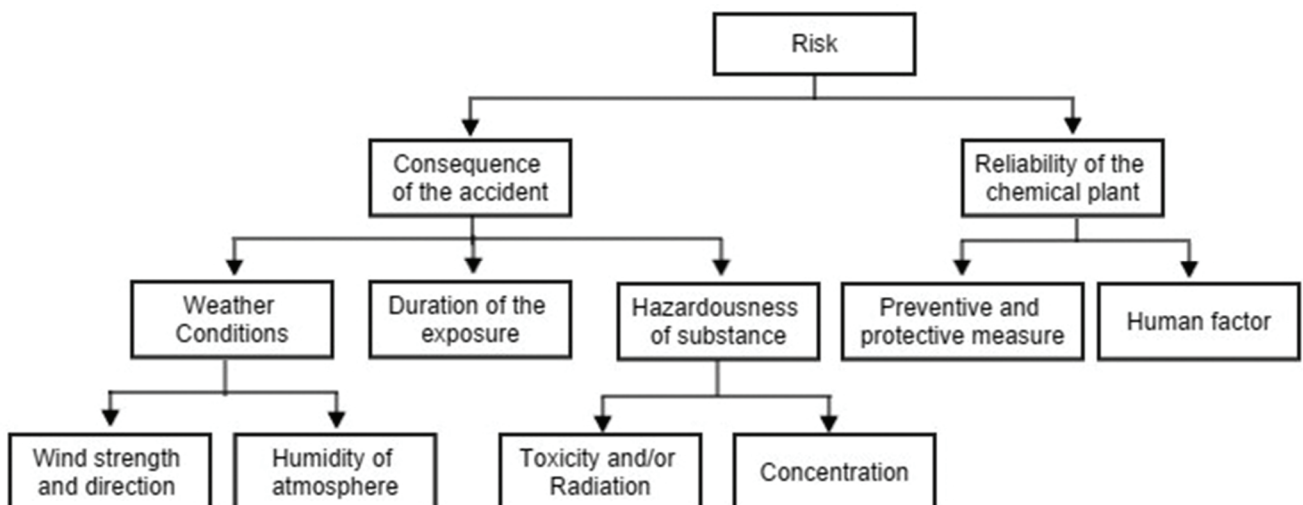


Fig. 4. Risk assessment scheme for releases of gas substances in chemical plants.

A. Input Definition

As we can see in Fig. 4, there are many factors, which participate in the risk assessment of release of chemical substances at a plant. At a minimum, two variables must be defined in order to achieve a proper risk assessment: “consequences of the accident” and “reliability of the chemical plant”. Each of these variables depends on other factors, which in their turn depend on the following factors. Having a big number of variables, for better risk assessment it is needed to have rather much information, which we do not have at the moment or it might be expensive to get, because of timing, finances and other costs. For example, there are many parameters, such as “concentration of substance”, “duration of the exposure”, “weather conditions” etc., which can be found while using special measures, also such things as “probability of accidents”, “human factor” cannot be measured with the help of special instruments. Therefore, the fuzzy model was rearranged in order to be based on a reduced number of input variables: in practice, all sub-variables (as it is shown in Fig. 3) have been grouped into two main variables. The fact that the information necessary by fuzzy logic is mostly more qualitative than quantitative, makes this method a good candidate to be used in order to perform the preliminary estimation of ecological risks. Besides, making a fuzzy model, in this case chemical plant managers will help to use methodology in an easier and more available and understandable way. As it is shown in Tables I and II, fuzzy sets were made for all inputs.

TABLE I
CRITERIA OF RELIABILITY OF THE CHEMICAL PLANT

Reliability of the chemical plant (V)	Criterion
H – High	- The plant fully corresponds to the safety requirements - The staff has been taught
M – Medium	- In general, the plant corresponds to necessary safety conditions - The level of the staff’s preparation is not enough
L – Low	- The plant does not correspond to necessary safety requirements - The staff has no instruction

TABLE II
CRITERIA OF THE CONSEQUENCES OF THE ACCIDENT

Consequence of the accident (G)	Criterion
A	In case of release of chemicals, there is a high probability (50%) that it will lead to the fatal result – human deaths
B	In case of release of chemicals, injury to people might have a low probability of fatal result
C	In case of release of chemicals, minor damage to people’s health may happen
D	In case of release of chemicals, there is a small probability that it will bring only minor damage to people’s health
E	In case of release of chemicals, no damage to people’s health will be observed

B. Output Definition

The main result of this fuzzy model is the definition of category of risk. The function of risk in the given example is presented in the form of equation (1).

$$R = GV \quad (1)$$

where G : consequences of the accident, V : reliability of the chemical plant.

As it is shown in Table III, for output fuzzy sets were created.

TABLE III
CRITERIA OF RISK LEVEL

Risk level (R)	Criterion of evaluation
V – Very High	A high probability of fatal result at the plant and nearby territory
IV – High	A low probability of fatal result, but a high probability of damage to people’s health
III – Medium	A slight probability of health damage and fatal result
II – Low	No risk of death, but slight probability of health damage
I – Very Low	No risk of death and health damage

C. Fuzzification

In the *fuzzification* process, the membership functions defined on the input and output variables are applied to actual values, to determine the degree of truth for each rule. In this paper, the Gaussian type of membership function was selected. The fuzzy sets and their membership function are demonstrated in Fig. 5(a), (b), (c) for each variable used in the fuzzy risk assessment.

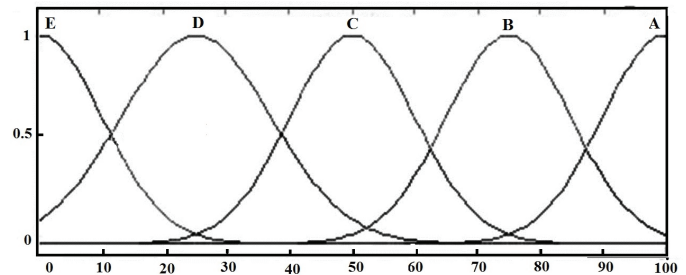


Fig. 5. (a) Graphs of membership functions of consequence of the accident.

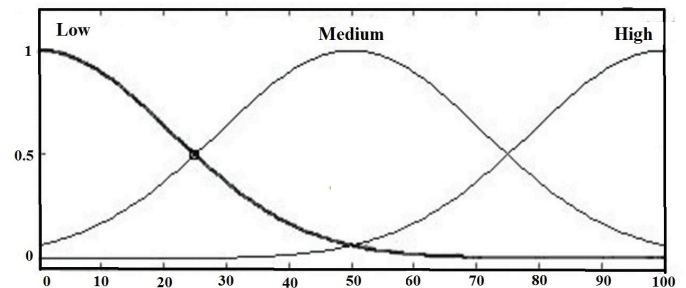


Fig. 5. (b) Graphs of membership functions of reliability of the chemical plant.

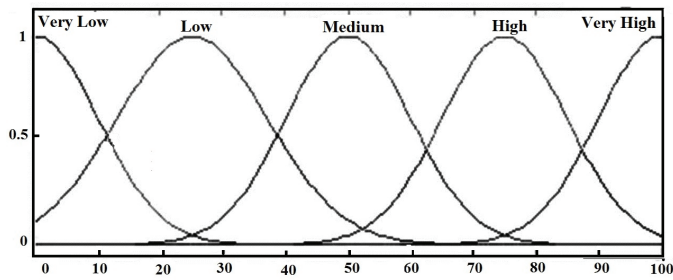


Fig. 5. (c) Graphs of membership functions of risk.

D. Setting up the Rules

Fuzzy logic is a decisional system based on linguistic rules [11]. Therefore, a next step is to connect fuzzy sets by rules. One type of fuzzy inference rule is called the max-min inference rule (Mamdani's methodology).

The most used rules are: "if X and Y then Z ", where X and Y is the premise and Z is the consequence. As an example: "if *reliability of the chemical plant (V)* is *High* AND *consequences of the accident (G)* is "In case of releases of chemicals, there is no damage to people" THEN *Risk level (R)* is *Very Low*".

All the rules can be seen in Table IV. For example: if $G=A$ and $V=M$, then $R=IV$.

TABLE IV
A SYSTEM OF RULES OF FUZZY INFERENCE

Consequences of a possible accident (G)	Reliability of the chemical plant (V)		
	H	M	L
A	III	IV	V
B	III	III	IV
C	II	III	III
D	II	II	III
E	I	II	II

E. Defuzzification

The *defuzzification* process, that is, the conversion of the fuzzy output set (represented as a surface as is shown in Fig. 6 and Fig. 7) to a crisp number, can be done by using different methodologies (Max, mean-max membership principle, centroid method, medium, weighted average method, center of the largest area method and other methods) [12]. In this paper, the most common and widely used *defuzzification* method has been selected: centroid method. The results can be seen at the top of Fig. 6. According to the centroid method or center of gravity method, the crisp value of the output variable is computed by finding the variable value of the center of gravity of the membership function for the fuzzy value. If you think of the area as a plate of equal density, the centroid is the point along the horizontal axis about which this shape would balance. More complete description of the algorithm can be found in [13]. For example, as it is shown in Fig. 6, if parameter *Consequence* (*consequences of the accident (G)*) is 30 units, but parameter *Reliability* (*reliability of the chemical plant (V)*) – 15, then Risk level will be 43.

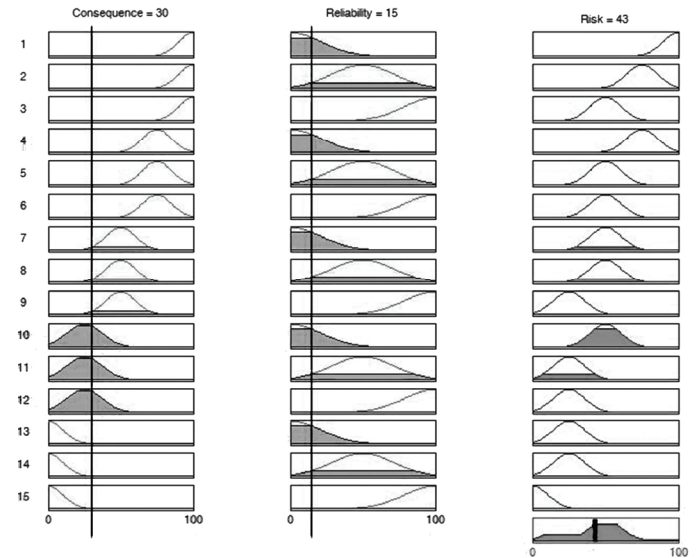


Fig. 6. Application of the centroid method to the inputs.

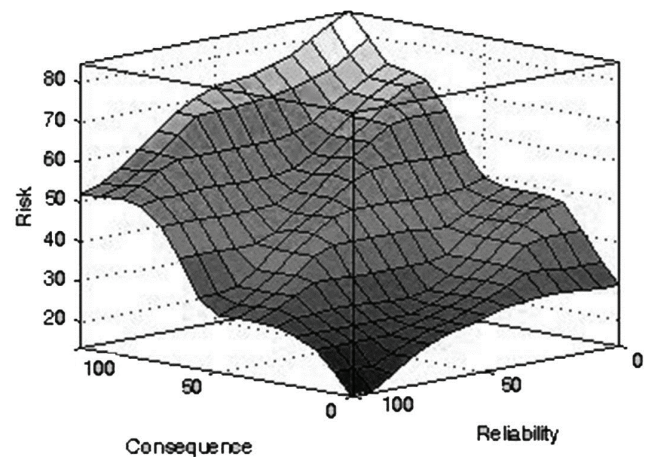


Fig. 7. Three-dimensional representation of the risk, consequence of the accident and reliability of the chemical plant.

IV. CONCLUSION

The application of the fuzzy logic to the assessment of risk of environment and people's health both at the plant territory and outside it in case of potential accident and releases of gaseous substance allows making a complex algorithm of analysis more affordable in order to obtain a risk assessment given incomplete and reduced input data.

The main variables of the model are as follows:

- Consequences of the accident;
- Reliability of the chemical plant.

For input and output variables, membership functions have been established. A set of rules to link input and output variables has been set up and from there by the *defuzzification* method a crisp value can be obtained. The use of this model allows for a realistic preliminary assessment of the risk of accidental chemical releases. But, it is also important to highlight the model's user-friendly tool design and the

effectiveness of achieving very appropriate results in a short time period.

The methodology can be used not only by public authorities but also by the same plant managers, since it is a method that allows for the evaluation of the risk level of the site and also it enables one to see whether the safety measurements are suitable. This application can be used as a preliminary risk assessment tool, able to highlight critical situations and the need for more in-depth and complete analysis; it can also be used in the case of necessity to be able to make a thoughtful decision to reduce a risk level.

REFERENCES

- [1] B. Hennig, L. Ormsbee, C. J. McClain, B. A. Watkins, B. Blumberg, L. G. Bachas, W. Sanderson, C. Thompson, W. A. Suk, "Nutrition can modulate the toxicity of environmental pollutants: implications in risk assessment and human health," *Environmental Health Perspectives*, vol. 120 (6), pp. 771–774, Jun. 2012. <http://dx.doi.org/10.1289/ehp.1104712>
- [2] K. R. Smith, "Comparative environmental health assessments," *Annals Of The New York Academy Of Sciences*, Ann. N. Y. Acad. Sci., Oct. 2008, vol. 1140, pp. 31–39. <http://dx.doi.org/10.1196/annals.1454.032>
- [3] J. K. Lein, "Expressing environmental risk using fuzzy variables: a preliminary examination," *Environ Prof.*, 14, 1992, pp. 257–267.
- [4] B. Martin, A. Pearson, B. Bauer, *An Ecological Risk Assessment of Wind Energy*, The Nature Conservancy Helena, Montana, pp. 146, 2009.
- [5] L. Zadeh, "Fuzzy Sets" *Information Control*, 1965, 8, pp. 338–353. [http://dx.doi.org/10.1016/S0019-9958\(65\)90241-X](http://dx.doi.org/10.1016/S0019-9958(65)90241-X)
- [6] M. Ganesh, *Introduction to Fuzzy Sets and Fuzzy Logic*. PHI Learning Pvt. Ltd, 2006, 256 p.
- [7] G. J. Klir, Y. Bo, *Fuzzy Sets and Fuzzy Logic*. Prentice Hall, 1995, 592 p.
- [8] T. E. McKone, A. W. Deshpande, "Can Fuzzy Logic Bring Complex Environmental Problems into Focus?" *Environ. Sci. Technol.*, vol. 39 (2), pp. 42A–47A, 2005. <http://dx.doi.org/10.1021/es0531632>
- [9] L. A. Zadeh, "The Role of Fuzzy Logic in the Management of Uncertainty in Expert Systems," *Fuzzy Sets and Systems*, vol. 11, issue 1–3, pp. 197–198, 1983. [http://dx.doi.org/10.1016/S0165-0114\(83\)80081-5](http://dx.doi.org/10.1016/S0165-0114(83)80081-5)
- [10] L. Zadeh, "Fuzzy logic and approximate reasoning" *Synthese*, vol. 30, 1975, pp. 407–428. <http://dx.doi.org/10.1007/BF00485052>
- [11] R. E. Bellman, L. A. Zadeh, "Decision making in a fuzzy environment," *Management Science*, vol. 17, no. 4, pp. B141–B164, 1970. <http://dx.doi.org/10.1287/mnsc.17.4.B141>
- [12] G. J. Klir, and B. Yuan, *Fuzzy Sets, Fuzzy Logic and Fuzzy System*, World Scientific, Singapore, 1996, p. 840.
- [13] T. J. Ross, "Properties of membership functions, fuzzification, and defuzzification" in *Fuzzy Logic with Engineering Applications*, 3rd ed., John Wiley & Sons, Ltd, Chichester, United Kingdom, 2010, pp. 89–116. <http://dx.doi.org/10.1002/9781119994374.ch4>

Andrejs Radionovs is a Doctoral Student at Rezekne Higher Education Institution (Latvia). He received the M.Sc. degree in Computer Science from Daugavpils University (Latvia). His research interests include information systems, fuzzy information, uncertain information and in particular, fuzzy set theory and its applications in risk assessment.
Address: Rezekne Higher Education Institution, 90 Atbrīvošanas Av., Rezekne LV-4600, Latvia.
E-mail: a.radionovs@gmail.com.

Oļegs Uzhga-Rebrov is a Professor at the Faculty of Economics, Rezekne Higher Education Institution (Latvia). He received the Doctoral degree in Information Systems from Riga Technical University in 1994. His research interests include different approaches to processing incomplete, uncertain and fuzzy information, in particular, fuzzy set theory, rough set theory, as well as fuzzy classification and fuzzy clustering techniques and their applications in bioinformatics.
Address: Rezekne Higher Education Institution, 90 Atbrīvošanas Av., Rezekne LV-4600, Latvia.
E-mail: ushga@ru.lv.

Andrejs Radionovs, Oļegs Uzhga-Rebrovs. Izplūdušās loģikas pielietojums risku analizē

Riska novērtēšana ir svarīgs uzdevums daudzās cilvēku darbības jomās: ekonomiskajā, tehniskajā, ekoloģiskajā u.c. Pēdējo gadu laikā pieaugošais uztraukums par ekoloģiskā stāvokļa pasliktināšanos un potenciālo risku paaugstināšanos, kas ir saistīts ar daudzām cilvēka aktivitātēm un jauno tehnoloģiju izmantošanu, rada palielinātu interesi par ekoloģisko risku novērtēšanu. Ja ir pietiekams daudzums pamatinformācijas, tad sākotnējo riska novērtējumu var balstīt uz statistikas metodēm. Diemžēl ekoloģiskā riska novērtēšanai trūkst statistiskās informācijas. Tādējādi riska novērtējums jāveic informācijas trūkuma, nepietiekamības vai nenoteiktības apstākļos. Vēsturiski pirmais nenoteiktības veids, kuram bija izstrādāti teorētiskie un praktiskie pamati, bija gadījuma notikumu iestāšanās izredzes (varbūtējā nenoteiktība). Pašlaik varbūtības teorija nodrošina spēcīgu matemātisko aparātu nenoteiktības vadībai dažādās cilvēku darbības jomās. Taču varbūtības teorija un statistika nevar palīdzēt atspoguļot izplūdušus subjektīvus priekšstatus par relevanto reālo lielumu nenoteiktību (izplūdušās nenoteiktības). Lai samazinātu trūkstošās objektīvās informācijas ietekmi, bieži izmanto ekspertu vērtējumu par iespējamem zaudējumiem un notikuma varbūtību. Eksperti, pamatojoties uz savu kompetenci, pieredzi un dažreiz intuīciju, var sniegt datu analīzei nepieciešamos datus. Ir izstrādātas dažādas metodes izplūdušo varbūtību izmantošanai un aprēķināšanai. Pie šīm metodēm pieder: intervālu varbūtības metode, otrās kārtas varbūtības metode, izplūdušās loģikas metode utt. Šajā rakstā izplūdušās loģikas metode tiek pielietota ekoloģiskā riska analizē, gadījumā, ja notikusi toksiskas vielas noplūde ķīmiskajā rūpnīcā. Izplūdušās loģikas pieeju var izmantot ķīmisko rūpnīcu konsultanti riska jautājumos, lai pieņemtu nepieciešamos lēmumus tādās situācijās, kad izplūdušās ķīmiskās vielas var būt bīstamas ne tikai videi, bet arī cilvēku veselībai.

Андрей Радионов, Олег Ужга-Ребров. Применение нечеткой логики для анализа рисков

Оценка рисков является важной задачей во многих областях человеческой деятельности: экономической, технической, экологической и т. д. Растущее беспокойство по поводу ухудшающегося состояния окружающей среды и возрастание потенциальных рисков, связанных со многими видами деятельности человека и использованием новых технологий, вызывают стремительное увеличение интереса к оценке и управлению экологических рисков. В случае наличия достаточной исходной информации предварительная оценка риска осуществляется на основе статистических методов. К сожалению, наличие достаточного статистического материала в оценке экологических рисков является скорее исключением, чем правилом. В таком случае, оценка риска должна проводиться в условиях отсутствия, недостаточности или неопределенности информации. Исторически первым типом неопределенности, для которого были разработаны теоретические и практические основы, были шансы наступления случайных событий (вероятностные неопределенности). В настоящее время теория вероятности предоставляет мощный математический аппарат для управления неопределенностью в различных областях человеческой деятельности. Для снижения влияния недостатка объективной информации часто используются оценки экспертов о потенциальных потерях и вероятности события. Эксперты на основе своих профессиональных знаний, опыта и, иногда, интуиции могут представить необходимые для анализа данные. Разработаны методы для получения и использования неопределенных вероятностных оценок, к ним относятся: интервальные вероятности, вероятности второго порядка, метод нечеткой логики и т. д. В данной статье описан процесс анализа экологического риска методами нечеткой логики на примере утечки токсических выбросов на химическом заводе. Заявленный подход на основе нечеткой логики может быть использован консультантами по оценке риска на химических заводах для принятия необходимых решений в ситуациях, когда химические выбросы могут нанести вред не только окружающей среде, но и здоровью людей.

Knowledge-based System Design Based on Generic Method Conception

Henrihs Gorskis¹, Arkady Borisov^{2, 1,2} *Riga Technical University*

Abstract – This paper proposes an approach to generic object concepts for problem solving method ontologies and knowledge-based systems. By using placeholders, whose main purpose is to define characteristics of the objects upon which the method acts without specifying a domain, method ontologies can become more reusable. The re-usability stems from the ability to merge such method ontologies more easily with fitting domains. The generic object concepts in the method are proposed to be used as a bridge to the domain ontology. The paper provides a comparison to similar methods in related studies.

Keywords – Domain ontology, intelligent system, problem solving method, reusable components.

I. INTRODUCTION

An intelligent system based on storing knowledge in the form of ontologies requires a way of using that knowledge for the purpose of solving a problem. This solution is a problem solving method. A system can be required to solve several different problems using a number of problem solving methods. This paper tries to define a way of reusing existing problem solving methods with various domain descriptions given in the form of ontology models. The main idea is to provide a detailed description of a system in which it would be possible to use several different domain ontology models. Each model would contain a different description of a domain. The system would provide a way of combining almost any domain ontology with a method ontology model. In this paper, we try to define what information would be necessary in the given ontologies and what information it would be possible to obtain from such a combination of two models. For the purpose of reuse, it is necessary to use both ontology models as distinct and different models in the intelligent system. However, during the steps of solving a given problem, these two models would be used as one. The domain ontology is a model that describes all concepts of a certain domain. The method ontology in its turn describes all concepts required for the execution of the method. The method ontology should be as far removed from any domain specific knowledge as possible and the domain ontology should be self-sufficient and descriptive.

II. RELATED WORK

There are several papers that propose similar systems and approaches of using methods and problem solutions with knowledge from an ontology. This paper was heavily influenced by another paper – “Reusable Ontologies,

Knowledge-Acquisition Tools, and Performance Systems: Protégé-II Solutions to Sisypheus-2”. The paper describes a knowledge-based system that configures elevators and uses the Protégé-II architecture for this purpose. The paper proposes the use of both application ontology and mapping rules [1]. In doing so, it created the third ontology, which is specific to both the domain and the method ontology. It also keeps track of the mapping that was used in creating the ontology. Even though the paper provides an almost finished framework for a knowledge system that provides the ability for reuse, there is still room for improvement. This paper proposes certain alternatives to the approaches presented in the mentioned paper.

The proposal of separating general domain knowledge from specific tasks and the difficulties of doing this are not new [2]. Modular approaches to ontology building, which would allow for reuse, are also proposed [3]. Another approach to this problem is to build a new ontology from existing ones [4]. However, there still seem to be no specific solutions for all problems associated with reusing knowledge from ontology models [5], [6].

The main problem is that by changing the context of the problem any domain knowledge can be required to change with it [2]. During creation of a domain ontology, the context of it is the requirement to be descriptive. When a specific problem needs to be solved, the context changes and another structure for the ontology may be more fitting. This can be achieved by mapping knowledge from several ontology models [7].

The approach described in the clinical context based flexible workflow (CONFlexFlow) showed how to integrate information about pathways into an ontology-based system [8]. It also provides an approach of using method specific knowledge in the form of Jess rules with the domain ontology. However, the given approach is very reliant on the existing information from the domain ontology and, therefore, less reusable with other domain ontologies.

The paper “A Method-Description Language” provides a language that is capable of describing method ontologies and can be used for creating a “reuse library” [9]. The ideas in that paper have similarities to the approach proposed in this paper, especially the notion of creating and using a library that contains solutions to problems. The described language is useful for mapping ontologies to each other, by providing Meta information for the mapping process.

III. DOMAIN ONTOLOGY

The domain ontology is used as storage for domain knowledge. It holds descriptions given for concepts that are important in the domain. By describing the concepts of the domain, the relationships and properties between them, the description of the domain itself is also obtained. The domain ontology is usually created by a domain expert. Using the structure of the concepts and by defining restraints and other relationships, the expert creates a framework for the domain knowledge. Once the basis for the ontology is created, the expert fills the ontology with instances, which are the domain knowledge. The class model structure is a meta-layer for the knowledge, while the instances are the knowledge. The domain ontology usually features the description of objects that are important in the domain. For example, forest ontology will have concepts for trees and other vegetation, which can be separated into several different types. The domain ontology describes the domain in such a way that it is possible to obtain knowledge about it. Fig. 1 shows an abstract illustration of the domain ontology. The concepts and relationships are important parts of the knowledge about the domain. However, the domain ontology itself is not designed for solving or finding any solutions to problems. This is different from the method ontology, which provides a description of action.

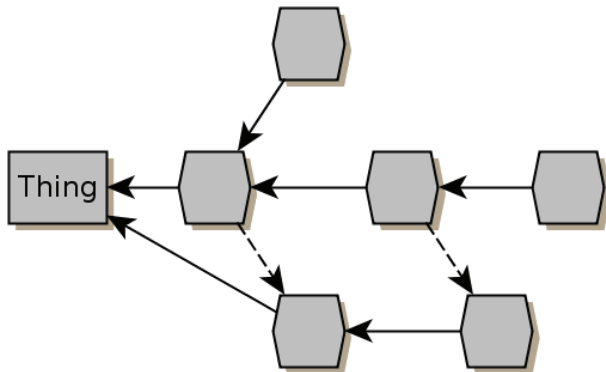


Fig. 1. Generalized visualization of a domain ontology.

IV. METHOD ONTOLOGY

The method ontology differs from the usual domain ontology. Instead of providing a description of a certain domain, that can be associated with any real world thing, it provides information about a method or action. The method ontology contains concepts, which are important for executing the provided method. Very important parts of the method ontology are concepts that describe the input and output of the method. The structure and properties of these classes provide the information necessary for preparing the method and provide information about the outgoing information that is created as a result of the execution of the method. Additionally, the method ontology contains concepts of things necessary during the execution. Those can be descriptions of variables and smaller actions, which are required for keeping track of changes. The method ontology describes these concepts in an abstract way, which is as far removed as

possible from any specific domain. The concepts it describes need to be able to transmit all the information necessary for using these concepts. This includes the hierarchy, properties and restrictions of any concepts that are taken from domain ontology for the specific purpose of using them with the method.

Fig. 2 shows an abstract visualization of a method ontology.

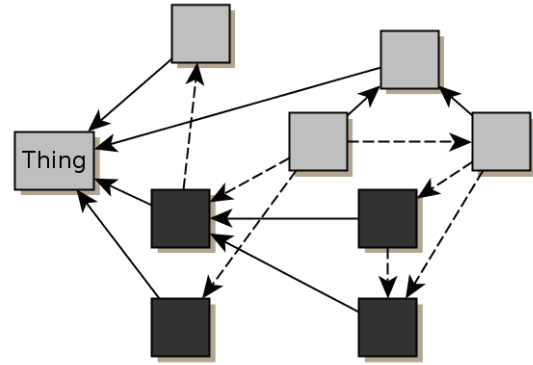


Fig. 2. Generalized visualization of a method ontology.

Note that in Fig. 2 the concepts of the method ontology are separated into two kinds. The light gray blocks symbolize concepts that are internal. They are used to describe actions and other objects and processes that are only used in the description of the method and are in no way connected to any other domains. The dark blocks symbolize objects of the method. The objects of the method are concepts the method acts upon. Their description is important since it is necessary to explain the hierarchy and properties of the things that are being used in the method. However, these objects should be described only in abstract terms. They will later be used for merging with a domain ontology that holds the information of actual concepts.

The method ontology is provided in addition to the method body. The method body is the description of the method itself. In contrast to the method ontology, it does not provide knowledge about the method, but is the pure actions of the method. The method body can be, for example, given as commands in a programming language. It can also be a general description of the actions and the order of the actions that have to be taken to solve a specific problem. Without the method ontology, the method body would not be understandable, and without the method body, the method ontology would provide a description of a method, but could not provide the order of actions.

V. GENERIC OBJECT CONCEPTS

A method is a description of actions, which are performed on some object or thing. For example, a sorting method describes how to put differently sized objects into order. A general description of such a method does not care about the object, which is being sorted. However, for the sorting method to work, the object which is sorted has to have a size, which can be compared, in order to determine the position of the

object in the list of sorted objects. The method ontology can, therefore, provide a general description of an object that is being sorted. This object has to have a property, which tells the size of the object. The method ontology can also provide several descriptions and a hierarchy of these objects. For example, a method ontology that describes the process of sorting some objects can provide the following information:

- The method sorts objects;
- These objects have a property or slot that provides information about the size of the object;
- There are small objects, which are a sub-concept of object and have a small size;
- There are medium objects, which are a sub-concept of object and have a medium size;
- There are large objects, which are a sub-concept of object and have a large size.

If the ontology provides a hierarchy of sub-objects, it can use it for its explanatory powers. A sub-object can be used in the description just like any other object, but it provides its unique properties. However, from the information above, we can see that such a description needs to be finished before it can be used. A generally described method will not provide fixed information about what size objects are to be considered small or large. This information has to be provided at the time, when a domain ontology is connected to the method. Different domains will have different ideas about what size objects are small or large.

Once the objects that are used in the method are defined, the method can provide a general explanation of how it uses these objects, what information is required at the beginning of the method, and what information is returned by the end of the method execution. The inputs and outputs of the method can also be generic objects, but they do not have to be. Let us take a look at an example. Let the example use an ontology that describes different foods. Food ontologies are common and are also used in medicine [10], [11] for the purpose of describing the different kinds and properties of food a person may consume. The example contains:

- The method ontology, which describes a sorting method. It has its generic object structure of small, medium and large generic objects that are used in the sorting;
- The domain ontology, which describes the food domain. It contains concepts like “Nut”, “Fruit”, “Bread” and others. These concepts have individuals like “Walnut”, “Hazelnut”, “Apple”, “Pear”, “White Bread” and others.

Concepts from the food ontology have to be mapped to the generic objects of the method ontology. This can be done by providing the information about the sizes of the foods and the category sizes. Or the information about the sizes can be skipped (as long as that information is not vital to the method itself) and the concepts can be mapped directly by the knowledge expert. The sorting method has to provide the description of the input. In this case, the method input is a list of generic objects (or sub-concepts of the generic object since

they can be viewed as the same). From this it is clear that a list of food is the input for the method. The method sorts the objects by their size and returns a sorted list of the same food as the result of the method.

Mapping domain object to these generic method objects should be easier if the method ontology can provide a clear description of what is expected from the concepts that are mapped to the generic objects.

VI. THE RELATION BETWEEN THE GENERIC OBJECT CONCEPTS AND THE DOMAIN CONCEPTS

The mapping and usage of these generic objects can be simplified if a new kind of relation is introduced to the ontology description of the concepts. This relation is “of functional equality”. This relation would signify to any user or process that uses the ontology description of a concept that any function, action or other process has to perform the same action with the related concept. If we have a relation of functional equality between the concept “apple” and “medium sized object” and a process has to check the size of the concept apple, the result should be that the apple is medium sized, even if that information is provided in the concept “medium sized object” and not in “apple” since both are functionally equal for the purpose of the method. Such an effect can be achieved by using deductive reasoning or by implementing a process of equal functionality (polymorphism). A similar effect can be achieved by using a simple “is a” relation, but that approach would make it necessary to perform additional reasoning, which has to be defined specifically for the connected concepts. Fig. 3 shows a connection between two concepts from two different ontologies, which have to be functionally equal for the purposes of the method.

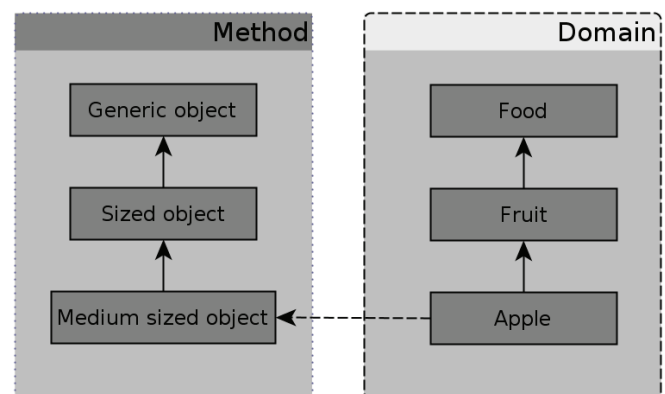


Fig. 3. Relations between ontologies.

VII. ONTOLOGY USE WITHIN THE SYSTEM

The descriptions of the two kinds of ontologies that are used in the system given in the previous sections are very important. The system is based on the presupposition that the method ontology provides generic objects that can be mapped to the knowledge of the domain ontology. It is based on the process of merging. Merging is important for explanation purposes, so the method is no longer abstract, but describes actions and objects of a certain field. During merging it is

necessary to define which abstract method concepts are equivalent to some of the domain concepts. The working of the system can be described with the following steps:

Step 1 – Obtaining the ontologies: The domain ontology is the description of the field, in which the problem arose. The method ontology should be picked based on the problem itself. A reuse library as mentioned in another paper [9] can be very helpful for finding fitting ontologies and making sure that they are compatible.

Step 2 – Mapping: Based on the description of the method ontology, every generic object used by the method has to be mapped to one or several concepts within the domain ontology. Fig. 4 shows how method objects are mapped to domain concepts. Note that in a certain situation it required knowledge be missing from the domain ontology. It can be necessary to add missing information to the ontology.

Step 3 – Creation of the application ontology: Similar to the process described in the related paper, an ontology model that holds all the information for the given task is created [1]. Based on the mapping information and any added knowledge, an application ontology is created. Fig. 5 shows the new ontology that resulted from the combination of the domain and the method ontology. The dotted block symbolizes an added concept. The light and dark octagons symbolize the concepts that resulted from the combining method and domain knowledge. It is important to note that these concepts keep all the properties from the concepts they were created from.

Step 4 – Problem solving: Using the application ontology and the method body, it is possible to solve the problem, in the context of the provided domain. This is possible since the concepts used in the method body were connected to the method ontology. By creating the application ontology, any missing information (the inputs) and the context of the problem (the domain) are provided to the method. The actual execution of the method is task for the system. This can be done either by using an API (Application Programming Interface) created for this task, which is capable of looking up the required information from the connected ontology or by generating the executable code by adding the information from the ontology to the method template. An API, for the purpose of using OWL (Ontology Web Language) already exists [12], but it would require additional functionality as described, in order to perform the required actions.

Step 5 – Explanation of the solution: Any solution found by the method can, at this point, be explained further by using the knowledge of the ontologies. If the user of the system requires more information about the solution, the system can provide it, by using connected concepts or by reasoning. Since the solution is described by using the application ontology, the descriptive powers of the ontology can be used.

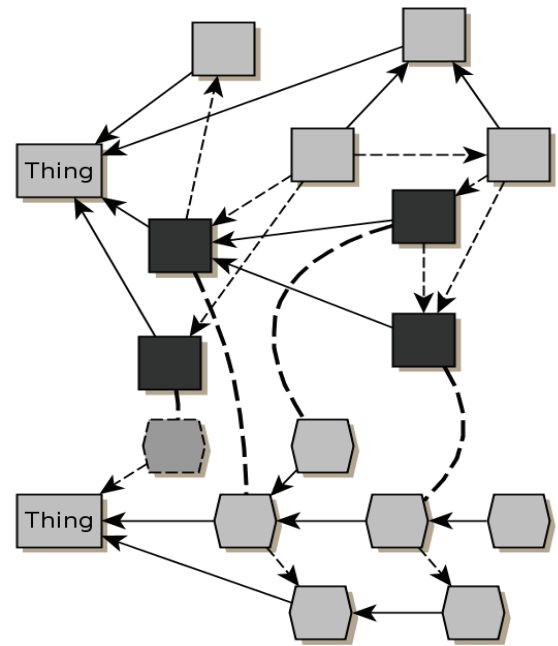


Fig. 4. Mapping the method ontology to the domain ontology.

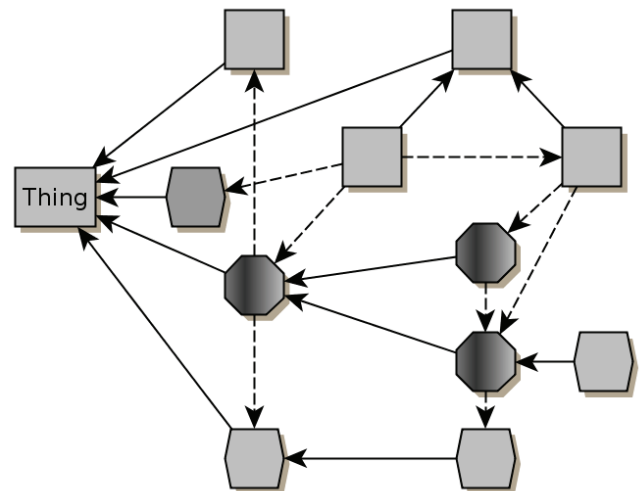


Fig. 5. The application ontology.

The main idea of this approach is the use of method objects in the description of the method ontology. By using these abstract objects, it is easier to define and perform mapping and merging of the two ontology models.

VIII. CONCLUSION

By combining method and domain ontology, it is possible to reuse the generally defined methods with any fitting domain ontology. This way we can reuse knowledge to solve specific problems and provide explanations for the solutions. The proposed is an approach for a knowledge-based system that can use ontologies in the described way. It is important to note that this approach is heavily based on the way the method ontology is created. The reusability of ontologies is achieved by having method object descriptions in the method ontology.

If the method ontology defines all the concepts it needs in order to perform, the mapping process can be simplified, and the method ontology can become very reusable. These generalized method objects provide the hierarchy and properties, which can be used in the method. Once these objects are combined with domain concepts, the new concepts inherit the hierarchy and properties of both. This way a clear interface for mapping or merging between ontologies is provided. This approach of using generalized non-specific concepts in the description of the method can be applied to any of the methods concepts. It is not limited only to the inputs and outputs of the method. If the method uses variables or objects during execution, which rely or can be described in the domain, these objects can be provided in a generalized way for mapping to the domain ontology.

The application ontology created as a result of combining two different ontologies and adding any required information is distinct from the domain and method ontologies. The reuse comes from using the basic ontologies in different combinations. The application ontology is a specific ontology usable only for solving the task it was created for. Only by creating such application ontologies for different situations, by using preexisting ontologies, reuse of knowledge can be achieved.

The ability to describe the obtained solution from the method comes from the idea that the execution of a method is connected to the ontological description of the concepts used in the method. Once the method ontology is replaced by the application ontology, the method has to be capable of using this new knowledge about the concepts that are used in the method. The generic method objects are also renamed and combined with concepts from the domain. This means that these concepts now contain information from two ontologies usable for solving the task. That means that the execution needs to be capable to look up the properties of concepts at runtime. For example, if a method is designed to sort different sized objects by their size, the application ontology can be created with a domain ontology that contains knowledge about fruit. And a certain fruit was mapped to a certain method related object that provides information about size. The method has to be capable of using the concept that describes the fruit as a sized object. At this point, it can become clear if a fundamental difference exists within these ontologies, which would require further work from the domain expert to make such ontologies compatible.

The language used in the method body to describe actions and the exact specifications of the execution are subjects for future work. They use concepts and information from the ontology, but they are not part of the ontology and do not necessarily use the ontology directly.

This paper is part of research in progress to create a specification of an intelligent system that will be capable of reusing ontology knowledge. The approach provided in this paper about reusing methods by creating combinations of different ontologies is part of this research.

REFERENCES

- [1] T. E. Rothenfluh, J. H. Gennari, H. Eriksson, A. R. Puerta, S. W. Tu and M. A. Musen, "Reusable Ontologies, Knowledge-Acquisition Tools, and Performance Systems. Protégé-II Solutions to Sisyphus-2" in *International Journal of Human-Computer Studies*. 2002, pp. 303–332. <http://dx.doi.org/10.1006/ijhc.1996.0017>
- [2] B. Chandrasekaran, "Generic Tasks in Knowledge-Based Reasoning: High-Level Building Blocks for Expert System Design," *IEEE Expert*, vol. 1, Issue 3, 1986, pp. 23–30. <http://dx.doi.org/10.1109/MEX.1986.4306977>
- [3] C. Torniai, M. H. Brush, N. Vasilevsky, E. Segerdell, M. Wilson, T. Johnson, K. Corday, C. Shaffer and M. Haendel "Developing an Application Ontology for Biomedical Resource Annotation and Retrieval: Challenges and Lessons Learned," in *ICBO*, 2011.
- [4] N. K. Leung, S. K. Lau, J. Fan and N. Tsang, "An integration-oriented ontology development methodology to reuse existing ontologies in an ontology development process," in *Proceedings of the 13th International Conference on Information Integration and Web-based Applications and Services*. ACM. December 2011, pp. 174–181.
- [5] D. Lonsdale, D. W. Embley, Y. Ding, L. Xu and M. Hepp, "Reusing ontologies and language components for ontology generation" in *Data & Knowledge Engineering*, vol. 69, no. 4, 2010, pp. 318–330.
- [6] E. Simperl, "Reusing ontologies on the Semantic Web: A feasibility study" in *Data & Knowledge Engineering*, vol. 68, no. 10, 2009, pp. 905–925.
- [7] J. J. Jung, "Reusing ontology mappings for query routing in semantic peer-to-peer environment," in *Information Sciences*, vol. 180, no. 17, 2010, pp. 3248–3257. <http://dx.doi.org/10.1016/j.ins.2010.04.018>
- [8] W. Yao, A. Kumar, "CONFlexFlow: Integrating Flexible clinical pathways into clinical decision support systems using context and rules" in *Decision Support Systems*, vol. 55, no. 2, 2013, pp. 499–515. <http://dx.doi.org/10.1016/j.dss.2012.10.008>
- [9] J. H. Gennari, W. Grosso, M. Musen, "A Method-Description Language: An initial ontology with examples". *Stanford Medical Informatics*. 1998, pp. 1–18.
- [10] J. Cantais, D. Dominguez, V. Gigante, L. Laera, V. Tamma, "An example of food ontology for diabetes control". *Proceedings of the International Semantic Web Conference 2005 workshop on Ontology Patterns for the Semantic Web*. 2005, pp. 1–9.
- [11] J. Graça, M. Mourão, O. Anunciação, P. Monteiro, H. S. Pinto, V. Loureiro, "Ontology building process: The wine domain" in: *Proc. of the 5th Conf. of EFITA*. 2005, pp. 1138 – 1145.
- [12] J. Henß, J. Kleb, S. Grimm, "A Protégé 4 Backend for Native OWL Persistence," *11th Intl. Protégé Conference*, June 23–26, 2009, pp. 1–4.

Henrihs Gorskis is a second-year Doctoral Student majoring in Information Technology at Riga Technical University (RTU). He received his *Mg. sc. ing.* degree in 2013. His research interests include data mining, ontology engineering, evolutionary computing and programming. In 2013 he participated in the AICT conference and also provided a paper on ontology engineering. He is especially fond of the Java programming language and uses it for both work and personal application development.
E-mail: henrihs.gorskis@rtu.lv.

Arkady Borisov received his Doctoral Degree in Technical Cybernetics from Riga Polytechnic Institute in 1970 and *Dr. habil. sc. comp.* degree in Technical Cybernetics from Taganrog State Radio Engineering University in 1986. He is a Professor of Computer Science at the Faculty of Computer Science and Information Technology, Riga Technical University (Latvia). His research interests include fuzzy sets, fuzzy logic and computational intelligence. He has 235 publications in the field. He has supervised a number of national research grants and participated in the European research project ECLIPS. He is a member of IFSA European Fuzzy System Working Group, Russian Fuzzy System and Soft Computing Association, honorary member of the Scientific Board, member of the Scientific Advisory Board of the Fuzzy Initiative Nordrhein-Westfalen (Dortmund, Germany).
E-mail: arkadijs.borisovs@cs.rtu.lv.

Henrihs Gorskis, Arkadijs Borisovs. Uz zināšanām bāzētas sistēmas izstrāde, lietojot vispārīgās metodes konceptus

Intelektuālās sistēmās var būt nepieciešams lietot vairākas ontoloģijas. Ontoloģijas glabā konceptuālās pamatzināšanas šādās sistēmās: lai atrisinātu dažādus uzdevumus ar dažādām pamatzināšanām, ir nepieciešams lietot dažādas ontoloģijas. Šajā darbā tiek piedāvāta pieeja, kas palīdz atkārtoti lietot dažādas ontoloģijas ar vienu un to pašu sistēmu jeb atrisināšanas metodi. Sistēmā tiek glabāta metodes ontoloģija. Tās ir konceptuālās pamatzināšanas par pielietojamo metodi un to, kādā veidā tiek atrisināts uzdevums. Lai atvieglotu metodes ontoloģiju apvienošanu ar ontoloģijām, kas apraksta problēmas nozari, tiek piedāvāts lietot vispārīgus konceptus par objektiem metodes aprakstā. Šiem konceptiem piemīt visas īpašības, kas ir nepieciešamas, lai metode varētu tos lietot problēmas atrisināšanas gaitā. Šādam aprakstam ir jābūt pēc iespējas vispārīgam un nesaistītam ar trešo problēmas apgabalu. Tajā brīdī, kad metodes ontoloģija tiek apvienota ar problēmas ontoloģiju, šie vispārīgie metodes objekti kalpo par apvienošanas punktiem. Savienojot vispārīgo metodes objektu ar konceptu no problēmas ontoloģijas, tiek izveidota saite ar citiem konceptiem no problēmas nozares. Vienlaikus tiek garantēts, ka saistītiem konceptiem piemīt visas īpašības, kas ir nepieciešamas, lai tos varētu lietot metodes izpildes laikā. Kad metodes un problēmas vides ontoloģijas ir apvienotas, tad izveidojas aplikācijas ontoloģija. Tā ir ontoloģija, kas kalpo par pamatu problēmas risināšanas procesā. Izmantojot apvienoto ontoloģiju, ir iespējams uz tās pielietot ontoloģijas iespējas, jauno zināšanu atklāšanai vai esošo faktu aprakstīšanai. Sakarā ar to, ka vispārīgie metodes koncepti tiek savienoti ar citas ontoloģijas konceptiem, šajā solī var atklāties, ka izvēlētais koncepts ir nederīgs, jo var izveidoties pretrunas ar vienu no ontoloģijām. Tiek piedāvāts lietot šos vispārīgos metodes konceptus, lai atvieglotu ontoloģiju apvienošanas procesu un padarītu ontoloģijas atkārtoti lietojamas. Tādējādi ontoloģiju apvienošanas procesā jau iepriekš ir norādīts veids, pēc kura ir iespējams ātrāk apvienot dažādas ontoloģijas.

Генрих Горский, Аркадий Борисов. Разработка системы, основанной на знаниях, используя общие концепты метода

В интеллектуальных системах иногда необходимо использовать несколько онтологий. В таких системах онтология хранит базовые концептуальные знания. Для решения различных задач с разными базовыми знаниями необходимо использовать целый ряд различных онтологий. В данной работе предлагается подход, который помогает повторно использовать разнообразные онтологии с той же системой или методом решения задачи. В системе хранится онтология метода. Она является концептуальной основой знаний, которые применяются в методе и описывают способ, с помощью которого эта задача решается. Для того чтобы облегчить слияние онтологии метода с онтологиями, описывающими проблему в отраслях, предлагается использовать общие концепты объектов, описанных в методе. Эти концепты имеют все качества, которые необходимы методу, для того чтобы они могли быть использованы в ходе решения задачи. Такое описание должно быть по возможности максимально общим и не связанным с третьей проблемной областью. Когда онтология метода объединяется с онтологией отрасли, эти общие объекты метода служат точками объединения. Объединяя общие объекты метода с концептами из онтологии проблемы, создаётся связь с другими понятиями отрасли. В то же время это гарантирует, что соответствующие понятия будут обладать всеми качествами, которые необходимы для использования во время выполнения метода. Когда онтологии метода и окружающей среды объединяются, создаётся онтология приложения. Эта онтология служит основой для процесса решения проблемы. Используя объединённую онтологию, можно применять функции онтологии для открытия новых знаний или описания существующих фактов. В связи с тем, что общие понятия метода связаны с другими понятиями онтологии, на этом этапе может выясниться, что выбранная концепция является недействительной, так как могут появиться противоречия в одной из онтологий. Предлагается использовать эти общие объекты метода для облегчения процесса слияния онтологий и сделать онтологии повторно используемыми. Таким образом, в процессе слияния онтологий уже существовало указание, по которому можно было бы быстрее комбинировать различные онтологии.

Development of Ontological Knowledge Model for Raw Materials Management Task

Darya Plinere¹, Arkady Borisov^{2, 1, 2} *Riga Technical University*

Abstract – Ontology application is widely used for different purposes and areas. This paper presents the difference between taxonomy and ontology, the steps of ontology development and an example of ontology development for raw materials management. Raw materials management is a process that is considered to be part of the overall supply chain management process and focuses on the tasks of securing the highest quality materials at the lowest rates possible, while also working within the company structure to make sure those materials provide the best possible benefit within the production process.

Keywords – Agents, JADE, ontology development, Protégé, raw materials management.

I. INTRODUCTION

Raw materials management is a process that is considered to be part of the general process of supply chain management, and focuses on solving the problems of providing high quality materials at the lowest rates possible, as well as working in the company in order to ensure that these materials provide the best advantage in the production process. Raw materials management begins with the acquisition and evaluation of the raw materials, goes through the use of these materials in the production process, and even includes evaluating the amount of waste that is present when these materials are produced into the finished product [1].

The task of raw materials management begins with the assessment and selection of the materials needed to produce the finished product. The manager devotes time and effort in order to find the right materials, using the company's resources to ensure that they have the appropriate level of quality, and then takes action to purchase the materials in the amounts desired [1].

As part of the purchasing process, the manager also attempts to reach an agreement in the best price possible of raw materials for the use in the production process. Here, the focus of raw materials management moves from the task of finding the right materials and focuses on the use of contractual arrangements complete with a reduced price or volume purchase pricing in order to make sure that the company has an appropriate supply of the right materials for the maintenance of operation [1].

The process of raw materials management is ongoing. At all times, the efforts to provide the proper quality level of raw materials and for the most competitive prices is in progress, this allows the manager to find new suppliers, and possibly to save the company more money on the front end. Always be aware of the options open to the company to provide the necessary materials for production purposes, the business can

minimize any disruptions in the production that could happen if the usual seller is suddenly unable to meet the demand for some reason [1].

There are some reasons why the development of ontology is needed:

- to share common understanding of the structure of information among people or software agents;
- to enable reuse of domain knowledge;
- to make domain assumptions explicit;
- to separate domain knowledge from the operational knowledge;
- to analyze domain knowledge [2].

II. PROBLEM STATEMENT

The field of microelectronics on the example of the company for the production of chips is studied in this paper. Chip production requires timely saturation of crystals in a raw material warehouse. The order of crystals is made by a purchasing department manager right now by telephone, fax and e-mail.

In order to achieve maximum utilization of production capacity and minimize storage and procurement costs, it was decided to use a new method of raw material procurement. The system must meet the following requirements:

1. to ensure continuous production of chips;
2. to perform the analysis of production and sales of chips in order to plan the subsequent production orders and ensure timely delivery of the missing elements.

To ensure continuous production of chips, it is required to make timely coordination between the supplier and the customer of the name and the quantity of component parts, as well as price and delivery time.

III. PROPOSED METHOD

Multi-agent systems are widely used for problems of supply chain management, in particular for the raw materials management task. One of the features of multi-agent system is that it offers a decentralized system. Global behavior of the whole system is determined by the individual behavior of a family of agents who follow their own rules of behavior, exist in a shared environment and interact with the environment and other agents. This makes a multi-agent software system most appropriate for the task. Agents use a common ontology or taxonomy for negotiating, thus allowing agents to understand the substance of the negotiations (Fig. 1).

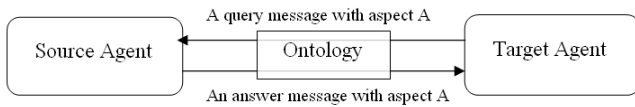


Fig. 1. Agent cooperation using a common ontology.

The implementation of the proposed method requires:

1. the development of an ontology or taxonomy;
2. agent system creation.

A. Taxonomy vs. Ontology

It was necessary to clearly define a taxonomy and ontology and understand whether it would be enough to use the taxonomy for the task.

Taxonomy – a set of concepts presented in a hierarchical structure, covering the domain area. As a rule, the concepts are arranged in a relationship “IS-A”, focusing on the wide to narrow concept, and implies the inheritance of class characteristics of its superclass.

Ontology – a set of concepts and relationships that represent a wide range of knowledge of domain area. In addition to the relationship “IS-A” it has an unlimited number of relationships between concepts, as well as properties with their restrictions and values and logical inference support [3].

The ontology is used for this task solving, not a taxonomy, because the taxonomy involves only the relationship of class-superclass, and in this task, there are such relations as *consists_of*, *has_proper_base* linking classes together. The other types of relationships are used, such as *employees_position*, *documents_name*, *order_status* and others.

B. Ontology Development Step by Step

Ontology is a formal, explicit specification of a shared conceptualization [4].

Formally, an ontology can be defined [5] as the tuple (1).

$$O = (C, H, I, R, P, A), \quad (1)$$

where

C – the set of entities of the ontology - concepts;

H – the set of taxonomic relationships between concepts, which define a concept hierarchy;

I – the set of relationships between ontology elements and its instances.

R – the set of ontology relationships that are neither “kind_of” nor “is_a”;

P – the set of properties of ontology classes;

A – the set of axioms [5].

The development of an ontology consists of:

1. definition of classes or concepts in the ontology;
2. location of classes in the taxonomic hierarchy (a subclass – superclass) if this possibility exists;
3. properties and a description of the allowed values of these properties;

4. identifying the restrictions of property values;
5. instances.

If it is determined that there is no suitable taxonomy/ontology, its development begins with definitions of key concepts used to express the necessary knowledge for a particular business (commercial) activities.

List of concepts for this task: crystal, plate, the number of crystals on the plate, appropriate base, crystal size, the quality of the plate, defect, valid, the coefficient of the output production of the plate, the production, the order of the plate, the return of plate, price, discount, delivery time, number of plates, currency, total cost, document, contract, invoice, specification, license, order status, seller, buyer, buy, sell, exchange, return, contractors, etc.

The next step is the definition of the class hierarchy, (see Table I). Describing terms and concepts, it is necessary to use the following rule:

If class A is a superclass of class B, then every instance of class B is an instance of class A.

TABLE I
CLASS HIERARCHY OF INITIAL ONTOLOGY

Class	Subclass
Crystal	
Plate	
Document	
	Invoice
	Contract
	License
	Specification
Order	
Base	
Company	
	Buyer_company
	Seller_company
Employees	
	Seller_employees
	Buyer_employees
Money	

The definition of the properties of classes is the next step in the development of the ontology. Properties of the class are a distinctive type of information for different instances of the same class.

The domain area’s terms have already been allocated in the classes; most of the remaining terms are likely to be the properties of these classes. It is needed to determine which class it describes for each property in the list. These properties become slots attached to classes.

The class “Crystal” has the following properties: name of the crystal, crystal size, appropriate base (the relationship between the classes “Crystal” and “Base”). The class “Plate” has following properties: plate name, consists of (the

relationship between the classes “Plate” and “Crystal”), coefficient of the output plate, quality of the plate, cost of the plate. The class “Documents” – document name, serial number of the document, participants of document; subclass invoice, contract, license, specification inherit in these properties. The class “Company” has the following properties: company’s name, company’s address, has employees. The property of the class “Base” – name of the base. Class “Employees” has the following properties: name, surname, personal identity number, position, works for. The class “Order” has the following properties: order’s number, order’s contact person, order’s status.

The next step is defining facets of the slots. The slots may have various facets, which describe the type of value, allowed values, the number of values and other properties of the slot. For example, the value of crystal’s name property is one string, i. e., crystal’s name is a slot with value type “String”. Appropriate base slot as values may be instances of the class “Base” and plate’s value slot has type “Float”. Numbered slots determine the specific list of allowed values for the slot. For example, we can establish that plate’s quality slot can take one of two possible values: valid or defect. Numbered slots in Protégé are of type “Symbol”.

Last step is the creation of instances [6].

TABLE II
CLASSES, PROPERTIES AND INSTANCES OF ONTOLOGY

Class	Property	Instances
Crystal, Plate, Document, Invoice, Contract, License, Specification, Base for Crystal, Company, Buyer_Company, Seller_Company, Employees, Buyer_Employees, Seller_Employees, Money	Crystals_quantity, appropriate_base, crystal_size, plate_quality, output_coefficient, order_status, currency, name, surname, address, crystal_name, order_number, consists_of, company_name, etc	Contract 145- 14/1, Invoice 14-10, ABCTrade, Waleo, Crystal Bi005, Base 301.8, Plate H005R, Armands Elksnis, Order 14/10

Horizontal links

The horizontal links [7] are properties or relations, which link individuals or classes that are not in hierarchical relationships. To create the horizontal links, the logical properties were used with one or more arguments. Examples of logical property are the following:

- 1) “Plate consists of crystals”, where plate and crystal are arguments, consists of – a property;
- 2) “Elksnis works for ABCTrade”, where Elksnis and ABCTrade are arguments, works for – a property;
- 3) “Plate has a certain number of crystals” where plate is an argument, has a certain number of crystals – a property.

C. Ontology Development in Protégé

Domain area’s ontology development and editing are implemented in Protégé ontology editor, see Fig. 2. See [2], [8], [9] as a guide of building ontology in Protégé.

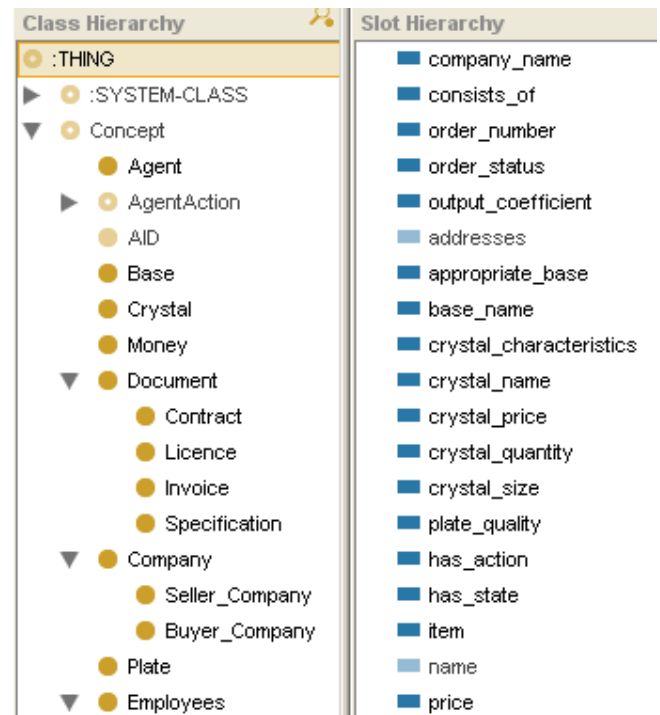


Fig. 2. Ontology development in Protégé ontology editor.

D. The Agent-based System Development

The proposed agent-based system consists of two agents – Seller Agent and Buyer Agent. Buyer Agent orders for the purchase of raw materials, negotiates with the Seller Agent for price and delivery time. The interaction between the agents is via a common knowledge ontology. For the interaction software, agents use the knowledge ontology *Ontology*, common to the agents involved in the negotiations. Negotiation requires knowledge of the size of the crystals, the possible bases and knowledge of the possible alternatives.

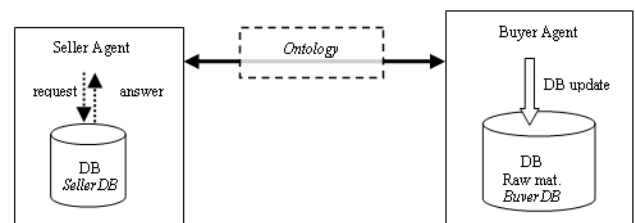


Fig. 3. Agent system with two agents and a common ontology.

The following infrastructure was used for constructing the system: Java and JADE – platform in which agents exist and interact, it gives agents the basic services necessary to their

existence, Protégé ontology editor for ontology development, Ontology Bean Generator to convert the domain ontology in JADE classes, MySQL for database support, Apache Ant to compile the code and NetBeans IDE as an integrated development environment. For the interaction of the agents ACL messages was used [10].

To implement the model, the following steps were taken:

- 1) domain ontology constructed and presented in a form suitable for use in JADE;
- 2) multi-agent system constructed in the JADE environment.

IV. PROPOSED METHOD APPLICATION

Purchase of the crystals before the application of methodologies is presented in Fig. 4.

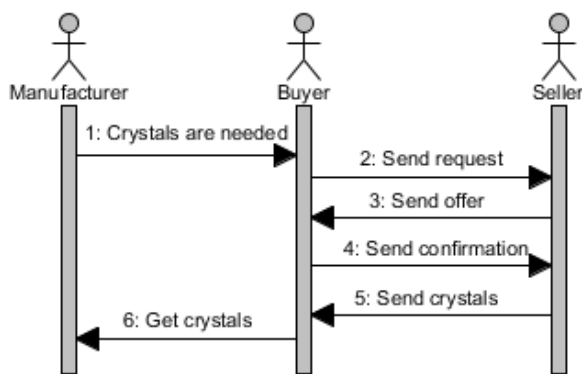


Fig. 4. Communication between buyer and seller managers in case of crystal requirement in raw material warehouse.

Production department needs crystals for chip assembly. Therefore, it is necessary to purchase crystals. Buyer manager sends a request to the seller manager by phone, by e-mail and/or fax. Request is received by a seller manager and the offer with price, quantity and the time of delivery is dispatched by e-mail and/or fax. Offer is received by a seller manager and the confirmation is sent to the offer by e-mail and/or fax. Confirmation is received by the seller manager and requested crystals are sent to the buyer manager. Invoice is sent inside the parcel, and also by e-mail and/or fax. The communication between the buyer and seller ends.

The purchase of crystals using the proposed method is presented in the following steps:

1. Seller agent and buyer agent use the same ontology for cooperation. It starts when raw material for chip production is needed. Buyer agent fulfills the user Profile ontology [11], [12] and it becomes an instance of domain ontology [11], [13].
2. Cooperation begins with the buyer agent's call for proposal (CFP) in the JADE environment. Seller Agent receives a call for proposal for the Buyer Agent with a detailed description of his order. Seller Agent offers to buy the batch from the warehouse. The offer (PROPOSE) is sent with the price, quantity and the time of delivery to the Buyer Agent via the JADE environment.

3. The offer (PROPOSE) is received.

Confirmation (ACCEPT-PROPOSAL) is sent to Seller Agent via the JADE environment.

4. Confirmation is received.

Requested crystals are dispatched to the buyer. Invoice is sent in the parcel. Electronic invoice is dispatched to Buyer Agent (INFORM) (see Fig. 5).

Interaction between agents is suspended until a new request from the Buyer Agent.

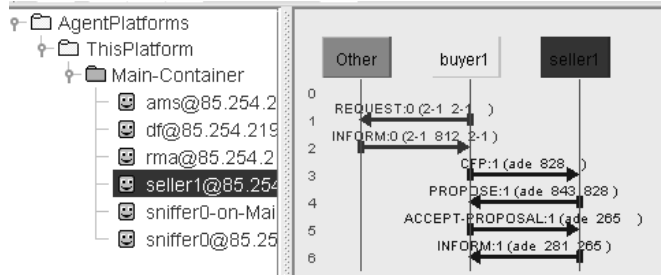


Fig. 5. Agent system with two agents and a common ontology.

V. CONCLUSION AND FUTURE RESEARCH

Ontology is widely used in different problem areas. Multi-agent system uses ontology for agent collaboration. Before ontology can be used, the development of it is needed. This paper presents a detailed description of ontological knowledge model development.

For this study, a multi-agent system was built for the raw materials management task. Implementation of the model in the software environment JADE has demonstrated the interaction of two agents from different parts of the supply chain. To match the name, quantity and price of crystals, agents use ACL messages, in which the domain ontology is a means of structuring knowledge and used for the sharing of knowledge between software agents. The effectiveness of a multi-agent system is proved by precision execution of tasks, quick and timely response to changes in the environment, and the lack of human factors. The study has shown that for the approval of purchase of crystals between the supplier and the customer has been spent just a few seconds using multi-agent system. If a question of time is not relevant, it is possible to the Buyer Agent to bargain and get a better price.

The effectiveness of ontology application in an agent system can be measured by ontology evaluation techniques, which is a future research topic.

REFERENCES

- [1] WiseGEEK, [Online] Available: <http://www.wisegeek.com/what-is-raw-materials-management.htm>
- [2] N. F. Noy, D. L. McGuinness, "Ontology Development 101: A Guide to Creating Your First Ontology," Stanford Knowledge Systems Laboratory Technical Rep. KSL-01-05 and Stanford Medical Informatics Technical Rep. SMI-2001-0880, 2001.
- [3] Wand [Online] Available: <http://www.wandinc.com/taxonomy101.aspx>
- [4] T. Gruber, "A translation approach to portable ontology specifications," *Knowledge Acquisition*, vol. 5, no. 2, pp. 199–220, 1993. <http://dx.doi.org/10.1006/knac.1993.1008>
- [5] I. Serra, R. Girardi, "A Process for Extracting Non-Taxonomic Relationships of Ontologies from Text," *Intelligent Information Management*, vol. 3 no. 4, 2011, pp. 119–124. <http://dx.doi.org/10.4236/iim.2011.34014>

- [6] J. Bermejo, "A Simplified Guide to Create an Ontology," *The Autonomous Systems Laboratory*, Universidad Politécnica de Madrid, 2007.
- [7] E. D'Avanzo, A. Lieto, T. Kuflik "Manually vs semiautomatic domain specific ontology building," Thesis in Information and Commercial Electronics, 2008.
- [8] M. Horridge, "A Practical Guide To Building OWL Ontologies Using Protégé 4 and CO-ODE Tools – Edition1.3," The University of Manchester, 2011.
- [9] C. Lagoze, "Building OWL Ontologies with Protégé," Cornell University, 2008, [Online]. Available: <http://www.cs.cornell.edu/courses/cs431/2008sp/Lectures/public/lecture-4-09-08.pdf>,
- [10] D. Plinere, A. Borisov, "A Negotiation-Based Multi-Agent System for Supply Chain Management," *IT and Management Science*. vol. 49, pp. 128–132, 2011.
- [11] M. Laclavik, Z. Balogh, M. Babik, L. Hluchý, "AgentOWL: Semantic Knowledge Model and Agent Architecture," *Computers and Artificial Intelligence*, 2006.
- [12] A. Sieg, B. Mobasher, and R. Burke, "Web Search Personalization with Ontological User Profiles," *Proc. 16th ACM Conf. Information and Knowledge Management*, CIKM .07, pp. 525–534, 2007.
- [13] M. Golemati, A. Katifori, C. Vassilakis, G. Lepouras, C. Halatsis "Creating an Ontology for the User Profile: Method and Applications,"

1st IEEE International Conference on Research Challenges in Information Science (RCIS), Morocco, 2007

Darya Plinere received the *Mg. sc. ing.* degree from Riga Technical University in 2006. She is a Researcher at the Faculty of Computer Science and Information Technology, Riga Technical University. Her research interests include multi-agent systems, ontology engineering and supply chain management.

E-mail: darja.plinere@rtu.lv.

Arkady Borisov received the Doctoral degree in Technical Cybernetics from Riga Polytechnic Institute in 1970 and *Dr. habil. sc. comp.* degree in Technical Cybernetics from Taganrog State Radio Engineering University in 1986. He is a Professor of Computer Science at the Faculty of Computer Science and Information Technology, Riga Technical University. His research interests include fuzzy sets, fuzzy logic, computational intelligence and bioinformatics. He has 235 publications in the field. He is a member of IFSA European Fuzzy System Working Group, Russian Fuzzy System and Soft Computing Association, honorary member of the Scientific Board, member of the Scientific Advisory Board of the Fuzzy Initiative Nordrhein-Westfalen (Dortmund, Germany).

Address: 1 Kalku Str., Riga, LV-1658; phone: +371 67089530

E-mail: arkadijs.borisovs@cs.rtu.lv.

Darja Plinere, Arkādijs Borisovs. Ontoloģiskā zināšanu modeļa izstrāde izejvielu pārvaldības uzdevumam

Šajā rakstā ir detalizēti aprakstīti ontoloģiskā zināšanu modeļa izstrāde. Ontoloģija plaši tiek izmantota dažādās jomās. Šajā rakstā tiek piedāvāti ontoloģijas izstrādes soļi izejvielu pārvaldības uzdevumam. Izejvielu pārvaldības uzdevums ir piegādes ķēdes kopīga procesa sastāvdaļa, kas koncentrējas uz tādiem uzdevumiem kā augstākas kvalitātes materiāla sagāde par iespējami zemāko cenu, tajā pašā laikā darbojas kompānijas struktūras ietvaros, lai pārliecinātos, ka šie materiāli nodrošina vislabāko labumu ražošanas ietvaros. Šajā darbā tiek pētīti – preču nosaukumi, daudzums un vienošanās starp klientu un piegādātāju par piegādes laiku. Uzdevuma risināšanai tika nolemts pielietot multi-aģentu sistēmu un izmantot ontoloģiju kā saziņas kanālu. Ontoloģija ir saziņas kanāls starp aģentiem, kas ļauj kopīgi izmantot zināšanas un piedāvā vienādu izpratni par viņu sarunu būtību. Lai izmantotu ontoloģiju, tā ir jāizveido. Ontoloģija ir priekšmeta apgabala formālais attēlojums, tā aprakstīšanai ir savi likumi. Dotajam uzdevumam tika izstrādāts ontoloģiskais zināšanu modelis, kā arī rakstā ir parādīta atšķirība starp taksonomiju un ontoloģiju. Aģentu sistēmas tālākai ontoloģijas izmantošanai, ontoloģijas izstrāde tika veikta ontoloģiju redaktorā *Protégé*, pēc tam tā tika pārveidota JADE ontoloģijas klasēs, izmantojot *Ontology Bean Generator*. Aģentu sistēma, konstruēta JADE vidē, parādīja mijiedarbību starp diviem aģentiem un atrada uzdevuma risinājumu. Piedāvātās pieejas efektivitāti nodrošina izpildes precizitāte, savlaicīga reaģēšana uz vides izmaiņām un cilvēka faktora trūkums. Ja laika faktors nav svarīgs, ir iespējams organizēt konkursu vai izsoli.

Дарья Плинере, Аркадий Борисов. Построение онтологической модели знаний для задачи управления материальными ресурсами

В данной статье детализированно описывается построение онтологической модели знаний. Онтологии широко применяются для различных предметных областей. В данной статье представлено пошаговое построение онтологии на примере задачи управления материальными ресурсами предприятия. Задача управления материальными ресурсами предприятия является частью общего процесса управления цепочками поставок и концентрируется на решении проблем обеспечения высокого качества материалов, по возможности, по самым низким ценам. В то же время работа ведется и внутри структуры компании, чтобы убедиться, что эти материалы обеспечивают наилучшую выгоду в рамках производственного процесса. В данной работе исследуется согласование между заказчиком и поставщиком наименования, количества и времени доставки заказа. Для решения поставленной задачи решено применить много-агентную систему и использовать онтологию в качестве канала общения. Онтология является каналом связи между агентами и позволяет совместно использовать знания и иметь одинаковое понимание о сути их переговоров. Для использования онтологии необходимо ее построить. Онтология – это формальное представление предметной области, и существуют правила для ее описания. Для данной задачи построена онтологическая модель знаний, а также в статье показана разница между таксономией и онтологией. Для дальнейшего использования онтологии агентной системой, построение онтологии выполнено в редакторе онтологий *Protégé*, затем онтология была преобразована в JADE классы с помощью *Ontology Bean Generator*. Агентная система, реализованная в среде JADE, продемонстрировала взаимодействие между двумя агентами разных частей цепи поставок и нашла решение поставленной задачи. Эффективность применения предложенного подхода доказывается точностью выполнения, своевременным ответом на изменение среды и отсутствием человеческого фактора. Если временной фактор не важен, становится возможным организовать конкурс или аукцион.

Optimization of the Fuzzy Investment Portfolio during the Time Period

Inna Sydoruk¹, Yuri Zaychenko², ^{1,2} *National Technical University of Ukraine
"Kiev Polytechnic Institute"*

Abstract – The problem of portfolio optimization under uncertainty is considered. For its solution the application of fuzzy sets theory is suggested. Fuzzy portfolio optimization problem during the time period is stated, its model is provided, investigated, and algorithm of its solution presented. This problem includes two main criteria – portfolio profitability and risk. A mathematical model of this problem is constructed and explored. For better estimation of stock profitability, Fuzzy Group Method of Data Handling (FGMDH) is suggested. The experimental investigations of the suggested approach are carried out. The results with optimal portfolios based on forecasted profitability are presented and its efficiency is evaluated.

Keywords – Forecasting, fuzzy portfolio, FGMDH, portfolio optimization, profitability, stock prices.

I. INTRODUCTION

Portfolio is a purposefully formed set of investment assets (real or financial investments) owned by an individual or a legal person. The aim of this set is to implement the pre-developed strategy and to achieve the investment objectives. The main objective of portfolio investment is to create for a set of investment assets such investment conditions that are inaccessible from the position of a single asset and possible only when it is combined with the other. This includes achieving the optimal combination for investors of such investment characteristics as profitability and risk level.

Portfolio analysis exists, perhaps, as long, as people think about making rational decisions connected with the use of limited resources. However, the occurrence moment of portfolio analysis started with a publication of pioneer work by Harry Markovitz (Markovitz's Portfolio Selection) in 1952. The model offered in that work, simple enough in essence, has allowed catching the basic features of the financial market, from the point of view of the investor, and has supplied the last with the tool for development of rational investment decisions.

The central problem in the Markovitz's theory is the portfolio choice, i. e., a set of operations. Thus, in the estimation, both separate operations and their portfolios are considered: profitability and risk of operations and their portfolios. The risk thus receives a quantitative estimation. The account of mutual correlation dependences between profitability of operations appears the essential moment in the theory. This account allows making effective diversification of portfolio, leading to essential decrease in risk of a portfolio in comparison with risk of the operations included in it. At last, the quantitative characteristic of the

basic investment characteristics allows defining and solving a problem of a choice of an optimum portfolio in the form of a problem of quadratic optimization.

However, the worldwide market crises during the last 20 years have shown that existing theories of optimization of share portfolios and forecasting of share indices have exhausted themselves, and essential revision of share management methods is necessary.

Thus, in the light of obvious insufficiency of available scientific methods for management of financial assets, the development of a fundamentally new theory of management of the financial system functioning in the conditions of essential uncertainty is needed. The great assistance to this theory was rendered by the theory of the fuzzy sets, which have been developed about half a century ago in fundamental works of Lotfi Zadeh.

In previous studies, we have discovered the application of classical probabilistic method and fuzzy set theory. We have considered direct, dual and multi-criteria optimization portfolio problems.

The purpose of the present article is the research and experimental analysis of direct optimization problem during the time period.

II. PROBLEM STATEMENT

Let us consider a share portfolio from N components and its expected behavior at time interval $[0, T]$. Each portfolio component is characterized by the financial profitability r_i , $i = 1, N$.

The holder of a share portfolio – the private investor, the investment company, mutual fund – operates the investments, being guided by certain reasons. On the one hand, the investor tries to maximize the profitability. On the other hand, it fixes a maximum permissible risk of an inefficiency of the investments. We will assume the capital of the investor be equal to 1. The problem of optimization of a share portfolio consists in the finding of a vector of share price distribution of papers in a portfolio $x = \{x_i\}$, $i = \overline{1, N}$ of the investor maximizing the income at the set risk level (obviously,

that $\sum_{i=1}^N x_i = 1$).

Weaknesses of a Classical Markovitz's Model

In the process of practical application of Markovitz's model, its drawbacks were found out:

1. The hypothesis about normality profitableness distributions in practice does not prove to be true.
2. Stationarity of price processes also not always is in practice.
3. At last, the risk of actives is considered a dispersion standard deviation of the prices of securities from the expected value, i.e., a decrease in profitableness of securities in relation to the expected value, and profitableness increase in relation to an average are estimated absolutely the same.

Though for the proprietor of securities these events are absolutely different.

These weaknesses of Markovitz's theory define the necessity of the development of essentially new approach of definition of an optimum investment portfolio.

III. FUZZY SET PORTFOLIO MODEL

Main principles and idea of the method are as follows:

- The risk of a portfolio is not its volatility, but possibility that expected profitableness of a portfolio will appear below some pre-established planned value.
- Correlation of assets in a portfolio is not considered and not accounted.
- Profitableness of each asset is not a random fuzzy number. Similarly, restriction on an extremely low level of profitableness can be both usual scalar and fuzzy number of any kind. Therefore, optimization of a portfolio in such a statement may mean in that specific case the requirement to maximize expected profitableness of a portfolio at a point of time T at the fixed risk level of a portfolio.
- Profitableness of a security on termination of ownership term is expected to be equal to r and is in a settlement range.

Denote for the i -th security:

$\tilde{r}_i$ – expected profitableness of i -th security;

r_{i1} – the lower border of profitableness of i -th security;

r_{i2} – the upper border of profitableness of i -th security.

$r_i = (r_{i1}, \tilde{r}_i, r_{i2})$ – profitableness of i -th security, is a triangular fuzzy number.

Then profitableness of a portfolio:

$$r = \left(r_{\min} = \sum_{i=1}^N x_i r_{i1}; \tilde{r} = \sum_{i=1}^N x_i \tilde{r}_i; r_{\max} = \sum_{i=1}^N x_i r_{i2} \right), \quad (1)$$

where x_i is a portion of i -th asset in portfolio, and

$$\sum_{i=1}^N x_i = 1, x_i \geq 0, i = \overline{1, N} \quad (2)$$

Critical level of profitableness of a portfolio at the moment of T is r^* .

IV. MATHEMATICAL MODEL OF A FUZZY OPTIMIZATION PROBLEM

To define the structure of a portfolio, which will provide the maximum profitableness at the set risk level, it is required to solve the following problem (1) – (6):

$$\{x_{opt}\} = \{x\} \mid r \rightarrow \max, \beta = \text{const}, \quad (3)$$

where r is profitableness, β is a desired risk, vector's components x satisfy (2).

As shown in (1) – (6), the most expected value risk degree of a portfolio is defined:

$$\beta = \begin{cases} 0, & \text{if } r^* < r_{\min} \\ R \left(1 + \frac{1-\alpha_1}{\alpha_1} \ln(1-\alpha_1) \right), & \text{if } r_{\min} \leq r^* \leq \tilde{r} \\ 1 - (1-R) \left(1 + \frac{1-\alpha_1}{\alpha_1} \ln(1-\alpha_1) \right), & \text{if } \tilde{r} \leq r^* < r_{\max} \\ 1, & \text{if } r^* \geq r_{\max} \end{cases} \quad (4)$$

where

$$R = \begin{cases} \frac{r^* - r_{\min}}{r_{\max} - r_{\min}}, & \text{if } r^* < r_{\max} \\ 1, & \text{if } r^* \geq r_{\max} \end{cases} \quad (5)$$

$$\alpha_1 = \begin{cases} 0, & \text{if } r^* < r_{\min} \\ \frac{r^* - r_{\min}}{\tilde{r} - r_{\min}}, & \text{if } r_{\min} \leq r^* < \tilde{r} \\ 1, & \text{if } r^* = \tilde{r} \\ \frac{r_{\max} - r^*}{r_{\max} - \tilde{r}}, & \text{if } \tilde{r} < r^* < r_{\max} \\ 0, & \text{if } r^* \geq r_{\max} \end{cases} \quad (6)$$

The profitableness of a portfolio is shown in (1).

V. MATHEMATICAL MODEL OF PORTFOLIO OPTIMIZATION DURING THE PERIOD

In this case we must define the structure of the portfolio, which provides the maximum average return for a given level of risk. Thus, we calculate the profitableness from (3) as:

$$\tilde{r} = \frac{1}{T} \sum_{t=1}^T \sum_{i=1}^N x_{ti} \tilde{r}_{ti},$$

where $\tilde{r}_{ti}$ – the expected profitableness of i -th security in time unit t , T – the number of time units. We should find an optimal portfolio from the following problem:

$$\tilde{r} = \sum_{i=1}^N x_i \tilde{r}_i \rightarrow \max, \quad (7)$$

$$\beta = \text{const}, \quad (8)$$

$$\sum_{i=1}^N x_i = 1, x_i \geq 0, i = \overline{1, N} \quad (9)$$

At a risk level variation β 3 cases are possible. We will consider in detail each of them.

1. $\beta = 0$

From (4) it is evident that this case is possible

$$\text{when } r^* < \frac{1}{T} \sum_{t=1}^T \sum_{i=1}^N x_{ti} r_{ti1}$$

We receive the following problem of linear programming (9) – (11):

$$\tilde{r} = \frac{1}{T} \sum_{t=1}^T \sum_{i=1}^N x_{ti} \tilde{r}_{ti} \rightarrow \max, \quad (10)$$

$$r^* < \frac{1}{T} \sum_{t=1}^T \sum_{i=1}^N x_{ti} r_{ti1}, \quad (11)$$

$$\sum_{i=1}^N x_i = 1, x_i \geq 0, i = \overline{1, N}, t = \overline{1, T} \quad (12)$$

The found result of the problem decision (10) – (12) vector $x = \{x_i\}$ $i = \overline{1, N}$ is a required structure of an optimum portfolio for the given risk level.

2. $\beta = 1$

From (4) it follows that this case is possible when

$$r^* \geq \frac{1}{T} \sum_{t=1}^T \sum_{i=1}^N x_{ti} r_{ti2}.$$

We receive the following problem of linear programming (13) – (15):

$$\tilde{r} = \frac{1}{T} \sum_{t=1}^T \sum_{i=1}^N x_{ti} \tilde{r}_{ti} \rightarrow \max, \quad (13)$$

$$r^* \geq \frac{1}{T} \sum_{t=1}^T \sum_{i=1}^N x_{ti} r_{ti2}, \quad (14)$$

$$\sum_{i=1}^N x_i = 1, x_i \geq 0, i = \overline{1, N}, t = \overline{1, T} \quad (15)$$

3. $0 < \beta < 1$

From (4) it is evident that this case is possible

$$\text{when } \frac{1}{T} \sum_{t=1}^T \sum_{i=1}^N x_{ti} r_{ti1} \leq r^* < \frac{1}{T} \sum_{t=1}^T \sum_{i=1}^N x_{ti} \tilde{r}_{ti}, \quad \text{or} \quad \text{when}$$

$$\frac{1}{T} \sum_{t=1}^T \sum_{i=1}^N x_{ti} \tilde{r}_{ti} \leq r^* < \frac{1}{T} \sum_{t=1}^T \sum_{i=1}^N x_{ti} r_{ti2}.$$

$$\text{a) Let } \frac{1}{T} \sum_{t=1}^T \sum_{i=1}^N x_{ti} r_{ti1} \leq r^* < \frac{1}{T} \sum_{t=1}^T \sum_{i=1}^N x_{ti} \tilde{r}_{ti}.$$

Then using (4) – (7) problem (7) – (9) is reduced to the following problem of nonlinear programming:

$$\tilde{r} = \frac{1}{T} \sum_{t=1}^T \sum_{i=1}^N x_{ti} \tilde{r}_{ti} \rightarrow \max, \quad (16)$$

$$\beta = \frac{1}{\frac{1}{T} \sum_{t=1}^T \sum_{i=1}^N x_{ti} r_{ti2} - \frac{1}{T} \sum_{t=1}^T \sum_{i=1}^N x_{ti} r_{ti1}} \left(r^* - \frac{1}{T} \sum_{t=1}^T \sum_{i=1}^N x_{ti} r_{ti1} \right) + \frac{1}{\frac{1}{T} \sum_{t=1}^T \sum_{i=1}^N x_{ti} r_{ti2} - \frac{1}{T} \sum_{t=1}^T \sum_{i=1}^N x_{ti} r_{ti1}}.$$

$$\cdot \left(\frac{1}{T} \sum_{t=1}^T \sum_{i=1}^N x_{ti} \tilde{r}_{ti} - r^* \right) \cdot \ln \left(\frac{\frac{1}{T} \sum_{t=1}^T \sum_{i=1}^N x_{ti} \tilde{r}_{ti} - r^*}{\frac{1}{T} \sum_{t=1}^T \sum_{i=1}^N x_{ti} \tilde{r}_{ti} - \frac{1}{T} \sum_{t=1}^T \sum_{i=1}^N x_{ti} r_{ti1}} \right), \quad (17)$$

$$\frac{1}{T} \sum_{t=1}^T \sum_{i=1}^N x_{ti} r_{ti1} \leq r^*, \quad (18)$$

$$\frac{1}{T} \sum_{t=1}^T \sum_{i=1}^N x_{ti} \tilde{r}_{ti} > r^*, \quad (19)$$

$$\sum_{i=1}^N x_i = 1, x_i \geq 0, i = \overline{1, N}, t = \overline{1, T} \quad (20)$$

$$\frac{1}{T} \sum_{i=1}^T \sum_{i=1}^N x_{ti} r_{ti2} > r^*, \quad (23)$$

$$\text{b) Let } \frac{1}{T} \sum_{i=1}^T \sum_{i=1}^N x_{ti} \tilde{r}_{ti} \leq r^* < \frac{1}{T} \sum_{i=1}^T \sum_{i=1}^N x_{ti} r_{ti2}.$$

$$\frac{1}{T} \sum_{i=1}^T \sum_{i=1}^N x_{ti} \tilde{r}_{ti} \leq r^*, \quad (24)$$

Then the problem (7) – (9) is reduced to the following problem of nonlinear programming:

$$\sum_{i=1}^N x_i = 1, x_i \geq 0, i = \overline{1, N}, t = \overline{1, T} \quad (25)$$

$$\tilde{r} = \frac{1}{T} \sum_{i=1}^T \sum_{i=1}^N x_{ti} \tilde{r}_{ti} \rightarrow \max, \quad (21)$$

$$\beta = \frac{1}{\frac{1}{T} \sum_{i=1}^T \sum_{i=1}^N x_{ti} r_{ti2} - \frac{1}{T} \sum_{i=1}^T \sum_{i=1}^N x_{ti} r_{ti1}} \left(r^* - \frac{1}{T} \sum_{i=1}^T \sum_{i=1}^N x_{ti} r_{ti1} \right) - \frac{1}{\frac{1}{T} \sum_{i=1}^T \sum_{i=1}^N x_{ti} r_{ti2} - \frac{1}{T} \sum_{i=1}^T \sum_{i=1}^N x_{ti} r_{ti1}} \cdot \left(r^* - \frac{1}{T} \sum_{i=1}^T \sum_{i=1}^N x_{ti} \tilde{r}_{ti} \right) \cdot \ln \left(\frac{r^* - \frac{1}{T} \sum_{i=1}^T \sum_{i=1}^N x_{ti} \tilde{r}_{ti}}{\frac{1}{T} \sum_{i=1}^T \sum_{i=1}^N x_{ti} r_{ti2} - \frac{1}{T} \sum_{i=1}^T \sum_{i=1}^N x_{ti} r_{ti1}} \right), \quad (22)$$

The R-algorithm of minimization of not differentiated functions is applied to the decision of problems (16) – (20) and (21) – (25). Vector $x = \{x_i\} \ i = \overline{1, N}$ is the required portfolio structure.

VI. EXPERIMENTAL INVESTIGATIONS AND RESULT ANALYSIS

As the input data closing prices of leading companies at the stock exchange NYSE, the authors used: Canon Inc. (ADR), McDonald's Corporation (MCD), PepsiCo, Inc (PEP), the Procter & Gamble Company (PG), SAP AG (ADR) (SAP) in the period from 13 December 2013 to 17 January 2014. The founded profitableness is presented in Table I.

TABLE I
PROFITABLENESS, %

Companies	CAJ	MCD	PEP	PG	SAP
Dates					
13 December 2012	1.650	1.337	2.128	2.843	1.566
20 December 2012	-0.060	-1.111	-1.000	-0.184	-2.247
27 December 2012	1.560	-0.633	-1.038	-0.861	-1.132
3 December 2014	0.880	0.484	0.808	1.890	2.655
10 January 2014	1.770	0.052	-1.483	0.422	1.042
17 January 2014	-1.270	-0.105	0.206	0.162	0.843

Further using the Fuzzy GMDH method [10] with triangular membership functions, linear partial descriptions and training

sample of 70%, the next profitableness values were forecasted for each week (Table II):

TABLE II
FORECASTED VALUES OF SHARES PROFITABLENESS, %

Companies	Profitableness				MAPE test sample	MSE Test sample
	Low bound	Forecasted	Upper bound	Real		
13 December 2012						
CAJ	1.498	1.636	1.774	1.65	2.5877	0.0163
MCD	1.112	1.351	1.59	1.337	2.8177	0.0325
PEP	2.015	2.156	2.297	2.128	2.7129	0.0233
PG	2.61	2.839	3.068	2.843	2.3729	0.0088

SAP	1.387	1.599	1.811	1.566	1.7877	0.0364
20 December 2013						
CAJ	-0.284	-0.036	0.212	-0.06	1.4456	0.0235
MCD	-1.303	-1.144	-0.985	-1.111	2.1045	0.0302
PEP	-1.265	-0.994	-0.723	-1	1.0498	0.0131
PG	-0.35	-0.171	0.008	-0.184	2.4122	0.0162
SAP	-2.309	-2.207	-2.105	-2.247	2.5935	0.0348
27 December 2013						
CAJ	1.356	1.594	1.832	1.56	2.1562	0.0212
MCD	-0.846	-0.647	-0.448	-0.633	1.8341	0.0287
PEP	-1.167	-1.006	-0.845	-1.038	2.6743	0.0189
PG	-1.134	-0.875	-0.616	-0.861	2.0452	0.0209
SAP	-1.305	-1.103	-0.901	-1.132	1.8243	0.0152
3 January 2014						
CAJ	0.696	0.914	1.132	0.88	2.4284	0.0283
MCD	0.218	0.497	0.776	0.484	1.9135	0.0134
PEP	0.604	0.796	0.988	0.808	1.899	0.0092
PG	1.662	1.863	2.064	1.89	2.2715	0.0088
SAP	2.539	2.683	2.827	2.655	1.8143	0.0256
10 January 2014						
CAJ	1.498	1.746	1.994	1.77	1.5463	0.0074
MCD	-0.16	0.027	0.214	0.052	1.5162	0.0147
PEP	-1.657	-1.512	-1.367	-1.483	2.1347	0.0382
PG	0.214	0.447	0.68	0.422	2.1756	0.0301
SAP	0.846	1.024	1.202	1.042	2.0015	0.0215
17 January 2014						
CAJ	-1.484	-1.246	-1.008	-1.27	2.5877	0.0163
MCD	-0.347	-0.118	0.111	-0.105	2.8177	0.0325
PEP	0.001	0.242	0.483	0.206	2.7129	0.0233
PG	0.041	0.17	0.299	0.162	2.3729	0.0088
SAP	0.675	0.867	1.059	0.843	1.7877	0.0364

By using the Fuzzy GMDH method, the portfolio optimization system stops to be dependent on the factor of expert subjectivity. Let us see the results of application of the suggested approach to determining an optimal invest portfolio to the date 17 January 2014. Let the critical profitableness level set by trader 0.7 %. Varying the risk level we obtain the following results for triangular MF presented in Tables III, IV

and Fig. 1. As we can see in Fig. 1, the dependence of profitableness–risk has descending type; the greater risk, the lesser profitableness is opposite to classical probabilistic methods. It may be explained by the fact that in the fuzzy approach by risk is meant the situation when the expected profitableness happens to be less than the given criteria level. When the expected profitableness decreases, the risk grows.

TABLE III
COMPONENTS OF OPTIMAL PORTFOLIO FOR TRIANGULAR MF WITH CRITICAL LEVEL $r^* = 0.7\%$

CAJ	MCD	PEP	PG	SAP
0.05482	0.00196	0.0027	0.00234	0.93818
0.06145	0.00113	0.00606	0.0039	0.92746
0.0698	0.00577	0.00235	0.00219	0.91989
0.06871	0.00228	0.0057	0.00244	0.92087
0.07567	0.00569	0.00106	0.00094	0.91664
0.07553	0.00002	0.0029	0.00208	0.91947
0.06774	0.00121	0.006	0.00234	0.92271
0.0764	0.001	0.00612	0.00464	0.91184
0.09072	0.00849	0.00655	0.0039	0.89034

TABLE IV
PARAMETERS OF OPTIMAL PORTFOLIO WITH CRITICAL LEVEL $R^*=0.7\%$

Low bound	Expected profitableness	Upper bound	Risk level
0.55133	0.74591	0.94049	0.2
0.53462	0.72954	0.92446	0.25
0.51544	0.71084	0.90624	0.3
0.51894	0.71431	0.90968	0.35
0.5045	0.70018	0.89587	0.4
0.50877	0.70425	0.89973	0.45
0.522	0.71731	0.91262	0.5
0.50197	0.69752	0.89308	0.55
0.46358	0.66014	0.8567	0.6

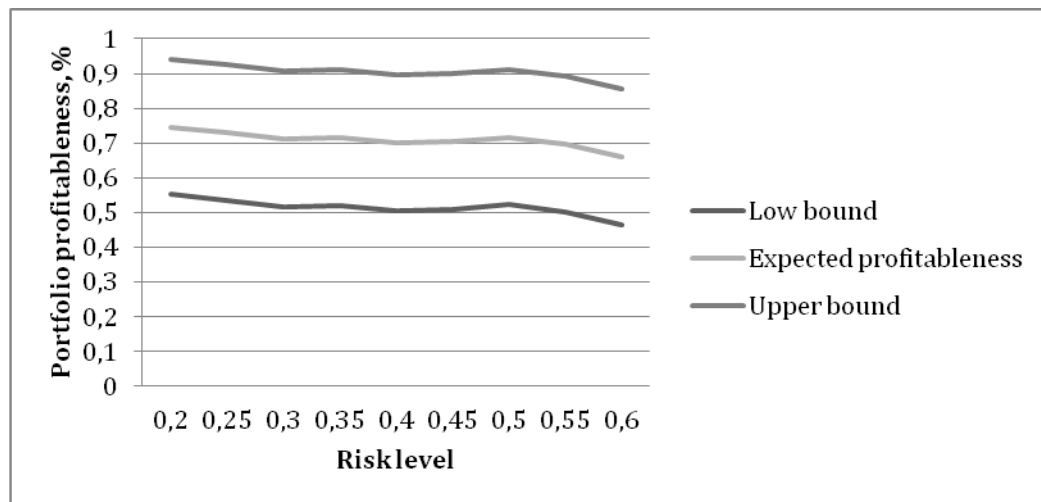


Fig. 1. Dependence of expected portfolio profitableness on risk level for triangular MF.

The calculated corridor of profitableness for optimal portfolio is [0.55133; 0.74591; 0.94049]. The main portfolio portion in this case goes to company SAP that can be explained by the

high level of its profitableness in comparison with other companies.

Now consider an optimal portfolio during 6 weeks (see Tables V, VI and Fig. 2).

TABLE V
OPTIMAL PORTFOLIO COMPONENTS DURING TIME PERIOD AND CRITERIA LEVEL 0.7%

CAJ	MCD	PEP	PG	SAP
0.00558	0.00123	0.00072	0.94771	0.06476
0.00871	0.00245	0.00051	0.92732	0.06101
0.00629	0.00595	0.00137	0.92718	0.05921
0.00651	0.00332	0.00058	0.91748	0.07211
0.00634	0.00893	0.00075	0.91475	0.06923
0.00877	0.00094	0.00032	0.92444	0.06553
0.00641	0.00056	0.00162	0.91957	0.07184
0.00746	0.00239	0.00054	0.90852	0.08109
0.00901	0.00654	0.00021	0.88581	0.09843

TABLE VI
OPTIMAL PORTFOLIO COMPONENTS DURING TIME PERIOD AND CRITERIA LEVEL 0.7%

Low bound	Expected profitableness	Upper bound	Risk level
0.49289	0.69582	0.89876	0.2
0.47619	0.68156	0.88693	0.25
0.47517	0.68102	0.88687	0.3
0.45507	0.66089	0.86671	0.35
0.46821	0.67435	0.88049	0.4
0.45918	0.66511	0.87104	0.45
0.45098	0.65674	0.8625	0.5
0.43601	0.64202	0.84803	0.55
0.42386	0.63087	0.83788	0.6

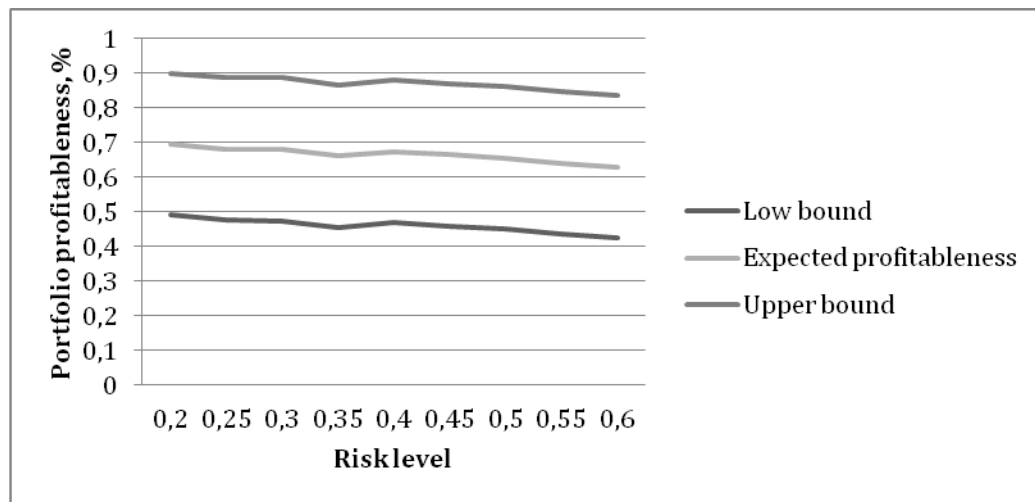


Fig. 2. The dependence of expected portfolio profitableness on risk level.

The calculated corridor of profitableness for optimal portfolio is [0.49289; 0.69582; 0.94049].

The main portfolio portion in this case goes to company PG that can be explained by the high level of its average profitableness in comparison with other companies.

VII. CONCLUSION

In this paper, the research in the field of portfolio management has been carried out. Fuzzy set theory has been used as a tool for getting an optimal portfolio. As a result of research, the mathematical model based on the fuzzy-set approach for finding of structure of the optimal investment portfolio has been received. On the basis of the theory of fuzzy sets, the algorithm of optimization of a share portfolio has been developed. As a result of research, the following conclusions have been made:

1. The dependence of profitableness–risk has a descending type; the greater risk, the lesser profitableness is opposite to classical probabilistic methods.
2. Portfolios during the time period and at the end of period have different structure and characteristics that can be explained by the different calculations of profitableness.

3. For improving the accuracy of the suggested fuzzy portfolio model, the fuzzy GMDH method has been applied to profitableness forecasting. The experimental investigations have proved its high efficiency.

REFERENCES

- [1] A.O. Nedosekin, "Monotonnyye portfel'i i ikh optimizatsiya", in *Audit i finansovyy analiz*, N2, 2002, pp. 68–72.
- [2] Sistema optimizatsii fondovogo portfelya (Simens Biznes Servisez Rossiya). [Online]. Available: <http://www.sbs.ru/index.asp?objectID=1863&lang=rus> [Accessed: Sept. 5, 2014].
- [3] A. O. Nedosekin, "Sistema optimizatsii fondovogo portfelya ot Simens Biznes Servisez", in *Bankovskiy tekhologii*, N5, 2003.
- [4] A. O. Nedosekin, "Optimizatsiya biznes-portfelya korporatsii". [Online]. Available: http://sedok.narod.ru/s_files/2003/Art_070303.doc.
- [5] P. Yu. Zaychenko, M. Yesfandiyarfard, "Analiz investitsionnogo portfelya dlya razlichnykh vidov funktsiy prinaldezhnosti", in *Sistemni doslidzhennya ta informatsiyni tekhologii*, N2, 2008, pp. 59–76.
- [6] P. Yu. Zaychenko, M. Yesfandiyarfard, "Nechetkiy metod induktivnogo modelirovaniya dlya prognozirovaniya kursov aktsiy v zadachakh portfel'noy optimizatsii", *Visnik Cherkas'kogo derzhavnogo tekhnologichnogo universitetu*, N1, 2008, pp. 9–14.
- [7] P. Yu. Zaychenko, M. Yesfandiyarfard, O. N. A. Ag. Gamish, "Analiz modeli optimizatsii nechetkogo portfelya", *Visnik natsional'nogo tekhnichnogo universitetu Ukraini "KPI"*. Informatika, upravlinnya ta obchisluyal'na tekhnika. Kiyv TOO "VEK+", N51, 2010, pp. 197–203.

- [8] P. Yu. Zaychenko, M. Yesfandiyarfard. "Analiz i sravneniye rezul'tatov resheniya zadachi optimizatsii investitsionnogo portfelya pri primeneni modeli Markovitsa i nechetko-mnozhestvennogo metoda", in *Proceedings of XIII-th International Conference KDS-2007 "Knowledge, Dialogue, Solution"*. June, 2007, Bulgaria, pp. 278–286.
- [9] P. Yu. Zaychenko, M. Yesfandiyarfard, "Optimizatsiya investitsionnogo portfelya v usloviyakh neopredelennosti na osnove prognozirovaniya kursov aktsiy", in *Proceedings of XIV-th International Conference KDS-2008 Knowledge, Dialogue, Solution*. June, 2008, Bulgaria, pp. 212–228.
- [10] P. Yu. Zaychenko, *Nechetkiye modeli i metody v intellektual'nykh sistemakh*. Uchebnoye posobiye dlya studentov vysshikh ucheb. Zavedeniy, Kiev, Slovo, 2008.
- [11] M. Z. Zgurovskiy, P. Yu. Zaychenko. *Modeli i metody prinyatiya resheniy v nechetkikh usloviyakh*. Kiev, Naukova dumka, 2011.
- [12] P. L. Vilenskiy, V. N. Livshits, Ye. R. Orlova, S. A. Smolyak, *Otsenka effektivnosti investitsionnykh proyektov*. Moskva, Delo, 1998.
- [13] S. A. Smolyak, "Uchet spetsifiki investitsionnykh proyektov pri otsenke ikh effektivnosti", *Audit i finansovyy analiz*, N3, 1999.
- [14] Yu. P. Zaychenko, Malikhekh Yesfandiyarfard, "Analiz i sravneniye rezul'tatov optimizatsii investitsionnogo portfelya pri primeneni modeli Markovitsa i nechetko-mnozhestvennogo metoda" in *Proceedings of X111-th International Conference KDS-2007 "Knowledge, Dialogue, Solution"*, vol. 1, pp. 278–287.



Inna A. Sydoruk graduated from the Institute for Applied System Analysis, NTUU "Kiev Polytechnic Institute". Her major fields of research are: information systems, fuzzy set theory, investment portfolio optimization, decision-making theory.
E-mail: isidoruk@ukr.net.



Yuriy P. Zaychenko is an Associate Professor with Kiev Polytechnic Institute. Since 1981, he has been a Professor the Institute for Applied System Analysis of Kiev Polytechnic Institute. His main fields of research are: decision- making under uncertainty and risk, fuzzy logic systems and fuzzy neural network and their applications in economy and finance, computational intelligence. Since 1995, he has been an Academician of the Ukrainian Academy of High School. He received a state prize award in science and technology in 2012.
E-mail: zaychenkoyuri@ukr.net, baskervil@voliacable.com.

Inna Sidoruka, Jurijs Zaičenko. Investīciju portfēļa optimizācija laika periodā

Investīciju portfēļa izmantošanas galvenais uzdevums ir uzlabot investēšanas nosacījumus, piešķirot vērtspapīru (VP) kopumam tādas investīciju raksturlielumus, kuri nav sasniedzami, izmantojot tikai atsevišķu VP, bet ir iespējami tikai gadījumā, kad tos izmanto kombinācijā. Mūsdienu pasaules ekonomisko procesu nestabilitāte rada neskaidrību investīciju lēmumu pieņemšanā. Sakarā ar nepieciešamību cīnīties pret to, šajā darbā tiek izmantota izplūdušo kopu pieeja, kur portfēļa risks ir nevis tā nestabilitāte, bet gan iespējamība, ka sagaidāmais portfēļa ienesīgums izrādīsies zemāks par kādu iepriekš definētu plānoto vērtību. Portfēļa aktīvu korelācija netiek apskatīta un ņemta vērā. Katra aktīva ienesīgums nav gadījuma veida izplūdušā vērtība (piemēram, trīsstūra vai intervāla veidā). Šajā darbā apskatīts investīciju portfēļa optimizācijas uzdevums konkrētā periodā. Ir jānosaka portfēļa struktūra, kura sniegs maksimālo vidējā ienesīguma līmeni periodā, ņemot vērā noteikto riska līmeni. Izejas dati tika iegūti no NYSE biržas, ņemot tādu ietekmīgu uzņēmumu kā Canon Inc. (ADR), McDonald's Corporation (MCD), PepsiCo, Inc. (PEP), Procter & Gamble Company (PG), SAP AG (SAP) akciju cenas izsolē pirms slēgšanas, laika posmā no 13.12.13. līdz 17.01.14. Ienesīguma prognozēšanai tika izmantota argumentu grupveida uzskaites izplūdušā metode, kas ļāva izvairīties no ekspertu subjektivitātes sistēmā. Rezultātā tika izveidots optimālais portfēlis sestās nedēļas beigām (17.01.14.) un visam laika posmam. Veikta iegūto rezultātu salīdzinošā analīze. Šādā veidā darbā tiek turpināts pētījums par investēšanas vadību, izmantojot kvalitatīvi jaunu izplūdušo kopu pieeju. Balstoties uz izveidoto matemātisko modeli un, izmantojot tā realizāciju programmatūras veidā, tika veikta noteiktajam laika periodam optimālā portfēļa meklēšana un iegūto rezultātu analīze.

Инна Сидорук, Юрий Зайченко. Оптимизация инвестиционного портфеля в течение периода времени

Основная задача портфельного инвестирования – улучшить условия инвестирования, предоставляя совокупности ценных бумаг (ЦБ) такие инвестиционные характеристики, которые недостижимы с позиции отдельно взятой ЦБ и возможны только при их комбинации. Нестабильность экономических процессов современного мира порождает неопределенность в процессе принятия инвестиционных решений. В связи с потребностью в борьбе с неопределенностью в работе применяется нечетко-множественный подход, где риск портфеля – это не его волатильность, а возможность того, что ожидаемая доходность портфеля окажется ниже некоторой установленной плановой величины; корреляция активов в портфеле не рассматривается и не учитывается; доходность каждого актива – это неслучайное нечеткое число (например, треугольного или интервального вида). В текущей работе рассмотрена прямая задача оптимизации инвестиционного портфеля для заданного периода. Нужно определить структуру портфеля, который обеспечит максимальный уровень средней доходности за период при заданном уровне риска. В качестве исходных данных были использованы цены закрытия на бирже NYSE таких влиятельных компаний, как Canon Inc. (ADR), McDonald's Corporation (MCD), PepsiCo, Inc (PEP), the Procter & Gamble Company (PG), SAP AG (ADR) (SAP) в период с 13.12.13 по 17.01.14. Для прогнозирования доходности был использован нечеткий метод группового учета аргументов (НМГУА), что позволило лишить систему субъективности экспертов. В результате был построен оптимальный портфель на конец шестой недели (17.01.14) и для всего периода времени. Проведен сравнительный анализ полученных результатов. Таким образом, в работе продолжено исследование в области управления инвестиционной деятельностью с использованием качественно нового нечетко-множественного подхода. На основе построенной математической модели был программно реализован поиск оптимального портфеля для заданного периода и проанализированы полученные результаты.

STORN: Solution to Traversal of Road Networks

Janis Kampars¹, Elina Shmite^{2, 1, 2} *Riga Technical University*

Abstract – The objective of the optimal traversal planning (OTP) is to calculate a route that would provide traversal of all streets in a predefined area. While solving the OTP problem, the total distance of the route should be minimized. This paper presents a solution that provides OTP and the execution of the corresponding plan using a series of loosely coupled web services. A mashup is created based on the provided web services and field tested using a data acquisition case in Riga, Latvia. Other possible application areas are street cleaning, package delivery and evacuation planning.

Keywords – Chinese postman problem, cloud computing, travelling salesman problem, web services.

I. INTRODUCTION

The purpose of this paper is the development of the Solution to Optimal Traversal of Road Networks (STORN). It will be implemented as a distributed system and comprise a series of web services that provide near optimal traversal of road networks using currently a popular Software as a Service (SaaS) delivery model. A mashup utilizing the mentioned web services will be created and tested by a company performing data acquisition in Riga, Latvia. The data acquisition process involves driving on all streets of Riga (bidirectional streets have to be traversed in both directions) and taking photos for later on-site processing. It is important to minimize the time needed for route generation and the total distance, while making sure that all streets have been covered.

The optimal traversal planning problem has other practical applications, some of which are snow plowing [1], police patrol planning [2], parcel delivery [3] and evacuation planning [4]. The STORN will be later adjusted to suit more use cases, promoting emergence of a wide range of optimal traversal planning mashups and development of thin clients for mobile computing platforms like iOS and Android.

The evolution of the technical paradigm called cloud computing (CC) has significantly affected the way information technology services are invented, developed, deployed and scaled [5], [6], and it strongly correlates with the chosen approach for implementing the STORN. Yang et al. have acknowledged the importance of CC in dealing with technical challenges in geospatial sciences and coined the term Spatial Cloud Computing (SCC) [7]. The STORN will follow the best practices of cloud application development and provide its functionality as a service. Study by Gartner [8] predicted that SaaS would become increasingly important and worldwide software revenues for SaaS delivery were said to grow by 19.4 % overall between 2008 and 2013. Web services provided by the STORN can be used to power mobile applications, following another technological trend and potential future of mobile application development – Mobile Cloud Computing (MCC) [9]. Although smartphones have

become the devices of the future for computing and service access, they are still constrained by CPU performance, memory and battery capacity. The MCC is used for extending the power of CC to mobile devices by offloading computing or data intense application from mobile devices to scalable and virtually unlimited cloud resources [10].

The development of the STORN involves several challenges – acquisition of the necessary spatial data, identification of suitable algorithms for optimal traversal planning and plan execution, data model transformations to fit the algorithmic requirements, definition of a scalable architecture that can be used to provide optimal traversal of road networks as a service.

The organization of this paper is as follows. Section II reviews the algorithms in the area of OTP that could be used in development of the STORN. Current research in the area of cloud application design is also discussed in this section. Section III presents the algorithmic and architectural solutions for the STORN. Section IV experimentally compares two algorithms for OTP. Section V concludes final remarks and possible directions for the future research.

II. STATE OF THE ART

This section reviews related studies in order to develop the needed algorithms for the STORN and choose an appropriate architecture for deploying it in the cloud.

A. Existing Algorithmic Approaches

If the road network is converted into a graph, where street fragments correspond to arcs (directed links) and street intersections are transformed into nodes, we can formalize the development of the OTP as a type of Arc Routing Problem (ARP). ARP is a subtype of vehicle routing problems, in which the tasks to be performed are located on arcs. This problem has been studied less than the node routing problem; however, during the last decade there has been impressive development summarized by [11]. The graph that corresponds to a road network is more similar to one examined in the Mixed Chinese Postman Problem (MCCPP) in terms that it contains arcs, edges and nodes; however, our use case requires that bidirectional streets (edges) have to be driven in both directions, while MCCPP only requires edges to be traversed in one direction. If the road network is transformed into a graph, where bidirectional streets correspond to two arcs (one for each direction), then it can be solved as a Directed Chinese Postman Problem (DCPP). In DCPP each arc has a cost (e. g., distance of the street fragment) and to solve a DCPP one has to find a route (Chinese Postman Tour or CPT) containing all of the arcs. In order to find the CPT, graph must be strongly connected (any node can be reached from any other node). If the graph is an Eulerian, there is an Euler circuit traversing

each arc exactly once, which is also the optimal CPT. A theorem states that a graph is Eulerian only if all of its nodes are balanced (each node has to have an equal number of inbound and outbound arcs). Finding an Euler circle is a relatively easy task [12]; unfortunately, graphs derived from street networks are rarely Euler and thus getting the CPT is more complex.

There are approaches that search for the CPT in a graph that does not require nodes to be balanced [13], while others solve the Chinese Postman Problem (CPP) by converting the initial graph to an Eulerian one (e.g., performing an Eulerian extension) by adding virtual arcs to the graph so that all nodes are balanced [12], [14]. While performing the Eulerian extension, the total cost of added virtual arcs must be minimized. There are also studies that concentrate solely on Eulerian extension of weighted, directed graphs and it is known to be an NP-hard problem [15]. The general version of DCPD does not take into account that some of the turns in the road network may be prohibited by the traffic regulations (e.g., left turn) and some maneuvers are more complex than others (driving straight is easier than making a left turn).

To deal with this issue, there are studies that recommend transforming the CPP into an equivalent Asymmetric Travelling Salesman Problem (ATSP) [14], [16]. Other studies [17] favor a more direct approach and state that transformation to ATSP is inefficient. To solve the ATSP a route containing all the vertices (e.g., the Hamiltonian cycle) must be found. ATSP is an NP-complete problem [14] and there is a wide range of algorithms for finding near optimal solutions, including branch and bound [18], patching [19], [20] and genetic [21] algorithms.

B. Architecting Applications for the Cloud

Designing applications for the cloud is a relatively new research area and it certainly lacks a methodological support. One of the best and technologically detailed sources for cloud application design patterns is whitepapers by cloud service providers. Amazon has published a document [23] containing best practices for designing cloud native applications. It also provides a variety of reference architectures for various tasks (e.g., web applications, batch processing, media serving); however, these are tightly bound to Amazon AWS platform and might be hard to implement on other platforms.

Hamdaqa et al. [24] present a more abstract version of cloud application architecture, in which cloud application is made of `CloudFrontTask` (accessible by an end-user), `CloudRotorTask` (a background process), `CloudCrossCuttingTask` (provides logging services and monitors cloud resources, which is important for scaling), and `CloudPersistenceTask` (provides persistent storage). Tasks are loosely coupled and communication between them is provided using queues.

Christian Inzinger et al. [25] present a methodology that addresses the complete development lifecycle of cloud applications. The methodology is based on iterative refinement of the application architecture based on requirements, from abstract representation of the application

that captures general business requirements to a concrete model of cloud-based application components.

Kwon et al. [26] present a method for transforming traditional applications into cloud-based services. The refactoring process consists of service extraction phase and service interface adaptation. The applicability of the approach is proven by transforming two third party Java applications into cloud-based applications.

Rimal et al. [27] classify architectural requirements for cloud computing systems according to the requirements of cloud providers, the enterprises that use the cloud and end-users. A relation between various cloud deployment models and requirements is also provided.

III. PROPOSED SOLUTION

This section describes STORN services, provides a high level architecture, a data model and algorithms used in implementing the STORN. We use freely available data from OpenStreetMap as our spatial data source.

A. STORN Web Services

A total of six web services (WS) are provided by STORN. User ID and API ID are mandatory parameters at all times; therefore, they are omitted from the web service parameter list.

Routing region selection WS

It is used to obtain a region ID for the required traversal region that is used as an input in other WS. Required input parameter is a name of the obtainable area. Optional parameters include a country code and administrative level of the area to narrow the search.

Routing region definition WS

It is used to define a custom region to make it available for a route definition. Required parameters include area name and area boundaries. The STORN contains user generated regions, which are accessible only by the owner, and public regions, which are based on Open Street Maps (OSM) data [22] and are accessible by all users.

Route generation WS

It is used to generate the actual road network traversal route and to obtain a route ID to be used in other WS. The only required parameter is a region ID. Navigational costs for each turn type in seconds can be provided optionally. If no navigational costs are set, there are no maneuver preferences (driving straight or making a left turn is equally favorable). Other optional parameters include custom vertex and arc tolerance distances for the active arc detection (see Section C).

Route retrieval WS

It is used to obtain spatial information of the route. The only required input parameter is a route ID. Optionally a filter can be specified to retrieve only the visited arcs of the route or unvisited ones. If no filter is set, the complete route is returned. It is also possible to get corresponding street names and hints, which are not returned by default, by using parameters *showStreetName* and *showHints*.

Route traversal WS

It is used to perform the route traversal procedure. It obtains the list of next arcs (the number of arcs can be adjusted by using an optional parameter) to be visited and hints. Required input parameters include a route ID and current coordinates of the driver. Coordinates are used to update the progress of the route traversal. This WS can be used to develop clients for mobile devices. During the research an HTML5 based frontend was implemented based on this service and the Route retrieval WS.

Route progress tracking WS

It is used to monitor the route traversal process by obtaining driver's current location, speed, information about deviation from the generated route and the length of the traversed and remaining distances. The only input parameter required is a route ID.

B. STORN Architecture

STORN has been designed to employ the benefits provided by cloud computing, and its architecture is given in Fig. 1.

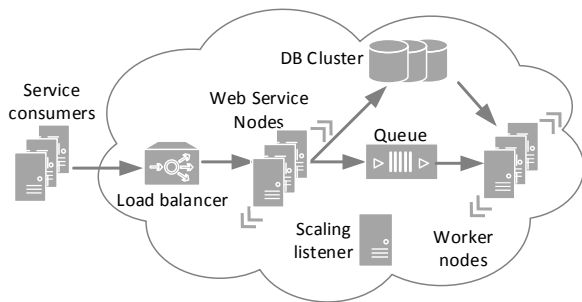


Fig. 1. High level architecture of the STORN.

Web services run on virtual machines that contain the necessary program logic and environment (e. g., Apache web server). The requests from service consumers are distributed to web services using a load balancer. Based on the nomenclature by [24] web service nodes are instances of CloudFrontTask. To provide high availability and performance, the minimum number of web service nodes is at least two at all times. The total number of web service nodes is being managed by a scaling listener (an instance of CloudCrossCuttingTask). When the average response time grows above a globally defined constant, new nodes are added. Similarly, the number of nodes is decreased when the average response time falls below a globally defined constant, but not lower than two nodes.

A Postgres-XL database cluster is used to provide a persistent data store. It is horizontally scalable, fully ACID and compatible with regular PostgreSQL and PostGIS. PostgreSQL/PostGIS is community's preferred database management system in working with OpenStreetMap data, which made compatibility with PostgreSQL/PostGIS an important criterion. Based on the nomenclature by [24] database cluster is an instance of CloudPersistenceTask.

Time and resource consuming tasks are performed by worker nodes (an instance of CloudRotorTask). Messages

describing the tasks are added to a queue by web service nodes. These messages are later pulled from the queue by worker nodes and the required processing is performed. In our case, route generation (initiated by the Route generation WS) is implemented using worker nodes. The scaling listener monitors the length of the queue and adjusts the number of worker nodes accordingly. To provide high availability, there are at least two worker nodes at all times. Worker nodes are implemented as virtual machines, which contain the application logic and environment to execute the needed tasks. During the task execution data are retrieved from the persistent data store and at the end of the process results are written back. Worker nodes are stateless, which means that unrecoverable errors during task execution pose no danger. After a certain visibility period the message containing the instructions for task execution will reappear in the queue and will be processed by another worker node.

C. Data Model

A local copy of Riga, Latvia, highway data is obtained and loaded into a PostgreSQL database. Initial data are loaded in four tables, which serve as a base for further data model transformations. These tables are *planet_osm_lines* (streets, roundabouts, links etc.), *planet_osm_points* (traffic lights, intersection points etc.), *planet_osm_relations* (used as a source for navigational restrictions) and *planet_osm_polygon* (used as a source for public regions). Initial tables are adjusted and transformed according to STORN requirements. Resulting database model is given in Fig. 2.

Table *graph_arcs* corresponds to the Eulerian graph and contains relations between two records from *osm_points*, which form an arc. For each arc $[u;v]$ a distance that equals to w_{uv} is stored. Each bidirectional arc is stored as two separate arcs – one for each direction. This table is used by Fleury's algorithm (See Section D). Tables *graph_e_nodes* and *graph_e_arcs* correspond to Expanded Line Graph and both are based on the *graph_arcs* table and used by Patched Cycle algorithm (See Section D).

Navigational restrictions are retrieved from *planet_osm_relations* and are available in two different forms. One of them defines only the allowed turns from a single or multiple ways (line geometry) via a specific node (point geometry) to a single or multiple ways. Restrictions are defined as string literals that start with a keyword "only" (e. g., "only_straight_turn"). The second approach for storing navigational constraints defines only prohibited turns and starts with a keyword "no" (e. g., "no_left_turn"). For our purposes, we transform every restriction record with many *from* or *to* entries into multiple restriction records with a single *from* and *to* entry.

Table *regions* contains OSM predefined set of administrative territories from *planet_osm_polygon*. Routing region definition WS users are able to add custom traversable regions to this table. If the region is user defined, user ID field contains the ID of the corresponding author (this field is left blank for publicly available regions retrieved from *planet_osm_polygon*).

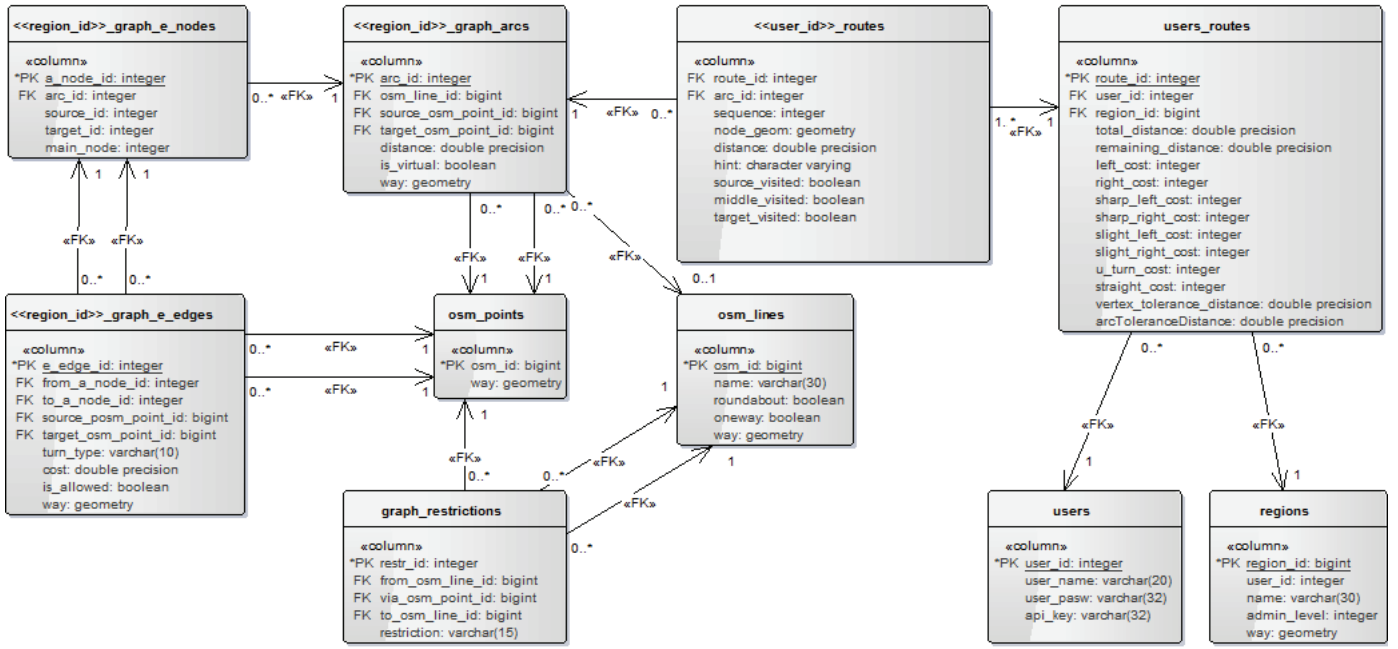


Fig. 2. Database entity relationship diagram of the STORN.

For each route a region ID, navigational costs and active arc detection tolerance parameters are stored in *users_routes*. Because of the different input parameters, tours created for the same regions are not the same and therefore cannot be reused as easily as graph tables.

Users can have more than one route each generated with different parameters. A separate *route* table is generated for each user. It contains the route definition and columns needed for storing route traversal progress.

D. Algorithms

We have based our solution on algorithms and graph models provided in [14]. In addition to navigational constraints, STORN encompasses prohibited turns, driving directions and route tracking.

If $[u;v]$ is the previous traversed arc and $[v;n]$ is the current traversable arc, we define a function $nav(u;v;n)$ to get a type of the turn to be made. Possible values for the turn type are “straight”, “left”, “right”, “sharp right”, “sharp left”, “slight right”, “slight left”, “u-turn” and “prohibited turn”.

When the function checks if the turn is prohibited, the input of source arc and target arc are used to search for affecting restrictions. A restriction is found if *From* way ends with a source arc and the *via* node matches both the source arc’s end point and target arc’s start point. If a restriction is found, its type is further reviewed. If the type is “only_” and *to* way starts with a target arc or the type is “_no” and *to* way does not start with a source arc, turn is allowed, otherwise – turn is prohibited.

We define function $ncost(u;v;n)$ to get a navigational cost for the turn made. Navigational costs used by the algorithms are set via the Route definition WS. Term path cost refers to both navigational costs and arc costs to traverse a given path.

Solving OTP is an optimization problem. Because of the presence of navigational constraints and prohibited turns in our solution we optimize time spent traversing the route. Two alternative algorithms for calculating OTP that work with two different graph models are described in this section – the greedy algorithm (GA) and the approximate algorithm (AA). We use Fleury’s algorithm for the GA with added heuristics to minimize the cost of the Eulerian cycle and we use simplified Patched Cycle algorithm [19] for the AA with added heuristics to minimize the cost of the Hamiltonian cycle.

Graph balancing approach

Both GA and AA require a balanced graph, so first we check if the graph is strongly connected and remove any vertices, which are creating dead-ends. Vertice removal is done by searching for vertices that contain only inbound or outbound arcs. The procedure is repeated while all vertices have at least one inbound and outbound arc.

Further the initial graph G is converted to an Eulerian graph G_e by balancing all of its vertices. Path costs needed by the algorithm are calculated by implementing a shortest path Dijkstra-like algorithm [28]. The classical version of Dijkstra is not applicable [29] in the given graph with prohibited turns as there are possible situations when an already visited node must be included in the graph traversal route to reach the destination node (otherwise no cycle can be found).

The greedy algorithm

Fleury’s algorithm removes a single arc at a time. If removing an arc from the graph makes it disconnected, the arc is called a bridge. Bridge is removed only if there is no other non-bridge arc to traverse. Heuristics are applied to make a minimum cost choice between two or more possible arc selections. The pseudocode of the GA is given below:

```

Algorithm(graph Ge, starting vertex s)
01. OTP = {s}; Bridge_children = {};
   Nonbridge_children = {};
02. while there exists an unvisited arc
03.   for each unvisited arc [s;n]
04.     if [s;n] = bridge arc of Pss
05.       append n to Bridge_children;
06.     else
07.       append n to Nonbridge_children.
08.   if Nonbridge_children is not empty
09.     from all Nonbridge_children vertices x
10.     m = the vertex with the minimum ncost(p;s;m)
11.   else if Bridge_children is not empty
12.     from all Bridge_children vertices y
13.     m = the vertex with the minimum ncost(p;s;m)
14.   else
15.     replace any virtual arc [u;v] with the
       vertices of shortest path Puv
16.     for each arc pair ([u;v], [v;n])
17.       if nav(v;u;n) = 'restricted turn'
18.         replace [v;n] with the vertices of
           shortest path Pvn.
19.   return OTP.
20. mark [s;m] as visited;
21. p = s; //previous parent p
22. s = m; //new parent s
23. bridge_children = {};
24. non_bridge_children = {};

```

In a formed tour we try to replace any detected restricted turns $[v;u;n]$ with the shortest path from v to n .

The approximate algorithm

To implement the AA graph G_e is converted to an Extended Line Graph (ELG) G^*1 . Instead of modeling intersections and the roads connecting them, ELG is used to model roads and turns from one to another [28]. For each arc $[u;v]$ in graph G_e we add an arc $[Uuv;Vuv]$ with a cost W_{uv} from G_e to G^*1 . After this step all the basic arcs are transferred. To add the turns for each vertex v_{uv} from G^*1 all the possible arcs $[Vuv;Vvx]$ with $ncost[u;v;x]$ are added.

AA consists of two phases. At first, the AP is solved to match every vertex with its best neighbor. Finally, the resulting subcycles are patched with the minimum cost to make a single cycle using the Karp-Steel patching [19], [20]. Solving an AP in a large dataset can be a very time-consuming task. To minimize the execution time thus contributing to performance improvements, we create subcycles with Nearest Neighbor Algorithm using the same heuristics as Fleury's Algorithm. Optimization is done only at the second phase solving an AP iteratively patching the two largest subcycles at a time. The pseudocode of the AA is given below:

```

Algorithm(graph G*1, starting vertex s)
01. Subcycles={}; CurrentCycle={};
02. while there exists an unvisited vertex
03.   if CurrentCycle is empty
04.     v = the arbitrary chosen, unvisited vertex;
05.     append v to CurrentCycle;
06.     mark v as visited;
07.   else
08.     from all the unvisited vertices x such that
       there exists an arc [v;x]
09.     c = the vertex with minimum  $W_{vx}$ .

```

```

10.   if c is null
11.     append CurrentCycle to Subcycles;
12.     CurrentCycle = {};
13.   else
14.     append c to CurrentCycle;
15.     v = c;
16.     mark c as visited.
17.     OTP={};
18.   while subcycles count != 1
19.     replace any virtual edge [u;v] with the
       vertices of shortest path Puv
20.     SC1; SC2 = the two largest subcycles;
21.     [u;v] = the arc from SC1 and [n;m] = the arc
       from SC2 with the minimum sum of  $W_{um} + W_{nv} -$ 
        $W_{uv} - W_{nm}$ ;
22.     delete [u;v] from SC1;
23.     delete [n;m] from SC2;
24.     append [u;m], [n,v] to SC1;
25.     join SC1 with SC2.
26.     for each arc [Uvu,Uun]
27.       if nav(v;u;n) = 'restricted turn'
28.         replace [Uvu;Uun] with the vertices of
           shortest path PuvuUun.
29.   order SC1 starting from s;
30.   OTP = SC1;
31.   return OTP;

```

Hint generation approach

Hints are short keywords that show the next maneuver to be performed. They are used in combination with street names of the previous and current arcs, remaining route distance, distance to the next maneuver and other route related information. Hints are generated based on the following eight rules:

1. If there is no previous arc to determine a turn (first arc of the tour), the hint contains direction of the arc (e. g., west).
2. If the next arc is the only possible path from the current arc, no hints are generated.
3. If the next arc is included in a roundabout and current one is not, generate hint "enter roundabout".
4. If the last hint was "enter roundabout" it is determined which exit has to be taken. All allowed exits are counted starting from the roundabout entrance to the desired exit. A hint "exit-n" is generated, where n is the exit number.
5. If the current arc $[u;v]$ is on the same road as the next arc $[v;n]$ and $nav(u;v;n)$ equals "straight", "slight left" or "slight right", no hint is generated. It is natural to continue driving on the same street straight forward or with a slight turn, therefore no guidance is needed.
7. If the current arc $[u;v]$ is not on the same road as the next arc $[n;v]$ and $nav(u;v;n)$ equals "straight", "slight left" or "slight right", generate hint "continue".
8. If for the current arc $[u;v]$ and next arc $[v;n]$ $nav(u;v;n)$ equals "left", "right", "sharp left", "sharp right" or "u-turn", the hint accordingly equals to $nav(u;v;n)$.

Active arc detection approach

During the route traversal, the progress is saved in the routes table. To detect which arcs have been traversed, an active arc detection algorithm is defined. Active (current) arc is the arc which according to the previously generated route is being traversed at the moment and has not been marked as

finished. Three edges on the arc are identified – start, middle and end. Each of these can be marked as traversed, when visited by a driver. Active arc detection starts with checking whether the driver's distance to the traversable arc is less than a predefined constant – arc tolerance. If it is true, it is further evaluated if the driver's current distance to the next traversable edge (arc start, middle or end) is less than a predefined distance – vertex tolerance. If it is true, the edge (e. g., middle) is marked as finished. Edges can be marked as finished only by going from the start to the end. When all three edges are finished, the arc is marked as finished. Arc and vertex tolerance constants are specified by the route generation WS and used for dealing with problems caused by spatial data and GPS inconsistencies.

IV. EXPERIMENTAL RESULTS

The experiments to compare the GA and the AA were conducted on a 1491 node large dataset with arc total distance of 72364.80 m, including virtual arcs – 85176.65 m. While calculating the path cost, the average driving speed is defined as 50 km/h.

The gathered results from using GA and AA on two different graph models are presented in Table I, the corresponding turn type counts are presented in Table II. The effectiveness of both approaches is evaluated in terms of path cost. Since both solutions use the same balanced graph, minimizing path cost also minimizes turn costs (chooses preferred maneuver types when possible).

TABLE I
GREEDY AND APPROXIMATE SOLUTION COMPARISON

	Greedy solution		Approximate solution	
	Expanded route	Replaced restrictions	Applied patching	Replaced restrictions
Distance (m)	85176.65	101954.76	85176.65	103705.00
Turn costs (sec)	17075.00	17935.00	14020.00	12715.00
Prohibited turns	33	17	25	0
Path cost (h)	6.46	7.03	5.61	5.34

Total path distance for both approaches before restriction replacement is equal as both algorithms at this phase visit corresponding graph arcs or nodes (ELG nodes) at a single time. Turn costs are lower for the AA because of the applied optimization while patching subcycles.

Prohibited turns are not fully replaced in GA. Using the simple graph model inserting shortest path replacements or expanding virtual arcs can create new prohibited turns in the start and end of the path. Since the ELG model has the information about the previous turn made, it is possible to replace every restriction and expand virtual arcs without creating new restrictions.

Both algorithms favor driving straight, therefore straight driving has mainly been chosen in both cases. GA right and left turn count is almost the same, while AA has used more right turns. GA with Fleury's algorithm has more situations where there are possible better turns to make but you have to

choose a more expensive non-bridge arc just to finish the algorithm execution.

TABLE II
TURN COUNT COMPARISON

Turn type	Turn cost	Greedy solution	Approximate solution
Straight	0	1746	1428
Right	5-15	203	232
Left	55-65	204	141
U-turn	80	3	8
Prohibited turn	240	17	0

AA takes more time to execute because solving the optimal assignment while patching is a time-consuming task; however, because of the transformed graph model and even using a partial optimization, AA path cost is lower than GA, thus the resulting route is considered to be more efficient.

V. CONCLUSION

This paper develops the STORN that provides OTP and the execution of the corresponding plan using a series of loosely coupled web services. The high level architecture and database model for the solution are defined. Two alternative OTP generation algorithms were tested and the AA was proved to be superior.

A mashup is created based on the provided web services and field tested using a data acquisition case in Riga, Latvia. Currently available feedback has shown that vertex and arc tolerance distances play an important role in city traversal and need to be adjusted by field testing (constants may vary for different cities). These are currently specified manually; however, a web service providing automatic calibration of these parameters should be developed. Another issue is dealing with road obstacles and construction works that are not considered during route planning and are discovered only while traversing the route. Currently the user interface of the mashup contains a fast-forward button to manually mark the next arcs as visited, but a more complex approach should be used (e. g., automatic route recalculation by using real-time data provided by the driver).

REFERENCES

- [1] G. Liu, Y. Ge, T. Z. Qiu, and H. R. Soleymani, "Optimization of snow plowing cost and time in an urban environment: A case study for the city of Edmonton," *Canadian Journal of Civil Engineering*, vol. 41, 2014, pp. 667–675. <http://dx.doi.org/10.1139/cjce-2013-0409>
- [2] S. S. Chawathe, "Organizing Hot-Spot Police Patrol Routes," in *Intelligence and Security Informatics*, 2007 IEEE, 2007, pp. 79–86.
- [3] L. Xintong and P. Feng, "The shortest path and spatial decision support system implementation in the context of parcel delivery," in *Geoinformatics, 2010 18th International Conference*, 2010, pp. 1–6.
- [4] M. Saadatseresht, A. Mansourian, and M. Taleai, "Evacuation planning using multiobjective evolutionary optimization approach," *European Journal of Operational Research*, vol. 198, 2009, pp. 305–314. <http://dx.doi.org/10.1016/j.ejor.2008.07.032>
- [5] S. Marston, Z. Li, S. Bandyopadhyay, J. Zhang, and A. Ghalsasi, "Cloud computing – The business perspective," *Decision Support Systems*, vol. 51, Apr. 2011, pp. 176–189. <http://dx.doi.org/10.1016/j.dss.2010.12.006>
- [6] J. Gao, P. Pattabhiraman, X. Bai, and W. T. Tsai, "SaaS performance and scalability evaluation in clouds," in *6th IEEE International Symposium on Service-Oriented System Engineering, SOSE 2011, Irvine, CA, 2011*, pp. 61–71.

- [7] C. Yang, M. Goodchild, Q. Huang, D. Nebert, R. Raskin, Y. Xu, et al., "Spatial cloud computing: How can the geospatial sciences use and help shape cloud computing?," *International Journal of Digital Earth*, vol. 4, 2011, pp. 305–329. <http://dx.doi.org/10.1080/17538947.2011.587547>
- [8] A. Benlian and T. Hess, "Opportunities and risks of software-as-a-service: Findings from a survey of IT executives," *Decision Support Systems*, vol. 52, pp. 232–246, Dec. 2011. <http://dx.doi.org/10.1016/j.dss.2011.07.007>
- [9] S. S. Qureshi, T. Ahmad, K. Rafique, and I. Shuja Ul, "Mobile cloud computing as a future for mobile applications – Implementation methods and challenging issues," in *2011 IEEE International Conference on Cloud Computing and Intelligence Systems, CCIS2011, Beijing*, 2011, pp. 467–471.
- [10] M. Shiraz, A. Gani, R. H. Khokhar, and R. Buyya, "A review on distributed application processing frameworks in smart mobile devices for mobile cloud computing," *IEEE Communications Surveys and Tutorials*, vol. 15, pp. 1294–1313, 2013. <http://dx.doi.org/10.1109/SURV.2012.111412.00045>
- [11] A. Corbérn and C. Prins, "Recent results on arc routing problems: An annotated bibliography," *Networks*, vol. 56, pp. 50–69, 2010.
- [12] H. Thimbleby, "The directed Chinese postman problem," *Software – Practice and Experience*, vol. 33, pp. 1081–1096, 2003. <http://dx.doi.org/10.1002/spe.540>
- [13] E. Minieka, "Chinese postman problem for mixed networks," *Management Science*, vol. 25, pp. 643–648, 1979. <http://dx.doi.org/10.1287/mnsc.25.7.643>
- [14] L. Kazemi, C. Shahabi, M. Sharifzadeh, and L. Vincent, "Optimal traversal planning in road networks with navigational constraints," in *GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems*, 2007, pp. 138–145.
- [15] F. Dorn, H. Moser, R. Niedermeier, and M. Weller, "Efficient algorithms for eulerian extension and rural postman," *SIAM Journal on Discrete Mathematics*, vol. 27, pp. 75–94, 2013. <http://dx.doi.org/10.1137/110834810>
- [16] D. Soler, E. Martínez, and J. C. Micó, "A transformation for the mixed general routing problem with turn penalties," *Journal of the Operational Research Society*, vol. 59, pp. 540–547, 2008. <http://dx.doi.org/10.1057/palgrave.jors.2602385>
- [17] J. Clossey, G. Laporte, and P. Soriano, "Solving arc routing problems with turn penalties," *Journal of the Operational Research Society*, vol. 52, pp. 433–439, 2001. <http://dx.doi.org/10.1057/palgrave.jors.2601052>
- [18] S. Almoustafa, S. Hanafi, and N. Mladenović, "New exact method for large asymmetric distance-constrained vehicle routing problem," *European Journal of Operational Research*, vol. 226, pp. 386–394, 2013. <http://dx.doi.org/10.1016/j.ejor.2012.11.040>
- [19] G. Jäger and W. Zhang, "A SAT based effective algorithm for the directed Hamiltonian cycle problem," in *5th International Computer Science Symposium in Russia, CSR 2010 vol. 6072 LNCS, ed. Kazan*, 2010, pp. 216–227.
- [20] M. Turkensteen, D. Ghosh, B. Goldengorin, and G. Sierksma, "Iterative patching and the asymmetric traveling salesman problem," *Discrete Optimization*, vol. 3, pp. 63–77, 2006. <http://dx.doi.org/10.1016/j.disopt.2005.10.005>
- [21] B. Freisleben and P. Merz, "Genetic local search algorithm for solving symmetric and asymmetric traveling salesman problems," in *Proceedings of the IEEE Conference on Evolutionary Computation*, 1996, pp. 616–621. <http://dx.doi.org/10.1109/ICEC.1996.542671>
- [22] "Downloading data," Aug. 9, 2014. [Online]. Available: http://wiki.openstreetmap.org/wiki/Downloading_data [Accessed: Oct. 5, 2014].
- [23] J. Varia, "Architecting for the Cloud: Best Practices," 2001. [Online]. Available: http://media.amazonwebservices.com/AWS_Cloud_Best_Practices.pdf [Accessed: Sept. 20, 2014].
- [24] M. Hamdaqa, T. Livogiannis, and L. Tahvildari, "A reference model for developing cloud applications," in *CLOSER 2011 – Proceedings of the 1st International Conference on Cloud Computing and Services Science*, 2011, pp. 98–103.
- [25] C. Inzinger, S. Nastic, S. Sehic, M. Vogler, F. Li, and S. Dustdar, "MADCAT: A methodology for architecture and deployment of cloud application topologies," in *8th IEEE International Symposium on Service Oriented System Engineering, SOSE 2014, Oxford*, 2014, pp. 13–22.
- [26] Y. W. Kwon and E. Tilevich, "Cloud refactoring: Automated transitioning to cloud-based services," *Automated Software Engineering*, vol. 21, pp. 345–372, 2014. <http://dx.doi.org/10.1007/s10515-013-0136-9>
- [27] B. P. Rimal, A. Jukan, D. Katsaros, and Y. Goeleven, "Architectural Requirements for Cloud Computing Systems: An Enterprise Cloud Approach," *Journal of Grid Computing*, vol. 9, pp. 3–26, 2011. <http://dx.doi.org/10.1007/s10723-010-9171-y>
- [28] R. Geisberger and C. Vetter, "Efficient routing in road networks with turn costs," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* vol. 6630 LNCS, ed. 2011, pp. 100–111.
- [29] D. Luxen, "Flexible route guidance through turn instruction graphs," in *MapInteract 2013 – Proceedings of the 1st ACM SIGSPATIAL International Workshop on MapInteraction*, 2013, pp. 78–83.

Jānis Kampars obtained his Doctoral Degree in IT from RTU, Latvia in 2012. He works as an Assistant Professor at RTU and his areas of interest are application integration, web development and cloud computing.
E-mail: Janis.Kampars@rtu.lv.

Elīna Šmite has earned her B. Sc. (in the year 2013) degree in IT from RTU, Latvia. She has been working at RTU since 2013 as a Junior Researcher. Her areas of interest are web/mashup technologies.
E-mail: Elina.Smite@rtu.lv.

Jānis Kampars, Elīna Šmite. RCTI: izbraucama ceļu tīkla risinājums

Optimāla izbraukšanas plāna (OIP) mērķis ir noteikt maršrutu, kas nodrošinātu visu ielu izbraukšanu iepriekš noteiktā apgabalā. Risinot OIP problēmu, īpaša vērība jāpievērš kopējā maršruta izbraukšanas laika minimizēšanai. Svarīgs faktors ir maršruta ģenerēšanas laiks, kas strauji pieaug līdz ar ģeogrāfiskā apgabala palielināšanu. Ģenerējot maršrutu, ir jāņem vērā, ka dažādi pagriezieni var būt aizliegti un ka divvirzienu ielas ir nepieciešams izbraukt abos virzienos. Iespējamās jomas OBP izmantošanai ietver ielu tīrīšanu, pasūtījumu piegādi, evakuācijas plānošanu, policijas patrulēšanas plānošanu un citas. OIP noteikšanai no OpenStreetMaps tiek iegūti telpiskie dati, kas tiek pārveidoti grafa veidā (ielas segments atbilst grafa lokam, bet krustojums – virsotnei). Šāda grafa pilnīgai šķērsošanai ir nepieciešams apmeklēt visas tā malas vismaz vienu reizi. Rakstā tiek aplūkoti un eksperimentāli novērtēti divi atšķirīgi OIP izbraukšanas plāna iegūšanas algoritmi. Balstoties uz šiem algoritmiem, tiek definēts Risinājums ceļu tīkla izbraukšanai (RCTI). RCTI ir mākonī bāzēta sistēma, kurai piekļuve tiek nodrošināta ar vairāku tīmekļa servisu palīdzību. Tīmekļa serviss var izmantot jaunu izbraucamo reģionu definēšanai, maršrutu ģenerēšanai, maršrutu izbraukšanas atbalsta nodrošināšanai un maršruta izsekošanai. Rakstā ir definēta RCTI arhitektūra un datubāzes modelis. RCTI prototipa darbība ir tikusi testēta reālos apstākļos Rīgā, Latvijā, veicot maršruta ģenerēšanu un daļēju izbraukšanu.

Янис Кампарс, Элина Шмите. РПДС: Решение дорожных сетей для проезда

Цель плана оптимального проезда (ПОП) – обнаружить маршрут, который обеспечит быстрое проезд по всем улицам на заранее определенном участке. Решая проблему ПОП, особое внимание следует обратить на минимизацию времени проезда по всему маршруту. Ещё один важный фактор – время разработки маршрута, которое быстро возрастает вместе с увеличением соответствующего географического участка. Разрабатывая маршрут, нужно принимать во внимание, что некоторые повороты могут быть запрещены, и что на двусторонних улицах движение идёт в двух направлениях. Возможные области применения ПОП: чистка улиц, доставка посылок, планирование эвакуации, планирование маршрута полицейского патруля и другие. Для определения ПОП с OpenStreetMaps собираются пространственные данные, которые превращаются в графу (сегмент улицы соответствует дуге графы, а пересечение – вершине). Для полного обхода такой графы необходимо посетить все её стороны, по крайней мере, один раз. В статье рассмотрены и экспериментально оценены два разных алгоритма получения плана проезда ПОП. Основываясь на этих алгоритмах, определяется решение для проезда дорожных сетей (РПДС). РПДС – это система, расположенная в облаке, доступ к которой обеспечен с помощью множества Web-сервисов. Web-сервисы можно использовать для определения новых участков для проезда, разработки маршрута, обеспечения поддержки проезда и отслеживания маршрута. В статье определена РПДС архитектура и модель базы данных. Функционирование прототипа РПДС протестировано в реальных условиях – был разработан маршрут и частичный проезд в городе Рига (Латвия).

Evaluation of the Ontological Knowledge Model

Darya Plinere¹, Arkady Borisov²,^{1,2} *Riga Technical University*

Abstract – An ontology is an explicit formal conceptualization of some domain of interest. Ontology application is widely used in multi-agent systems. Ontology provides the channel through which software agents interact; therefore, if the ontology definitions have not been sufficiently evaluated, communication between software agents may not succeed. Ontology evaluation guarantees to end users the correctness and completeness of ontology definitions and software. The goal is to detect wrong, incomplete or missed definitions in the ontology. The study is based on multi-agent system application for raw materials management task. The common ontology is used for agent cooperation; the quality of ontology is measured with ontology evaluation techniques.

Keywords – Application-based ontology evaluation, data-driven ontology evaluation, golden standard, ontology evaluation.

I. INTRODUCTION

Ontology evaluation task is to measure the quality of the ontology. Ontology makes sure that the meaning of the data exchanged between and within systems is consistent and shared by computers (expressed by formal models) and humans (as given by their conceptualization). Ontology allows all participants to speak a common language [1].

Ontology, as well as all engineering artifacts, needs a thorough evaluation. However, ontology evaluation poses some unique challenges: due to the declarative ontology nature developers can not just compile and run it like most of other software artifacts. This is data that have to be shared between various components and used for potentially different tasks [1].

Its role is to provide a formal semantics to terms in order to use them in the machine processable way. Ontology allows to share and formalize conceptualizations, and thus to enable humans and machines to easily understand the meaning of the data exchanged. This allows automatically aggregate, use and reuse distributed data sources, thereby creating an environment in which agents and applications can work together for the benefit of the user to still inexperienced level [1].

The central role of ontology in the agent system makes the evaluation of the ontology an important and worthwhile task: mistakes or omissions in the ontology can lead to not realizing the full potential of the data exchanged in the application. Good ontology leads to a higher degree of reuse and better interaction [1].

Some examples of disadvantages of the low-quality ontology: ontology readability may be affected if the vocabulary or syntax contains errors; reasoners may not be able to infer the answers in case of conflicting semantics. Unspecified ontology prevents automated ontology mapping approaches. On the other hand, high quality ontology can be easily reused, can be featured more easily in the existing

application, and will be easier to detect and actively omit the errors in the data [1].

II. RAW MATERIALS MANAGEMENT TASK FOR ONTOLOGY EVALUATION TECHNIQUE APPLICATION

This paper presents a study based on the use of multi-agent system for the task of raw material management. The example from the field of microelectronics, in particular, the company producing chips is discussed in this paper. Chip production requires raw materials, for example, crystals timely saturation in the warehouse. Today, the purchase of crystals is made by the purchasing department manager by phone, fax and e-mail.

Multi-agent systems are widely used for the tasks of supply chain management, particularly for raw materials management tasks. Multi-agent system provides a decentralized system; the global behavior of the system is determined by individual behavior of agents who use their own behavior rules, exist in a shared environment and interact with the environment and other agents. This makes multi-agent system most suitable for this task. Agents use a common ontology or taxonomy for negotiations, thus allowing agents to understand the substance of the negotiations. Ontology in the multi-agent system provides the channel through which software agents interact; therefore, if the ontology definitions have not been sufficiently evaluated, communication between software agents may not succeed. The outputs of application or its performance depend on used ontology in it. Ontology evaluation guarantees to end users the correctness and completeness of ontology definitions and software [2].

The following infrastructure was used for constructing the system: Java and JADE – platform in which agents exist and interact, it gives agents the basic services necessary to their existence, Protégé ontology editor for ontology development, Ontology Bean Generator to convert the domain ontology in JADE classes, MySQL for database support, Apache Ant to compile the code and NetBeans IDE as an integrated development environment [2].

III. CLASSIFICATION OF ONTOLOGY EVALUATION APPROACHES

Ontologies are not artifacts in a narrow sense, but are expressed by ontology documents, which in turn are artifacts. Evaluation methods are descriptions of procedures that assess a specific quality of an ontology. Since methods cannot assess an ontology directly (since they are not artifacts), methods always directly evaluate ontology documents.

Only indirectly it is possible for an evaluation method to assess an ontology (i. e., by assessing the ontology document that expresses the ontology) [1].

Various approaches to the evaluation of ontologies have been considered in the literature, depending on what kind of ontologies is evaluated and for what purpose. Broadly speaking, most evaluation approaches fall into one of the following categories [3], [4]:

- those based on comparing the ontology to a “golden standard”, which may itself be an ontology [5], [6];
- those based on using the ontology in an application and evaluating the results [7];
- those involving comparisons with a source of data (e. g., a collection of documents) about the domain to be covered by the ontology [8];
- those where evaluation is done by humans, who try to assess how well the ontology meets a set of predefined criteria, standards, requirements, etc. [9], [10], [11].

TABLE I
EVALUATION APPROACHES IN DIFFERENT LEVELS

Level	Golden standard approach	Application-based approach	Data-driven approach	Assessment by humans
Lexical, vocabulary, concept, data	+	+	+	+
Hierarchy, taxonomy	+	+	+	+
Other semantic relations	+	+	+	+
Context, application		+		+
Syntactic	+			+
Structure, architecture, design				+

In addition to the categories of evaluation above, ontology evaluation approaches can be grouped based on the level of evaluation, as described below. An ontology is a fairly complex structure and it is often more practical to focus on the evaluation of different levels of the ontology separately rather than trying to directly evaluate the ontology as a whole. The individual levels have been defined variously by different authors, but these various definitions tend to be broadly similar and usually involve the following levels:

- Lexical, vocabulary, or data layer;
- Hierarchy or taxonomy;
- Other semantic relations;
- Context or application level;
- Syntactic level;
- Structure, architecture, design.

Table I summarizes which approaches from the categories above are commonly used for which of these levels. The detailed description of ontology evaluation approaches based on the level of evaluation is described in [3].

IV. AN OVERVIEW AND APPLICATION OF THE APPROACHES

A. Data-driven Approach

Data-driven approach is that, where the ontology can be evaluated by comparing it to textual documents about the domain that is to be covered by ontology. The authors [8] proposed to compare one or more ontologies with corpus of documents, rather than to compare one ontology to another in order to find the most appropriate one. They extracted a set of relevant domain-specific terms from the corpus of documents, applying Latent Semantic Analysis and clustering method. Then they used WordNet to add two levels of hypernyms to each term in a cluster. And finally, the set of terms is identified in the corpus, mapped to the ontology. The amount of overlap between the domain-specific terms and the terms appearing in ontology can then be used to measure lexical keyword coverage by ontology labels. The ontology can be penalized for terms present in the corpus and absent in the ontology, and for terms present in the ontology, but absent in the corpus. They also proposed a “tennis measure” for two ontologies with the set of identical concepts, which have different organization of concepts, thus have different distance from each other [8], [3].

The document of raw material procurement process is available. Five different ontologies were constructed for the evaluation approach. Having the corpus and five ontologies, the comparison of each one to the corpus was done.

First of all, two questions were asked: how many words from the corpus were in the ontology? How many were not? The overlap between the domain-specific terms and the terms appearing in ontology used to measure lexical keyword coverage by ontology labels. We used the precision and recall method. Precision (1) in this context is the percentage of the ontology lexical entries that also appear in the corpus, relative to the total number of ontology words. Recall (2) is the percentage of corpus lexical entries that also appear as concept identifiers in ontology, relative to the total number of corpus lexical entries.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (1)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2)$$

Where TP = true positive, FN = false negative, FP = false positive, TN = true negative.

Table II shows the results of precision and recall method for five developed ontologies.

TABLE II
PRECISION AND RECALL METHOD RESULTS FOR FIVE ONTOLOGIES

	1	2	3	4	5
Precision	64%	41%	59%	72%	32%
Recall	84%	50%	72%	89%	41%

The next step was to find similarity in concept spelling. One set of concepts was compared to the set of corpus terms. The second step was performed by Levenshtein edit distance method. Edit distance – the minimum edit distance between two strings is the minimum number of editing operations needed to transform one into the other, like insertion, deletion or substitution.

The Levenshtein edit distance formula and examples can be found in [12].

TABLE III
LEVENSHTEIN EDIT DISTANCE FOR FIVE ONTOLOGIES

	1	2	3	4	5
Levenshtein edit distance	16	30	14	16	0

The results in Table II and Table III have shown that the fourth ontology is the most appropriate for a current task of raw material procurement.

B. Application-based Approach

Ontology in the multi-agent system provides the channel through which software agents interact; therefore, if the ontology definitions have not been sufficiently evaluated, communication between software agents may not succeed. The authors [7] propose evaluating ontologies by putting them into the application and evaluating the results. Good ontology is one, which produces good results on the given task. This approach has several drawbacks: 1) it can be seen that the ontology is good or bad when used in a particular way for a particular task, but it is difficult to generalize this observation; (2) the ontology could be only a small component of the application and its effect on the outcome may be relatively small and indirect; (3) comparing different ontologies is only possible if they can all be plugged into the same application.

Five ontologies were developed in Protégé ontology editor, then converted by Ontology Bean Generator in JADE classes; NetBeans IDE as an integrated development environment was executed in order to evaluate the produced results.

Five ontology-based agent systems have shown the following results:

1. The use of the first, second, third and fifth ontology is not enough for solving the problem.
2. The fourth ontology has shown the best evaluation results, cooperation between agents is achieved (see Fig. 1).

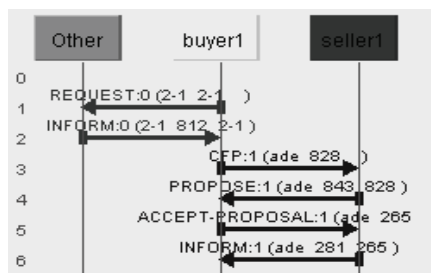


Fig. 1. Agent cooperation in JADE, using the fourth ontology.

C. Golden Standard Approach

The idea of golden standard approach is in comparison with concepts in the evaluated ontology to “golden standard”, which can be another ontology, considered a good representation of the concepts of the problem domain under consideration, or it could be taken from a corpus of documents, or prepared by a domain expert. Evaluation of the ontology can also be based on the precision and recall method, comparing ontology with a human-provided golden standard. In this context, precision is the fraction of the labels that also appear in the golden standard relative to the total number of labels. Recall is the percentage of the golden standard lexical entries that also appear as labels in the ontology, relative to the total number of golden standard lexical entries. Similarity between strings can be measured by the Levenshtein edit distance. The authors [5] propose several measures, such as the semantic cotopy of two hierarchies, for comparing structural aspect of two ontologies. Evaluation of an ontology on the semantic aspect can also be based on precision and recall measures [1]. The drawback of golden standard approach is the requirement for a lot of manual human work. However, once the golden standard is defined, comparison of two ontologies can proceed entirely automatically [3].

The fourth ontology was taken as a golden standard and was compared with others.

TABLE IV
PRECISION AND RECALL MEASURES FOR GOLDEN STANDARD APPROACH

	1	2	3	5
Precision	100%	68%	86%	41%
Recall	77%	68%	86%	32%

The results depicted in Table IV and Levenshtein edit distance have shown that the first ontology and the third ontology have better results than the second and the fifth ontology, having the fourth ontology as a golden standard.

D. Predefined Criterion Ontology Evaluation

This approach of ontology evaluation deals with the selection of a good ontology from a variety of ontologies as a decision-making problem. For the evaluation of ontology, several decision criteria or attributes must be defined and a numerical score for each criterion is given. An overall score for the ontology is then calculated as a weighted sum of its per-criterion scores. A drawback is that a lot of manual involvement by human experts may be needed.

Ontology evaluation can address a number of several different criteria. Therefore, the first task of the evaluator is to choose the criteria relevant for the given evaluation and then the proper evaluation methods to assess how well the ontology meets these criteria [1].

Different criteria and methods were analyzed by [1]. Five criteria have been chosen from this literature survey: accuracy, clarity, completeness, conciseness and consistency (see Fig. 2).

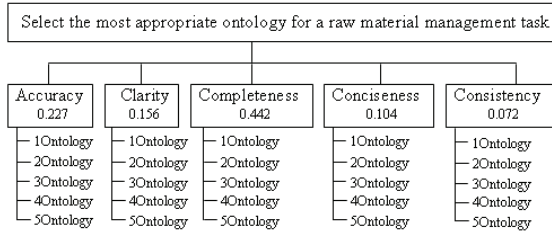


Fig. 2. A hierarchy tree with criteria calculated eigenvector.

• **Accuracy:** A higher accuracy comes from correct definitions and descriptions of classes, properties, and individuals.

• **Clarity:** Definitions should be objective and independent of the context. When a definition can be stated in logical axioms, it should be. All entities should be documented with a natural language.

• **Completeness:** Completeness measures if the domain of interest is appropriately covered. All the knowledge that is expected to be in the ontology is either explicitly stated or can be inferred from the ontology.

• **Conciseness:** Conciseness is the criteria that states if the ontology includes irrelevant elements with regard to the domain to be covered or redundant representations of the semantics.

• **Consistency:** Consistency describes that the ontology does not include or allow for any contradictions.

The authors [9] offer Ontometric – applying the Analytic Hierarchy Process (AHP) in the Ontology Choice. It can be used to: 1) select the most appropriate ontology among various alternatives or, 2) decide on the suitability of a particular ontology for the project.

Taking into account the general steps of AHP, described in [7], they have adapted the method to be used in the reuse of ontologies:

Step 1: specify the objective of the problem – “Select the most appropriate ontology for a raw materials management task”;

Step 2: build a hierarchy tree in this way: the root node is the objective of the problem, the intermediate levels are the criteria, and the lowest level contains the alternatives.

Step 3: for each set of brother nodes, make the pairwise comparison matrices with the criteria of the decision tree. For each comparison matrix, an eigenvector must be calculated.

Step 4: for each alternative ontology, assess its characteristics. For each one of these characteristics, the engineer should establish a scale of appropriate ratings.

Step 5: lastly, combine the vectors of weights obtained in step 3 with the values of the alternatives. Finally, the suitable ontology is chosen based on the results obtained.

Please refer to [9] for a better understanding of AHP method and Ontometric method.

TABLE V
APPLIED AHP METHOD FOR SELECTING BEST ONTOLOGY

	Final Result
Ontology1	0.575
Ontology2	0.305

Ontology3	0.551
Ontology4	0.852
Ontology5	0.340

The results in Table V have shown that the fourth ontology is the most appropriate one.

V. RESULTS

The study of ontology evaluation techniques has shown that evaluation approaches can be categorized by evaluation means: data-driven, application-driven, by humans or comparing to golden standard, or based on the level of evaluation: lexical, hierarchical, syntactical etc. Five different ontologies were developed for the multi-agent system for raw materials management task:

- 1) The first ontology consists only of hierarchical relationships – taxonomy;
- 2) The second ontology with grammatical mistakes;
- 3) The third ontology is incomplete;
- 4) The fourth ontology is complete;
- 5) The fifth ontology refers to a chip manufacturing domain area.

These five ontologies were evaluated by four approaches, and the following results were achieved:

1) Data-driven approach showed that the most appropriate ontology for the task was the fourth one, because of its highest indicators of precision and recall measures comparing corpus terms with ontology concepts and because of Levenshtein edit distance measure in concept spelling similarity.

2) Application-based approach could show the performance of these five ontologies applied in a real multi-agent system. As a result – only the fourth ontology could be used in an agent system, because of its completeness, others – due to incompleteness and grammatical mistakes did not show good performance.

3) The fourth ontology was taken as the expert-provided golden standard and was compared with other four ontologies. Better results were shown by taxonomy and incomplete ontology.

4) Predefined criteria ontology evaluation had the following result – the fourth ontology was the most appropriate ontology using the five criteria in its evaluation by means of an analytical hierarchy process.

VI. CONCLUSION

This paper has proposed a study of ontology evaluation techniques. Practical experiments were based on the developed ontologies for a multi-agent system, solving raw materials management task.

The ontology evaluation results have shown that only complete, with high accuracy developed ontology can be applied in ontology-based systems for its performance; otherwise, ontology application cannot offer its full potential in an agent system.

REFERENCES

- [1] D.Vrandečić, "Ontology Evaluation," PhD thesis, Karlsruher Institut für Technologie, 2010.
- [2] D. Plinere, A. Borisov, "A Negotiation-Based Multi-Agent System for Supply Chain Management," *IT and Management Science*, vol. 49, pp.128–132, 2011.
- [3] J.Branc, M.Grobelnik, D. Mladenec, "A survey of ontology evaluation techniques," *Conference on Data Mining and Data Warehouses (SiKDD 2005)*, 2005.
- [4] J. Branc, D. Mladenec, M. Grobelnik, "Automatic Evaluation of Ontologies," In: Kao, A., Poteet, S. (eds.), *Text Mining and Natural Language Processing*, Springer, 2006.
- [5] A. Meadche, S. Staab, "Measuring Similarity between Ontologies," *European Conference Knowledge Acquisition and Management (EKAW)*, pp. 251–263, Oct. 1–4, 2002, Madrid, Spain.
- [6] K. Dellschaft, S. Staab "On How to Perform a Gold Standard Based Evaluation of Ontology Learning," *In International Semantic Web Conference*, pp. 228–241, 2006.
- [7] R. Porzel and R. Malaka, "A Task-Based Approach for Ontology Evaluation," *ECAI Workshop Ontology Learning and Population (OLP)*, Aug. 2004, Valencia, Spain.
- [8] C. Brewster, H. Alani, S. Dasmahapatra, and Y. Wilks, "Data Driven Ontology Evaluation," *International Conference on Language Resources and Evaluation (LREC 2004)*, May 24–30, 2004, Lisbon, Portugal.
- [9] A. Lozano-Tello, A. Gómez-Pérez, "Ontometric: A method to choose the appropriate ontology," *Journal of Database Management*, vol. 15, issue 2, pp. 1–18, 2004. <http://dx.doi.org/10.4018/jdm.2004040101>
- [10] A. Gómez-Pérez, "Some ideas and examples to evaluate ontologies," Knowledge Systems Laboratory, Stanford University, 1994.
- [11] A. Gómez-Pérez, "Towards a framework to verify knowledge sharing technology," *Expert Systems with Applications*, no. 4, pp. 519–529, 1996. [http://dx.doi.org/10.1016/S0957-4174\(96\)00067-X](http://dx.doi.org/10.1016/S0957-4174(96)00067-X)
- [12] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics Doklady* vol. 10, no. 8, 1966, pp. 707–710.

Darya Plinere received her *Mg. sc. ing* from Riga Technical University in 2006. She is a Researcher at the Faculty of Computer Science and Information Technology, Riga Technical University. Her research interests include multi-agent systems, ontology engineering and supply chain management. E-mail: darja.plinere@rtu.lv.

Arkady Borisov received his Doctoral degree in Technical Cybernetics from Riga Polytechnic Institute in 1970 and *Dr. habil. sc. comp.* degree in Technical Cybernetics from Taganrog State Radio Engineering University in 1986. He is a Professor of Computer Science at the Faculty of Computer Science and Information Technology, Riga Technical University. His research interests include fuzzy sets, fuzzy logic, computational intelligence and bioinformatics. He has 235 publications in the field. He is a member of IFSA European Fuzzy System Working Group, Russian Fuzzy System and Soft Computing Association, honorary member of the Scientific Board, member of the Scientific Advisory Board of the Fuzzy Initiative Nordrhein-Westfalen (Dortmund, Germany). Address: 1 Kalku Str., Riga, LV-1658; phone: +371 67089530 E-mail: arkadijs.borisovs@cs.rtu.lv.

Darja Plinere, Arkādijs Borisovs. Ontoloģiskā zināšanu modeļa novērtēšana

Šajā rakstā ir izpētīta ontoloģijas novērtēšana, uzbūvēta izejvielu pārvaldības uzdevumam, pielietojot multi-aģentu sistēmu. Ontoloģija aģentu sistēmas nodrošina ar vienādu izpratni par informāciju, ar kuru apmainās aģenti, risinot uzdevumu. Ontoloģija ir kanāls, caur kuru aģenti mijiedarbojas. Līdz ar to, sistēmas efektīgai darbībai ir nepieciešams novērtēt uzbūvētās ontoloģijas kvalitāti un tās piemērotību uzdevumam. Ontoloģiskā zināšanu modeļa novērtēšanas pētījuma rezultātā tika identificētas četras pieejas novērtēšanai – vairāku kritēriju, salīdzinājums ar zelta standartu, balstoties uz datiem un balstoties uz pielietojumu pieeju. Pieejas var arī klasificēt attiecībā uz novērtējuma līmeni – sintakses, hierarhijas, attiecības u.c. Lai veiktu pētījumu, tika uzbūvētas piecas ontoloģijas, un katra no tām tika pakļauta novērtēšanai četrās pieejās. Pirmajai ontoloģijai ir tikai hierarhiskās attiecības (taksonomija), otrā ir – ontoloģija ar gramatiskām kļūdām, trešā – neapver visu priekšmeta apgabalu (nepabeigta), ceturta – aptver visu priekšmeta apgabalu un piektā – ir ontoloģija, kas saistīta ar ražošanu, nevis ar izejvielu pasūtījumu. Pieeja, kas balstīta uz datiem, salīdzināja pasūtījuma procesa dokumentu un uzbūvēto ontoloģiju. Tika izmantotas šādas metodes – precizitāte un pilnīgums, lai novērtētu priekšmeta apgabala segumu ar ontoloģiju, kā arī Levenšteina attālumu, lai noteiktu, cik lielā mērā ontoloģijas koncepti sakrīt ar dokumenta terminiem. Pieejā, kas balstīta uz pielietojumu, ontoloģija ir iekļauta aģenta sistēmā, un aģentu mijiedarbības rezultāti tika pētīti. Tikai ontoloģija, kas konstruēta ar augstu precizitāti un kas sedz visu priekšmeta apgabalu, var dot rezultātus no tās izmantošanas. Pēc tam tika izdarīta salīdzinošā analīze ar zelta standartu un noteikta daudz-kritēriju ontoloģijas kvalitāte. Salīdzinājums ar zelta standartu tika veikts ar tām pašām metodēm, kā pieejā, kas ir balstīta uz datiem. Daudz-kritēriju pieejā tika izmantots algoritms – analītiskais hierarhiskais process. Visas četras pieejas novērtēja ceturto ontoloģiju kā vispiemērotāko ontoloģiju šim uzdevumam.

Дарья Плинере, Аркадий Борисов. Оценка качества онтологической модели знаний

В данной статье исследуется оценка качества онтологии, построенной для решения задачи управления материальными ресурсами предприятия с помощью многоагентной системы. Онтология в агентной системе обеспечивает одинаковое понимание агентами информации, которой они обмениваются для решения поставленных задач. Онтология представляет собой канал, через который агенты взаимодействуют. Следовательно, для функционирования системы необходимо оценить качество предлагаемой онтологии и ее пригодность к данной задаче. В результате исследования оценки качества онтологической модели знаний были выявлены четыре подхода оценки качества – многокритериальный, сравнение с золотым стандартом, основанный на данных и основанный на применении. Подходы также можно классифицировать относительно уровня оценки – синтаксис, иерархия, отношения и т.п. Для проведения исследования было построено пять онтологий, каждая из них подвергалась оценке четырьмя подходами. Первая онтология имеет только иерархические отношения (таксономия), вторая – онтология с грамматическими ошибками, третья – не покрывающая предметную область (неполная), четвертая – покрывающая всю предметную область, и пятая – онтология, относящаяся к производству, а не закупке сырья. Подход, основанный на данных, сравнил документ, описывающий процесс закупки сырья, и построенные онтологии. Использовались методы точности и полноты для оценки покрытия онтологией предметной области, а также расстояние Левенштейна для определения того, насколько концепты онтологии совпадают с терминами документа. В подходе, основанном на применении, онтологии были включены в агентную систему, и исследовались результаты переговоров между агентами. Только онтология, построенная с высокой точностью и покрывающая предметную область, способна дать результат от ее применения. Затем был произведен сравнительный анализ с золотым стандартом и проведена оценка качества онтологии многокритериальным способом принятия решения. Сравнение с золотым сечением было произведено теми же методами, что и в подходе, основанном на данных. В многокритериальном подходе использовался алгоритм – аналитический иерархический процесс. Во всех результатах четвертая онтология была оценена как наиболее подходящая для поставленной задачи.

Ontology-based Image Representation

Arthur Stephenko¹, Arkady Borisov², ¹ *Ventspils University College*, ² *Riga Technical University*

Abstract – This article presents an overview of ontology based digital image representation. An ontology is a specification of a conceptualization to create a vocabulary for exchanging information, where conceptualization mean a mapping between symbols used in the computer (i.e., the vocabulary) and objects and relations in the real world. In this paper, digital image semantic annotation by ontology and a novel ontological approach that formalizes concepts and relations with respect to image representations for data mining – the Image Representations Ontology (IROn) – are examined.

Keywords – Digital image, IROn, ontology.

I. INTRODUCTION

The interpretation of an image by the computer is a highly complex task. There is a huge gap between the human and computational understanding of images and its interpretation. The representation of images by low-level features (such as color and texture) is described as well as the use of high-level features such as ontologies. An image is typically a representation of the objects present in a real life scene [1]. In the process of image acquisition, much of the original information present on the real object is lost (e. g., third dimension, motion), there is mixture of several factors in the value of a pixel (e. g., texture, lighting, geometry), introduction of false values (e. g., noise, chromatic aberration), and alteration of the original information (e. g., geometric distortion, blurring) [2]. In addition, fine details of the object are missing because of the limited resolution of any camera. Therefore, the image is only the approximation of a real object and there is always an error that separates the image and the real object.

Images can be used to infer information about the real object [1]. Image content does not make sense by itself. An image is inherently ambiguous and does not provide information about its content. Without a subject matter, an image does not allow making a distinction between relevant and irrelevant information. What is relevant for one application may be irrelevant for another. Furthermore, there is no intrinsic relevant information. For example, apparently simple information such as object edges is difficult to accurately extract without knowledge about the scene. Edges are usually modeled as sharp changes in image intensity, but this is also the case for noise, shadows, and texture elements [2].

In order to be able to perform the computational processing of images, it is necessary first to convert them into digital format. This is done by representing the image as a set of discrete numbers in a particular order. In a digital photo, the image is represented as a matrix of values, called pixels. Mathematically, an image can be represented as a function

$$I = f(x, y) \quad (1)$$

where I is the image intensity, f is a function that varies with the position x and y in matrix [1].

Image processing is the low-level part of a more global image analysis and computer vision system. Then high-level part uses image-processing results to perform interpretation, visualization, storage, or transmission of the image data. Image processing includes many categories of mathematical manipulations, such as image restoration, image enhancement, image compression, image reconstruction, image segmentation, and object detection. An image class is a set of images that share many features, on condition that these features are meaningful for the application [2].

Remote sensing images are used in many decision-making domains. The addition of semantics to geographical information management is critical for the improvement of semantic interoperability and the usability of this type of images. An important step to strengthen the semantic interoperability of these images includes making clear the semantics associated with geographical information [8]. The semantic representation of geographical data provides a formal semantic description that cannot be expressed in current models of geographic data [6].

For example, an image analysis can be used for analysis of long-term agricultural or forest landscape changes [7]. The same landscape area is compared year by year, always at the same season, to detect changes of area. The image class is created of color satellite images of landscape with the same spatial resolution. The objective of the image processing here is to segment the input images to isolate each potential vegetation area into a unique region. The resulting regions or classes will then be transmitted into a classifier, which is trained to identify various categories of geographic objects: field, forest, hedge, city, etc. [9]. The performance of the classifier distinctly depends on the accuracy of the image segmentation. To obtain the best classification results, usually an annotated image database is used to fit the parameters that best separate each class presented in the data set [2].

II. BASIC ONTOLOGY CONCEPTS

Several different terminologies (e. g., RadLex, LOINC, ICD-10 and SNOMED) are available and can be used for digital image semantic annotation. The terminologies can be organized into hierarchies or ontologies. An ontology is a formal representation of a set of concepts within a domain and the relationships between these concepts [10]. Therefore, ontologies are an effective means to formally specify and constrain knowledge, so ontologies can be viewed as a means for semantic image annotation. The use of ontologies allows performing more complex tasks than it would be possible through a simple list of terms. They have proved their utility

in various data mining applications, especially in annotating text to make it machine interpretable.

More challenging research perspectives arise when ontologies are used to annotate images where the information is encoded in numeric pixel values rather than in words and language grammar. Ontology-based semantic image annotation can contribute to image management tasks such as indexing and sharing of images and regions of interest by providing a common semantic reference to align and query the heterogeneous data available [1].

Current approaches to bridge the gap between the pixel-based foundational representation and high-level image semantics, such as content-based image retrieval (CBIR), include the utilization of taxonomies describing 2D spatial relations between the imaged objects and hence linking image features with semantics [3]. Therefore, indexing of images is implemented by combining low-level features (intensity, texture, color, shape, size, etc.) with features of high-level image semantics, such as our understanding of real-world objects. For example, an ontology of objects (Fig. 1) can be used to map low-level features with high-level features of image [1].

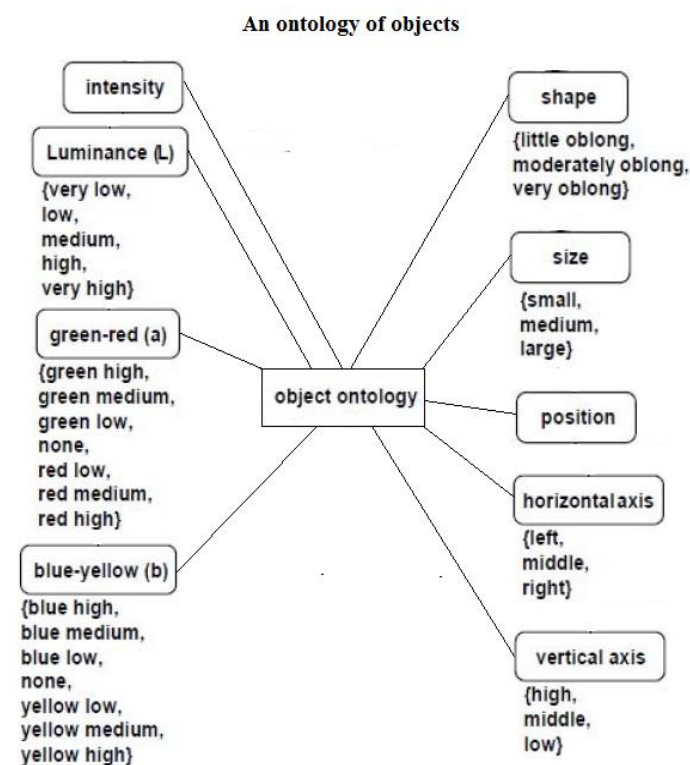


Fig. 1. Ontology-based image representation.

Ontologies must be clear both to humans for semantic annotation and to machines for cogitative and error checking, unintuitive rules for image processing give errors. Common rules allow for alignment ontology with other ontologies and, therefore, enable cross-domain operation with data. A reference ontology or domain ontology tries to optimize representational suitability to its subject matter. An application ontology is constructed for specific practical objectives.

Relational entity is anything that exists, including things, processes, functions, qualities, beliefs, actions, images, etc. The basic idea of relational entities is representation: for a person or interpretant an image entity represents some type of reality outside the image – an object, therefore, they connect viewers to reality. Instances represent what particular exists in reality – exists in numerous instances, e. g. databases, inventories, images. Types are connected to their instances and exist in objective reality – types of image, types of imaging process, etc. Types are ontologies, terminologies, catalogs, etc. Images are continuous representations – they represent instances in reality, but ontologies represent types in reality and the relations between these types.

Top-level ontologies proposed for the formalization of low level features include the VDO ontology, the COMM ontology, FMA, OBO ontologies, FuGO, SNOMED, UMLS Semantic Network, NCI Thesaurus, ICF, ISO Terminology Standards, HL7-RIM, BFO, DOLCE and other ontologies. Top-level ontologies have been shown to work in many different domains. Top-level ontology structure is shown in Fig. 2.

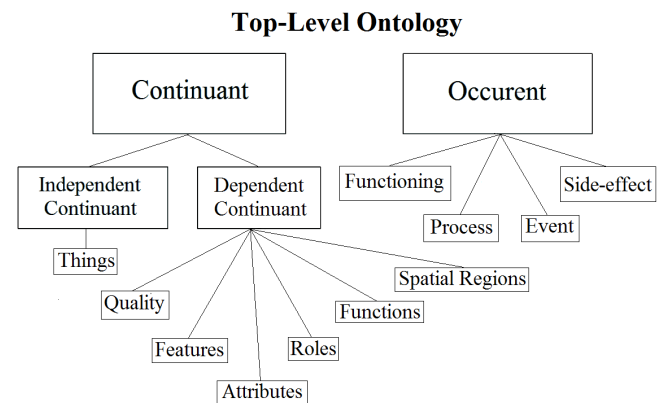


Fig. 2. Top-level ontology.

The ontologies are divided into two versions: continuant ontologies, and occurrent ontologies. Continuants are something existing in time and they preserve their identity through a change. Occurents are temporal parts, they unfold themselves in successive phases and they exist only in their phases [12]. Subtypes of continuants include independent continuants, dependent continuants, spatial regions and environments. All occurents are dependent entities – they are dependent on continuants as their carriers. Some dependent entities are monadic – they have one single independent carrier, but other dependent entities are relational – they have more than one independent carrier. Some dependent continuants are also realizable.

III. ONTOLOGY-BASED IMAGE REPRESENTATION

Images are complex representations, but they are not made of representational units. Representational units are terms, icons, bar codes, alphanumeric identifiers etc., which relate to entities in reality and which are not made of additional sub-

representations; therefore, representational units are the smaller parts in the domain of representations. The sub-representations of an image are regions. For an image, annotation terms representing types to detect knowledge about specific instances are used:

- This image as a whole;
- Regions in the image;
- Qualities of these regions;
- That part of the world the image represents as representative of a type as this specific instance.

Image interpretation is the mapping of regions (image segments) to pairs (instance, type) or the image segmentation to instances in reality identified as instances of specific types. Actual entities present in the image are regional segments created by delineation. These entities are the so-called image patterns. In addition, other kinds of segmentations exist, such as segmentations according to quantitative and qualitative spatial coordinates [4].

The image class definition can be done either by means of sample images or through a linguistic description. A priori information is represented by sample image parts. Two types of sample image parts are taken into account: clusters (blobs) that represent a region in the image, which delineates one object of interest or an image area and patches that separate one characteristic part of one object of interest. Therefore, an object of interest is described by a series of patches with their spatial relations. A priori information can also be represented by a linguistic description. The description language is generally made over a domain ontology. Therefore, the description of an image class is an application ontology [2].

The semantic representation of remote sensing images is characterized by the semantic definition of geographical objects presented in the image and respective relations [6]. The spatiotemporal relationships are used to obtain space and time related information of an image. The image contents are analyzed, and the relationship between the content (image objects) is obtained by means of spatial relationship. If the spatial and the temporal dimensions are included into the part of ontology, the image representation level is raised to the new dimension. The spatial information describes the regions of space. The digital image consists of various objects (or pixels), and these objects are linked by using spatial relationships.

The Region Connection Calculus (RCC) is used for representing high-level (qualitative) information in spatial representation. The Region Connection Calculus takes regions rather than points as a fundamental notion. The RCC abstractly describes regions in Euclidean space (or in a topological space) by their possible relations to each other. The RCC includes eight basic relations that can exist between any two regions. These basic relations are as follows:

- X is disconnected from Y (DC);
- X is externally connected with Y (EC);
- X is equal to Y (EQ);
- X partially overlaps Y (PO);
- X is tangential proper part of Y (TPP);
- X is tangential proper part of Y inverse (TPPi);

- X is non-tangential proper part of Y (NTPP);
- X is non-tangential proper part of Y inverse (NTPPi).

DC (disconnected) means that the two regions do not have a common point. EC (externally connected) means that the two regions have only common point borders. PO (partially overlapping) means that the two regions have common interior points. But TPP (tangential proper part) means that one region is a subset of the other region and they have some common points along the margins [5]. Fig. 3 shows eight basic relations of the Region Connection Calculus (RCC).

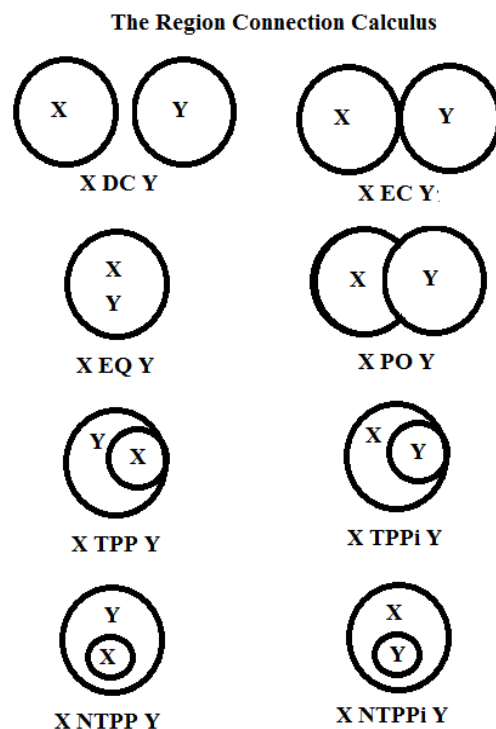


Fig. 3. The region connection calculus.

The association between geographic objects is represented in the image; knowledge takes into account the recovery process, and the topological relationships between the objects are represented [6]. The topological relationship is also used to tender quantitative information in spatial representation. The relevant topological relationships are as follows:

- Proximity describes how close two or more objects are;
- Orientation describes the location and direction of an object;
- Connectivity means how two objects are linked with each other;
- Adjacency explains whether two objects are next to each other or not;
- Membership means whether an object belongs to a particular group or not.

Temporal relationships are described by instances of relations, whose validity is a function of time [5]. Fig. 4 shows a hierarchy of temporal relationships using an event.

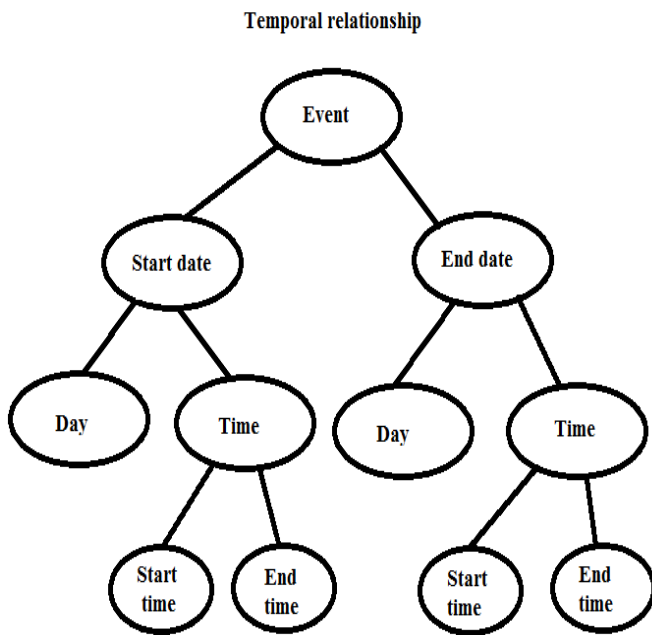


Fig. 4. Temporal relationship.

IV. THE IMAGE REPRESENTATION ONTOLOGY (IRON)

In paper [3], a novel ontological approach that formalizes concepts and relations with respect to image representations for image mining – the Image Representations Ontology (IRON) – was presented. It provides descriptors for pixels, image regions, image features, and clusters. This approach continues previous ontology approaches by including application of spatial relations between clusters in multidimensional feature spaces. The subject matter of images, image-related entities and vector space-related entities were used.

Images can be represented by the following image related entities:

- Whole images;
- Pixel, as the smaller part of an image;
- Image regions of interest (ROIs) that include self-connected regions of pixels, where these regions can have arbitrary shapes;
- Image features as properties of ROIs, images or image collections such as time-series images. Features include descriptors of intensity (histograms, mean, standard deviation, skewness, kurtosis), descriptors of texture (energy, entropy, contrast, homogeneity, correlation, variance, Laplacian transformation, gradient transformation, angular second moment, flat texture, inverse difference moment, etc.), descriptors of shape (curvature, area, perimeter, etc.) [11], position, and size;

- Image Attributes (DICOM attributes, gamma values, background color, textual metadata information), and their values as describing the whole image together with the process of the capture of the image.

The following vector space-related entities were included:

- Vector spaces with any dimensionality;
- Vectors that can be created by a set of image features, such as histograms;
- Clusters of feature vectors extracted from an image region.

Two kinds of relations: Representational Relation Assertions (RRAs) and Spatial Relation Assertions (SRAs) were used. The both of them were not modeled as object properties but as classes that are structure that is more complex. RRAs relate a given image, an image region, an image feature, or a set of features to the object it represents. SRAs relate clusters to each other. Such relations can also include vectors because a single vector can be viewed as a cluster of unitary number of elements. With the proposed ontology, clusters can be defined in n -dimensional feature spaces, where $n > 0$. Directional relations between the clusters indicate when one cluster is on the right or on the left of another cluster across axis m . Such relations are marked as “Right of across m axis” and “Left of across m axis”, where $m = 1, \dots, n$. There are also two-valued and three-valued SRAs (for example, “A is between B and C across axis m ”). SRAs include also distance relations, such as “Close to” and “Far from”, and topological relations, such as “Intersects with”, “Is interior to” and “Is exterior to”. The SRAs can be improved by fuzzy logic so that the ambiguity of the real world relations can also be captured.

Concrete domains like numeric values cannot be expressed in Web Ontology Language with DL class (OWL-DL). Many interpretation classes and features require the reference to numeric values. This can be represented by OWL-DL using XML schema. Comparing approach in [3] with published image ontologies it has been claimed that it is the one that most strictly implements principles of formal ontology. IRON was mainly coined based on the idea of using the formalization of 2D spatial relations to prevent uncertainty in image interpretation. However, the 2D spatial relations between the imaged objects can be defined by visual observations, while the spatial relations in a multi-dimensional space cannot be visually observed. In IRON this information can be provided by the “ground truth” information extracted from images annotated by domain experts. The “ground truthing” refers to the process of gathering the proper objective data. IRON can also be used to describe image content with 2D spatial relations similarly to the current imaging ontologies; for example, by using the pixel coordinates as features. Therefore, IRON is in general more used than current imaging ontologies.

V. CONCLUSIONS

In this paper, digital image semantic annotation by ontology and a novel ontological approach that formalizes concepts and relations with respect to image representations for data

mining – the Image Representations Ontology (IROn) – were examined.

An image is typically the representation of the objects present in a real life. The interpretation of an image by the computer is a highly complex task. To be able to perform the computational processing of images, it is necessary first to convert them into digital format. More challenging research perspectives arise when ontologies are used to annotate images where the information is encoded in numeric pixel values. An ontology is a specification of conceptualization to create a vocabulary for exchanging information, where conceptualization means mapping between symbols used in the computer (i. e., the vocabulary) and objects and relations in the real world.

Therefore, ontologies can be viewed as a means for semantic image annotation. Ontology-based semantic image annotation can contribute to image management tasks such as indexing and sharing of images and regions of interest by providing a common semantic reference to align and query the heterogeneous data available. Current approaches to bridge the gap between the pixel-based foundational representation and high level image semantics such as CBIR include the utilization of taxonomies describing 2D spatial relations between the imaged objects and, hence, linking image features with semantics and, therefore, indexing images by combining low-level features (intensity, texture, color, shape, size, etc.) with features of high level image semantics, such as our understanding of real-world objects.

The basic idea of relational entities is representation: for a person or interpretant an image entity represents some type of reality outside the image – an object, therefore they connect viewers to reality. Instances represent what general exists in reality – exists in numerous instances, e. g. databases, inventories, images. Types are connected to their instances and exist in objective reality – types of image, types of imaging process, etc. Types are ontologies, terminologies, catalogs, etc. Images represent instances in reality, but ontologies represent types in reality and the relations between these types.

A novel ontological approach that formalizes concepts and relations with respect to image representations for image mining – the Image Representations Ontology – was examined. It provides descriptors for pixels, image regions, image features, and clusters. This approach extends previous ontology approaches by including application of spatial relations between clusters in multidimensional feature spaces: representational relation assertions and spatial relation assertions. In addition, the subject matter of images, image-related entities and vector space-related entities were used in this approach.

Comparing the Image Representations Ontology with published image ontologies it has been claimed that it is the one that most strictly implements principles of formal ontology.

REFERENCES

[1] R. A. Moreno, M. d. S. Rebelo and M. A. Gutierrez, "Representation and Indexing of Medical Images," *Anales de Documentación*, vol. 14, no. 2, pp. 1–18, 2011. [Online]. Available: Dialnet, <http://dialnet.unirioja.es>. [Accessed Aug. 1, 2014].

- [2] R. Clouard, A. Renouf and M. Revenu, "An Ontology-Based Model for Representing Image Processing Application Objectives," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 24, no. 8, pp. 1181–1208, Dec. 2010. [Online]. Available: World Scientific, <http://www.worldscientific.com/>. [Accessed August 5, 2014]. <http://dx.doi.org/10.1142/S0218001410008354>
- [3] D. K. Iakovidis, D. Schober, M. Boeker and S. Schulz, "An Ontology of Image Representations for Medical Image Mining," *9th International Conference: Information Technology and Applications in Biomedicine*, pp. 1–4, Nov. 4–7, 2009. [Online]. Available: DebugIT, <http://debugit.eu/>. [Accessed August 10, 2014].
- [4] B. Smith, "Image Ontology." [Online]. Available: The National Center for Biomedical Ontology, <http://www.bioontology.org/>. [Accessed Aug. 21, 2014].
- [5] N. Magesh and P. Thangaraj, "Image Ontology Construction using Spatial and Temporal Relationships," *Life Science Journal*, vol. 10, no. 3, pp. 1–9, 2013. [Online]. Available: Life Sciences, <http://www.lifesciencesite.com/>. [Accessed Sep. 27, 2014].
- [6] R. O. Santos, "Ontology-Based Topological Representation of Remote Sensing Images," *International Journal of Remote Sensing*, vol. 35, no. 1, pp. 16–28, 2014. [Online]. Available: Taylor & Francis Online, <http://www.tandfonline.com/>. [Accessed Sep. 28, 2014]. <http://dx.doi.org/10.1080/01431161.2013.858847>
- [7] D. Lu, M. Batistella and E. Moran, "Integration of Landsat TM and SPOT HRG Images for Vegetation Change Detection in the Brazilian Amazon," *Photogrammetric Engineering & Remote Sensing*, vol. 74, no. 4, pp. 421–430, Apr. 2008. [Online]. Available: ASPRS, <http://www.asprs.org/>. [Accessed Aug. 3, 2014]. <http://dx.doi.org/10.14358/PERS.74.4.421>
- [8] F. T. Fonseca, M. J. Egenhofer and P. Agouris, "Using Ontologies for Integrated Geographic Information Systems," *Transactions in GIS*, vol. 6, no. 3, pp. 231–257, June 2002. [Online]. Available: Wiley Online Library, <http://onlinelibrary.wiley.com>. [Accessed Aug. 7, 2014]. <http://dx.doi.org/10.1111/1467-9671.00109>
- [9] D. Li, K. Di and D. Li, "Land Use Classification of Remote Sensing Image with GIS Data Based on Spatial Data Mining Techniques," *XIXth ISPRS Congress Technical Commission III: Systems for Data Processing, Analysis and Representation*, pp. 238–245, Amsterdam, The Netherlands, July 16–23, 2000. [Online]. Available: ISPRS, <http://www.isprs.org/>. [Accessed Aug. 8, 2014].
- [10] D. Kohli, R. Sliuzas, N. Kerle and A. Stein, "An Ontology of Slums for Image-Based Classification," *Computers, Environment and Urban Systems*, vol. 36, no. 2, pp. 154–163, Mar. 2012. [Online]. Available: ResearchGate, <http://www.researchgate.net>. [Accessed Aug. 12, 2014].
- [11] S. Mhiri, S. Despres and E. Zagrouba, "Ontologies for the Semantic-Based Medical Image Indexing: An Overview," *Proceedings of the 2008 International Conference on Information & Knowledge Engineering*, Las Vegas, Nevada, USA, July 14–17, 2008. [Online]. Available: ResearchGate, <http://www.researchgate.net>. [Accessed Sep. 15, 2014].
- [12] M. Tsiknakis, M. Brochhausen, J. Nabrzysky, J. Pucacki, S. G. Sfaikanakis, G. Potamias, C. Desmedt and D. Kafetzopoulos, "A Semantic Grid Infrastructure Enabling Integrated Access and Analysis of Multilevel Biomedical Data in Support of Postgenomic Clinical Trials on Cancer," *IEEE Transactions on Information Technology in Biomedicine*, vol. 12, no. 2, pp. 205–217, Mar. 2008. [Online]. Available: IMBB, <http://www.imbb.forth.gr>. [Accessed Sep. 20, 2014]. <http://dx.doi.org/10.1109/TITB.2007.903519>

Arthur Stepchenko is the first-year Doctoral Student majoring in Information Technology at Riga Technical University (RTU). He received his Mg.sc.comp. degree in Computer Science from Ventspils University College (VUC) in 2013.

He is a Research Assistant at the Department of Space Technology of the Engineering Research Institute "Ventspils International Radio Astronomy Centre", Ventspils University College (VIRAC).

His research interests include remote sensing, digital image processing, machine learning, data mining and programming.

E-mail: arturs1312@gmail.com.

Arkady Borisov received his Doctoral Degree in Technical Cybernetics from Riga Polytechnic Institute in 1970 and Dr.habil.sc.comp. degree in Technical Cybernetics from Taganrog State Radio Engineering University in 1986.

He is a Professor of Computer Science at the Faculty of Computer Science and Information Technology, Riga Technical University. His research interests include fuzzy sets, fuzzy logic, computational intelligence and bioinformatics. He has 235 publications in the field.

He is a member of IFSA European Fuzzy System Working Group, Russian Fuzzy System and Soft Computing Association, honorary member of the Scientific Board, member of the Scientific Advisory Board of the Fuzzy Initiative Nordrhein-Westfalen (Dortmund, Germany).
E-mail: arkadijs.borisovs@cs.rtu.lv.

Artūrs Stepčenko, Arkādijs Borisovs. Uz ontoloģiju bāzēta attēla reprezentācija

Attēls parasti reprezentē reālās pasaules objektus. Attēla interpretācija ar datora palīdzību ir sarežģīts uzdevums, un, lai, būtu iespējams veikt skaitļošanas darbības ar attēlu, vispirms tas ir jāpārveido digitālā formātā. Digitālais attēls ir divdimensionāla matrica, kuru veido diskrētas vērtības, ko sauc par pikseliem. Ontoloģija ir konceptualizācijas specifika, lai izveidotu vārdnīcu informācijas apmaiņai, kur ar konceptualizāciju tiek saprasta reālās pasaules objektu un to attiecību attēlošana ar datorā izmantojamo simbolu (vārdnīcas) palīdzību. Tādējādi ontoloģiju var apskatīt kā līdzekli attēla semantiskai anotācijai. Pašlaik, lai mazinātu plaisu starp uz pikseliem bāzētu attēla reprezentāciju un augsta līmeņa attēla semantiku, izmanto taksonomiju, kas apraksta divdimensionālas telpiskās attiecības starp attēla objektiem, un tādējādi sasaista attēla īpašības ar semantiku. Attēlus var indeksēt, kombinējot attēla zema līmeņa īpašības (intensitāte, tekstūra, krāsa, forma, izmērs) ar attēla augsta līmeņa semantikas īpašībām, kā, piemēram, reālās pasaules objektu izpratne. Telpiskās un īslaicīgās attiecības raksturo attiecības starp attēla objektiem. Telpiskā informācija apraksta telpā esošos reģionus. Īslaicīgas attiecības tiek aprakstītas ar instanču attiecībām, kuru darbības ilgums ir funkcija, kas atkarīga no laika. Rakstā tika apskatīta jauna ontoloģiskā pieeja – Attēla reprezentācijas ontoloģija (ARO). Tā nodrošina aprakstītājus pikseliem, attēlu reģioniem, attēlu īpašībām un klasteriem. Šī pieeja paplašina iepriekšējās ontoloģiskās pieejas, iekļaujot telpisko attiecību pielietošanu starp klasteriem daudzdimensionālās attēlu īpašību telpās. Arī attēlu temats, ar attēlu saistītās vienības un ar vektoru telpu saistītās vienības tiek izmantotas šajā pieejā. Salīdzinot Attēla reprezentācijas ontoloģiju ar iepriekš publicētajām attēlu ontoloģijām, tika secināts, ka šī ontoloģija ļoti strikti īsteno formālas ontoloģijas principus.

Артур Степченко, Аркадий Борисов. Представление изображения на основе онтологии

Изображение обычно представляет объекты реального мира. Интерпретация изображения с помощью компьютера - сложная задача, и чтобы можно было совершать вычислительные операции с изображением, во-первых, изображение надо представить в цифровом формате. Цифровое изображение является двумерной матрицей, которая состоит из дискретных значений, называемых пикселями. Онтология является спецификацией концептуализации, цель которой - сделать словарь для обмена информацией, где под концептуализацией понимают отображение объектов реального мира и их отношений с помощью символов, используемых в компьютере (словаре). Поэтому онтология может рассматриваться как средство для семантической аннотации. Изображения могут быть индексированы, комбинируя свойства изображения низкого уровня (интенсивность, текстура, краска, форма, размер) со свойствами семантики изображения высокого уровня, как например понимание объектов реального мира. Пространственные и кратковременные отношения описывают отношения между объектами изображения. Пространственная информация описывает регионы, которые находятся в пространстве. Кратковременные отношения описываются с помощью отношений инстанций, продолжительность которых является функцией, зависящей от времени. В статье был рассмотрен новый онтологический подход, который формализует понятие и отношение для представления изображений – Онтология Представления Изображения (ОПИ). Она обеспечивает средства описания для пикселей, регионов изображений, свойств изображений и кластеров. Этот подход дополняет существующие онтологические подходы, включая применение пространственных отношений между кластерами в многомерных пространствах свойств изображений. В данном подходе также используются темы изображения, связанные с изображением единицы и пространством векторов единицы. Сравнивая Онтологию Представления Изображения с ранее опубликованными онтологиями изображений, был сделан вывод, что эта онтология очень строго соблюдает принципы формальной онтологии.

Fuzzy Classification System for Bioinformatics Data Analysis

Madara Gasparovica-Asite¹, Ludmila Aleksejeva², ^{1,2} Riga Technical University

Abstract – This article describes the fuzzy classification system developed by the authors and that is particularly applicable to bioinformatics data classification. The description focuses on the following steps in the system: 1) Data preprocessing; 2) Classifier training and construction of the rule base; 3) Classification of new records and 4) Evaluation of the results; it also explains the details of processes in each step as well as the processes of missing data replacement, reduction of the number of alternatives and functions, construction of membership functions and stretching of the induced rules. The article concludes with a justification of the methods and algorithms chosen for each process of the system.

Keywords – Classification system, data mining, data preprocessing, fuzzy logic.

I. INTRODUCTION

A classification algorithm defines the association of a new object with a predefined group of objects. The classification system proposed in this article embraces the complete process of data mining – from preparing the data to result evaluation because common classification algorithms only cover the classification process although the methods for data preprocessing and result evaluation are critical and very necessary.

The classification system described in this article is intended for bioinformatics data classification. The specific features of the bioinformatics data comparing to other types of data are their large dimensionality (containing from several hundred to few thousand attributes) and the atypical comparatively small number of records (commonly these data sets hold less than hundred records due to the expensive tests that are used to collect the data). The data usually hold the results of patient examination and haematological molecular tests using very expensive media that explains the large dimensionality and small number of the available records. The data often has missing values due to tests not having been carried out for specific patients. Therefore, the methods chosen for this system have to comply with the specific character of the bioinformatics data sets; this does not mean that this classification system may not be used with other types of data although the best results are expected in data sets with similar record/dimension proportions.

Section II describes the fuzzy classification system and gives the information about the parts of the above-mentioned system. Section III gives the justification for inclusion of the data preprocessing part, including 1) replacing the missing values, 2) reduction of the dimensionality, 3) construction of membership functions. Section IV gives the description of the base fuzzy classification algorithm and its extension – rule stretching. Section V provides the explanation of result

evaluation. The article is concluded with the explanation of the reasoning behind the chosen system structure and its processes as well as with insight and ideas for further research.

II. FUZZY CLASSIFICATION SYSTEM

The fuzzy classification system developed by the authors consists of four main parts (a full diagram of the proposed system is presented in Fig. 1):

- Data preprocessing (if necessary, replacement of missing values; attribute subset reduction; construction of membership functions).
- Classifier training and rule base creation using rule stretching.
- Classification of new records (classifier validation or testing phase);
- Result evaluation (overall accuracy of the classifier and its overall error).

If it is possible, an expert of the problem domain can be involved in the result evaluation phase. This can be helpful while evaluating significance of the obtained results for the corresponding domain.

III. DATA PREPROCESSING

Data preprocessing is a significant part of data mining because the quality of the prepared data often has an indicative influence on the classification results [1]. For example, if the missing values are not processed, some classification algorithms cannot be applied for the reason that they cannot classify data with missing values. If dimensionality reduction is not applied, the time necessary for calculations can be excessively long, which prolongs the whole process necessary to receive results, which can be unfavorable in some domains. Construction of membership functions determines the result of classification, whereas this process turns data values into fuzzy values.

A. Missing Value Replacement

The missing values of the initial data sets are processed using the missing data replacement methods. The techniques that can be applied while processing the missing data can be divided into three groups [2]:

- 1) Ignoring the missing data or deleting the records/attributes with the missing values from the data set;
- 2) Attribute, criterion evaluation – specific algorithms that can be used to evaluate the significance of the missing data;
- 3) Replacing the missing values.

If the missing values are to be replaced, the corresponding algorithms can be divided into these groups according to the chosen value calculation approach:

- Using a global constant – all of the missing values are replaced by a certain symbol or word;
- Using the mean value of an attribute – the missing values are replaced by the mean value of the corresponding attribute;

- Using the class mean value of an attribute – the missing values are replaced by the mean value of the corresponding attribute for the corresponding class (an improved version of using the mean value of an attribute).

Using the most probable value – the missing values are replaced using one of the data mining methods – regression, cluster analysis etc., for more information about various methods see [3].

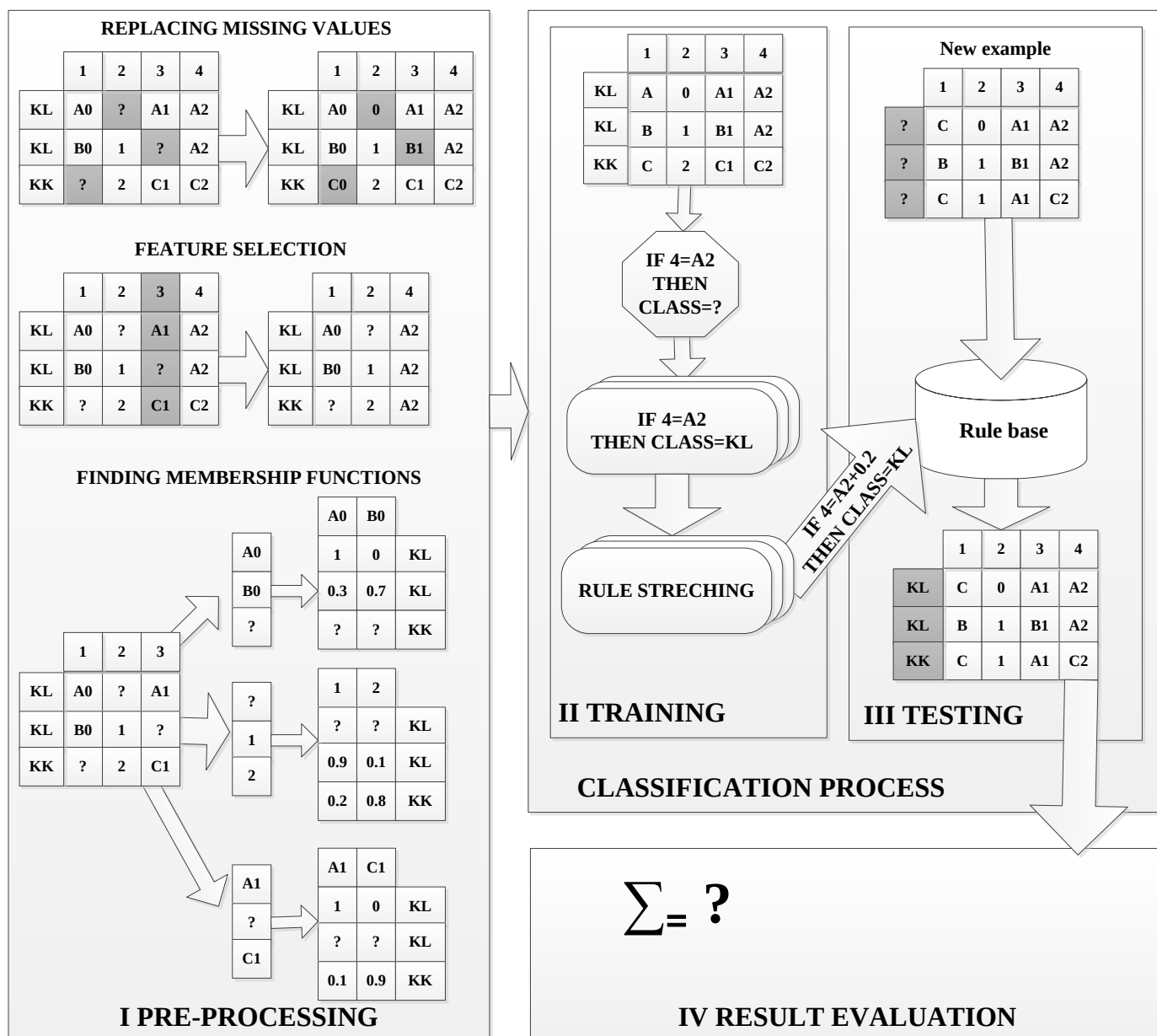


Fig. 1. Main processes of the classification system.

Taking into account the scope of data, one can conclude that in bioinformatics data imputation should be first considered from the biology point of view. In this case, the type of biological research (in vivo or in vitro) has to be taken into account. In the case of in vivo study (both preclinical and clinical), there is always a great variability in the obtained results because of the individual happening of physiological

processes of the individuals included into the research. Consequently, the value group of results obtained in the research can differ. To apply missing data imputation methods in in vivo studies a compulsory prerequisite is wide population studies that often cannot be carried out from clinical and biological point of view. In the case of in vitro studies, the research is performed in highly controlled laboratory

conditions, most often using standardized cell, microorganism or virus cultures. The precision and homogeneity of a group in such a research design is mostly affected by experience and skills of the experimenter. Respectively, in an optimally performed research one can expect foreseeable and homogenous results in the groups. In this case, missing value imputation methods in bioinformatics are applicable [3].

Therefore, the missing data replacement methods can be successfully used from the viewpoint of data mining; nevertheless, all risks must be considered beforehand by examining the possible changes that can be introduced to data. In the case when the initial data (with missing values) can be successfully processed without missing data replacement, from the viewpoint of biology it is advisable to leave the missing data. Therefore, several methods like Support Vector Machine Imputation (SVMi) [4] or Weighted K-Nearest Neighbor Imputation (WKNNi) [4] can be used if necessary, but the use should be considered for each bioinformatics data set individually [3].

B. Feature Selection

The bioinformatics data commonly have a large number of attributes and a small number of records. Therefore, processing of these data using classification algorithms is a very time- and resource-consuming process, because in order to induce rules the algorithm has to be trained on a very large and complex data set. Therefore, data pre-processing includes attribute selection techniques to reduce the number of attributes. These methods select only the attributes that are significant for data classification.

There are many advantages of feature selection and they allow carrying out the following activities: to avoid overfitting and improve model performance; to provide faster and more cost-effective models and to gain a deeper insight into the underlying processes that generated the data [4]. Feature selection techniques differ from each other in the way they incorporate search in the added space of feature subsets in the model selection. In the context of classification, the feature selection techniques can be conditionally divided into three categories according to the extent they can be linked to a classification model: 1) filter methods, 2) wrapper methods and 3) embedded methods [4]. Filter techniques assess the relevance of features by looking only at the intrinsic properties of the data. Wrapper methods embed the model hypothesis search within the feature subset search. In embedded technologies, the search for an optimal subset of features is built into the classifier construction [4].

A study was carried out to determine the impact of feature selection methods on results of bioinformatics data classification, to see more details about the study, consult [7]. It analyzed four different feature selection methods: Fast Correlation Based Filter solution (FCBF), Linear Forward Selection (LFS), Greedy Stepwise (GS), Genetic Search using three classification algorithms (GSUTCA); and indicated that the results of the original data set classification were one of the worst, but the best results were achieved using attribute selection with Greedy Stepwise and FCBF Search; therefore, using FCBF is beneficial for the complex and multidimensional bioinformatics data. Therefore, the proposed classification system uses FCBF [8] feature selection method.

C. Membership Function Construction

The membership function construction step is very important for all algorithms that work with fuzzy data, because the strength and efficiency of the feature selection method for a particular problem defines the success of the classification.

The authors have previously analyzed four fuzzy membership function construction methods – construction method using expert evaluations, membership function construction method using pairwise comparison, linguistic definitions of membership function using statistical data and the parametrical method for membership function construction. After a comparative empirical analysis was concluded in the previous research, the most promising method for this research field was the membership function construction method, which used expert evaluation, and the membership function construction method that used pairwise comparison [9], but due to the complex and extensive calculations that were necessary to find n-grade polynomial roots further studies used a simple construction algorithm for triangular membership function construction [10] with different numbers of intervals [11]. Potentially good results could be achieved by the membership function construction method based on the fuzzy classification algorithm [12]. However, this approach still needs more research.

The fuzzy classification system uses the membership function construction method that is based on triangular membership functions.

IV. TRAINING AND TESTING

During the process of training and testing, the classifier is trained and the example base is created, then the classifier is tested to evaluate the obtained classification results. The algorithm for the classification step used in the evaluation process is FuzzyBEXA, which is structurally based on the classic classification algorithm BEXA [13]. FuzzyBEXA algorithm expands the use of conditions (definitions) described in BEXA to fuzzy data. In the case of the fuzzy data classification algorithm FuzzyBEXA, the value of the algorithm is a conjunction in the range from 0 to 1, and a record can correspond to a conjunction with a very small membership indicator.

This type of situation can be undesirable; therefore, the algorithm introduces new variables: “alpha-cut” and “alpha-class cut” [13], more details about this classification algorithm are given in [15]. The induced classification rules are in the “IF ... THEN ...” form. A scheme of the algorithm is given in Fig. 2.

The analysis of similar studies and available literature pointed out FURIA as another very interesting algorithm for this task. FURIA (from *Fuzzy Unordered Rule Induction Algorithm*) learns fuzzy rules and unordered rule sets. The algorithm induces rules for each class separately using the “one class – other classes” dividing strategy. The main strength of this algorithm is the rule stretching method solving the pressing problem that new records, which should be classified, could be outside the space covered by the previously induced rules [16]. The rule stretching strategy

proposed by the authors is also used in this proposed fuzzy classification algorithm by stretching rules and expanding the range of application for these rules.

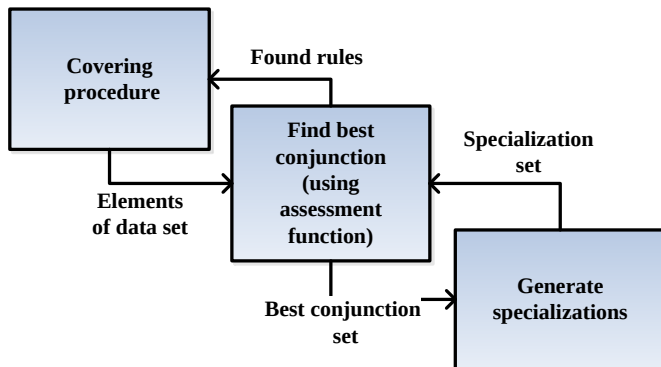


Fig. 2. FuzzyBEXA algorithm scheme.

The 10-fold cross-validation was used to correctly evaluate classification results. Therefore, the training and testing parts cannot be separated and will be described in this section together [12].

V. RESULT EVALUATION

Fig. 3 shows the description of the algorithms and methods included in each component of the system. The figures are followed by a detailed description of the reasoning behind inclusion of each component in the proposed fuzzy classification system and its significance in data mining.

The classification result evaluation can be carried out using five criteria [17]:

- Accuracy – the accuracy of the classification algorithm;
- Performance – how much of the calculation resources are necessary for the classifier to carry out classification;
- Robustness – the ability of the classifier to be error-resistant;
- Scalability – the ability of the classifier to maintain the degree of generalization while working with large amounts of data;
- Interpretation – the ability of the classifier to explain the obtained results.

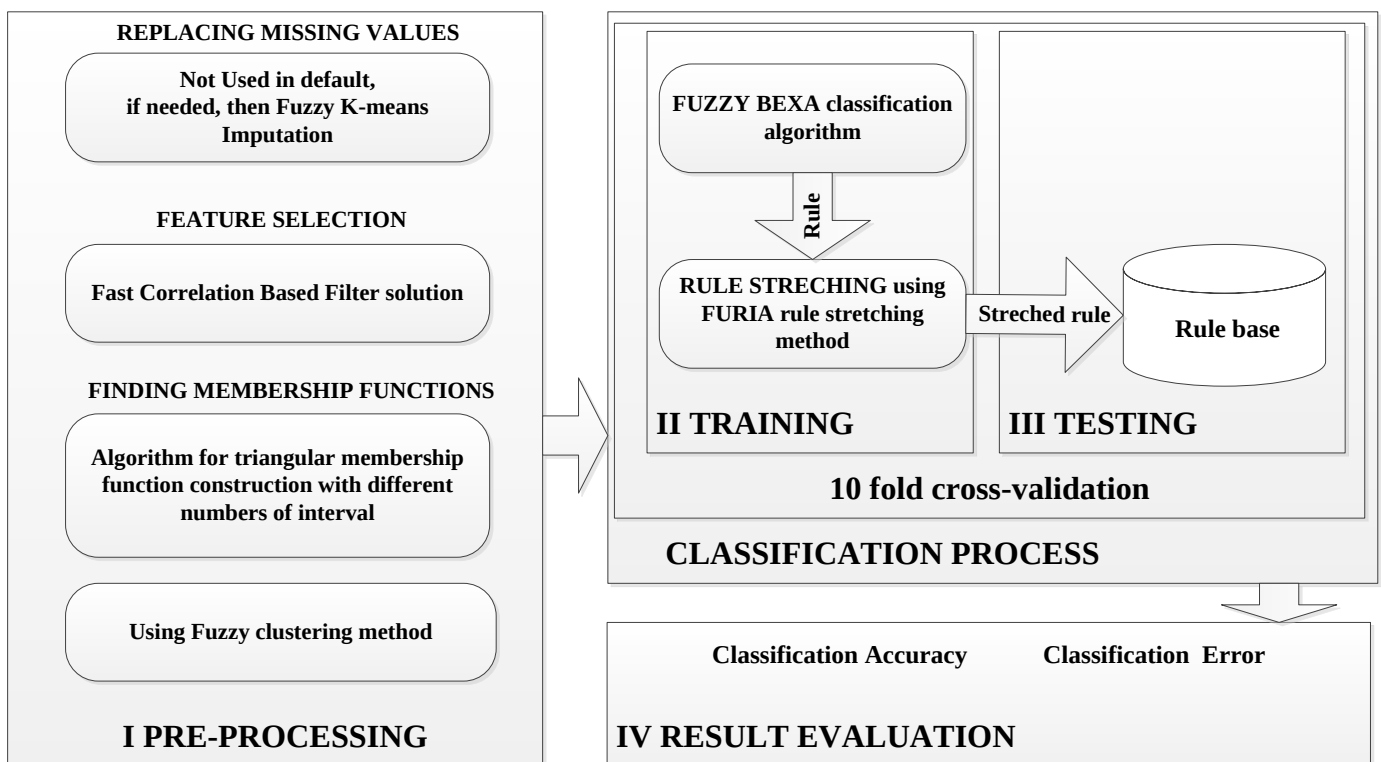


Fig. 3. Classification system algorithms.

As for the first – accuracy, it is the accuracy of the obtained IF-THEN classification rules to assign the correct class to a new record that has not been included into a training set. The accuracy is measured using the confusion matrix, that is, a matrix where the number of classes exists. The columns represent the class labels allocated by a classifier and the rows correspond to the actual classes. The confusion matrix for a two-class problem is shown in Table 1. The matrix is filled

with classification results – either correct or misclassified according to classes. The true positive rate (TP) shows the number of the positive records that were classified as positive; the false negative rate (FN) shows the number of positive records that were misclassified (assigned the negative label); the false positive rate (FP) shows the number of negative records that were misclassified (assigned the positive label); the true negative (TN) rate shows the number of negative

records that were classified correctly (assigned the negative label). To assess the accuracy of the obtained classifier, the overall accuracy (AC) is used (see TABLE I) [18]. It can be calculated by dividing the number of correctly classified records by the number of records in the used test data set. Similarly, the overall classification error (ER) can be calculated as presented in TABLE I.

TABLE I
CONFUSION MATRIX

		Predicted class	
		Yes	No
Actual class	Yes	$f_{++}(TP)$	$f_{+-}(FN)$
	No	$f_{-+}(FP)$	$f_{--}(TN)$
Overall accuracy		$AC = \frac{TP + TN}{TP + FN + FP + TN}$	
Overall classification error		$ER = 1 - AC$	

As for the performance, robustness and scalability – these criteria for the proposed system will be assessed in empirical studies.

As for the interpretation – IF-THEN rules provide an easily perceivable link between input data and the resulting class, which confirms the successful interpretability for the models of this algorithm.

VI. CONCLUSION

The fuzzy classification system described in the article is intended for classification of bioinformatics data. The article gives a detailed description and reasoning behind inclusion of every part of the system: 1) data preprocessing, 2) classifier training and creation of the rule base, 3) classification of new records (classifier validation or testing phase) and 4) result evaluation. All algorithms used in the classification system processes were chosen according to the experimental analysis described in the following articles: [3], [7], [9], [11]–[13], [15]. The main conclusions and results are the following:

- The methods chosen for data preprocessing:
 - a. Missing data replacement (only where necessary) should be carried out using Support Vector Machine Imputation or Weighted K-Nearest Neighbor Imputation methods;
 - b. Feature subset reduction – Fast Correlation Based Filter solution;
 - c. Membership function construction – a simple algorithm for triangular membership function construction.
- Classification algorithm FUZZY BEXA was used for classification (rule stretching – FURIA rule stretching method).
- Result evaluation (calculation of overall classifier accuracy and overall classifier error).

The results of experiments proved the justified behind the choice of every technique for the appropriate process in the classification system. Although the studies conducted by the authors in the field of using fuzzy clustering algorithms in

membership function analysis show potential, this field requires extension to include it in this fuzzy classification system to allow users to choose between membership function construction methods, including the fuzzy clustering algorithm, which should be specifically adapted for using in membership function construction. Another field that requires empirical studies is proving the potential of the fuzzy classification system when compared to other data mining methods that are used for classification of bioinformatics data.

REFERENCES

- [1] N. Zhang and W. F. Lu, "An Efficient Data Preprocessing Method for Mining Customer Survey Data," *Industrial Informatics, 5th IEEE International Conference*, June 23–27, 2007, pp. 573–578 Vienna, Austria.
- [2] G.E.A.P.A. Batista and M.C. Monard, "An Analysis of Four Missing Data Treatment Methods for Supervised Learning," *Applied Artificial Intelligence*, vol. 17 no. 5, pp. 519–533, 2003. <http://dx.doi.org/10.1080/713827181>
- [3] M. Gasparoviča, G. Krievina, L. Aleksejeva, "Biological Interpretation of Metabolic Syndrome Data Missing Value Imputation and Classification," *Proceedings of Workshop on Data Mining in Life Sciences, DMLS'2012*, July 20, 2012, pp. 167–176, Germany, Berlin.
- [4] H. A. B. Feng, G. C. Chen, C. D. Yin, B. B. Yang, Y. E. Chen "A SVM Regression Based Approach to Filling in Missing Values," *9th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, KES2005. LNCS*, 2005, vol. 3683, pp. 581–587, Melbourne, Australia. Springer.
- [5] O. Troyanskaya, M. Cantor, G. Sherlock, et al. "Missing Value estimation Methods for DNA Microarrays," *Bioinformatics*, vol. 17, pp. 520–525, 2001. <http://dx.doi.org/10.1093/bioinformatics/17.6.520>
- [6] Y. Saeyns, I. Inza, P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007 <http://dx.doi.org/10.1093/bioinformatics/btm344>
- [7] M. Gasparoviča, L. Aleksejeva, "Feature Selection for Bioinformatics Data Sets – Is It Recommended?" *Proceedings of the 5th International Conference on Applied Information and Communication Technologies, AICT2012*, Apr. 26–27, 2012, pp. 325–335, Jelgava, Latvia.
- [8] L. Yu and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," *Proceedings of the Twentieth International Conference on Machine Learning, ICML-2003*, Aug. 21–24, 2003, pp. 856–863, Washington DC. AAAI Press: Menlo Park, California, 2003
- [9] M. Gasparoviča, L. Aleksejeva, I. Tuleiko, "Finding Membership Functions for Bioinformatics Data," *Proceedings of 17th International Conference on Soft Computing, MENDEL 2011*, June 15–17, 2011, pp. 133–140, Brno, Czech Republic.
- [10] R. Chutia, S. Mahanta and H. K. Baruah. "An Alternative Method of Finding the Membership of a Fuzzy Number," in *International Journal of Latest Trends in Computing*, vol. 1, no. 69, Issue 2, pp. 69–72, Dec. 2010.
- [11] M. Gasparoviča M., I. Tuleiko, L. Aleksejeva, "Influence of Membership Functions on Classification of Multi-Dimensional Data" *Proceedings of the Riga Technical University International Scientific Conference*, no. 49, vol. 5, pp. 78–84, 2011
- [12] M. Gasparoviča, L. Aleksejeva, V. Nazaruks, "Using Fuzzy Clustering with Bioinformatics Data," *Proceedings of the 6th International Conference on Applied Information and Communication Technologies, AICT2013*, Apr. 25–26, 2013, pp. 62–70. Jelgava, Latvia.
- [13] M. Gasparoviča, "Using Fuzzy Algorithms to Solve Classification Tasks," Master thesis, Riga technical University, Riga. Latvia, 2010 (in Latvian)
- [14] J. Zyl, "Fuzzy set covering as a new paradigm for the induction of fuzzy classification rules," PhD thesis, University of Mannheim, Germany, 2007.
- [15] M. Gasparoviča, L. Aleksejeva, V. Gersons, "The Use of BEXA Family Algorithms in Bioinformatics Data Classification," *Proceedings of the Riga Technical University International Scientific Conference*, vol. 5, no. 50, pp. 120–126, 2012.

- [16] J. Hühn, J., E. Hüllermeier, "FURIA: An Algorithm for Unordered Fuzzy Rule Induction," *Data Mining and Knowledge Discovery*, vol. 19, no. 3, pp.293–319, 2009. <http://dx.doi.org/10.1007/s10618-009-0131-8>
- [17] J. Han, M. Kamber, *Data Mining: Concepts and Technologies. 1st Edition*. San Francisco etc.: Morgan Kaufman, 2000.
- [18] P.N. Tan, M. Steinbach, V. Kumar, *Introduction to Data Mining*. Boston: Pearson Education, 2006.

Madara Gasparovica-Asite received her diploma of *Mg.sc.ing.* in Information Technology from Riga Technical University in 2010. Now she is a Doctoral Student at the study programme "Information Technology", Riga Technical University.

Since 2008 she has worked as a Senior Laboratory Assistant at Riga Technical University, and since 2010 she has been working as a Researcher at the Department of Modeling and Simulation, the Institute of Information Technology. Previous publications: Gasparovica M., Novoselova N., Aleksejeva L. *Using Fuzzy Logic to Solve Bioinformatics Tasks*, Proceedings of Riga Technical University. Issue 5, Computer Science. Information Technology and Management Science, Vol. 44, 2010, pp. 99–105; Gasparovica M., Aleksejeva L. *Using Fuzzy Unordered Rule Induction Algorithm for Cancer Data Classification*, Proceedings of the 17th International Conference on Soft Computing, MENDEL 2011, Czech Republic, Brno, June 15–17, 2011, pp. 141–147.

Her interests include decision support systems, data mining tasks and modular rules. She is a member of IEEE.

Address: 1 Kalku Str., LV-1010, Riga, Latvia.

E-mail: madara.gasparovica@rtu.lv.

Ludmila Aleksejeva received her *Dr. sc.ing.* degree from Riga Technical University in 1998. She is an Associate Professor at the Department of Modeling and Simulation, Riga Technical University. Her research interests include decision making techniques and decision support system design principles, as well as data mining methods and tasks, and especially collaboration and cooperation of the mentioned techniques.

Most important previous publications: Gasparovica M., Novoselova N., Aleksejeva L. *Using Fuzzy Logic to Solve Bioinformatics Tasks*, Proceedings of Riga Technical University. Issue 5, Computer Science. Information Technology and Management Science, Vol. 44, 2010, pp. 99–105; Gasparovica M., Aleksejeva L., Tuleiko I. *Finding Membership Functions for Bioinformatics Data* // Proceedings of the 17th International Conference on Soft Computing, MENDEL 2011, Czech, Brno, June 15–17, 2011, pp. 133–140; Aleksejeva L., Užga-Rebrovs O., Borisovs A. *Fuzzy Classification and Clustering*. Textbook. – Riga, RTU Press, 2012, 248 p.

Address: 1 Kalku Str., LV-1010, Riga, Latvia.

E-mail: ludmila.aleksejeva_1@rtu.lv.

Madara Gasparoviča-Asīte, Ludmila Aleksejeva. Izplūdušī klasifikācijas sistēma bioinformātikas datu analīzei

Šajā rakstā aprakstīta autoru izveidotā izplūdušī klasifikācijas sistēma, kas īpaši paredzēta bioinformātikas datu klasificēšanai. Dots bioinformātikas datu apraksts – liels atribūtu un neliels ierakstu skaits, kas liek arī atbilstoši klasifikācijas sistēmai rēķināties ar šīm bioinformātikas datu īpatnībām, kas izraisa atbilstošu datu pirmapstrādes metožu lietojumu. Rakstā dots sistēmas daļu apraksts (t.sk. shematiskais) un pamatojums, kāpēc katram no sistēmas moduļiem izvēlēts šāds algoritms vai metode. Datu pirmapstrādei izvēlētas tādas metodes: a) trūkstošo vērtību aizvietošanai (tikai, ja nepieciešams) izmantojamas atbalsta vektoru (SVM) vai svērtā k-tuvāko kaimiņu (WKNN) aizvietošanas metodes; b) atribūtu skaita samazināšanai – ātrais, uz korelāciju balstītais, filtra risinājums (FCBF); c) piederības funkciju konstruēšanai – vienkāršs trīstūrveida piederības funkciju konstruēšanas algoritms. Klasifikācijai un testēšanai izmantots FUZZY BEXA klasifikācijas algoritms, bet likumu stiepšanai – FURIA likumu stiepšanas metode, kas nodrošina papildus iegūto klasifikācijas likumu stiepšanu, tos paplašinot nodrošinot iespēju klasificēt jaunus ierakstus. Lai palielinātu precizitāti un nodrošinātu datu kopas objektīvu sadalījumu apmācības un testa datu kopās un izslēgtu to, ka iegūtais klasifikācijas rezultāts ir atkarīgs no nejaušības, tiek pielietota desmitkārtīgā šķēršvalidācija. Rezultātu novērtēšanai izmantots klasifikatora kopējās precizitātes un klasifikatora kopējās kļūdas aprēķins. Tā kā pēc autoru veiktajiem pētījumiem izplūdušo klasterizācijas algoritmu pielietojums piederības funkciju analīzē uzrāda perspektīvas, tad šajā virzienā jāpaplašina autoru izveidotā izplūdušī klasifikācijas sistēma, tajā paredzot iespēju lietotajam izvēlēties piederības funkciju konstruēšanai arī izplūdušās klasifikācijas algoritmu. Šo algoritmu speciāli jāpiemēro izmantošanai piederības funkciju konstruēšanā. Jāturpina arī praktiski eksperimenti, lai pierādītu izplūdušās klasifikācijas sistēmas klasifikācijas potenciālu, salīdzinot ar citām datu ieguves metodēm, ko pielieto bioinformātikas datu klasifikācijā.

Мадара Гаспаровича-Асите, Людмила Алексеева. Система нечеткой классификации данных биоинформатики

В данной статье описывается разработанная авторами система нечеткой классификации, специально предназначенная для анализа и классификации реальных данных биоинформатики. Определена специфика данных биоинформатики – большое число атрибутов и малое число записей. Данные требования заставляют систему классификации учитывать эти особенности данных биоинформатики, что приводит к использованию соответствующих методов предобработки. Дано описание модулей системы, в том числе и схематическое, и обоснование выбора тех или иных алгоритмов и методов для каждого из модулей. Для предобработки данных выбраны следующие методы: а) для замещения пропущенных величин (только если необходимо) – методы опорных векторов SVM или взвешенных К-ближайших соседей (WKNN); б) для снижения числа атрибутов – использование быстрого фильтра, основанного на корреляции FCBF; и в) для конструирования функций принадлежности – элементарный алгоритм построения треугольных функций. Для классификации и тестирования применяется алгоритм нечеткой классификации FUZZY BEXA, а для растягивания правил – метод FURIA, который обеспечивает дополнительное растяжение полученных правил классификации, расширяя тем самым возможности классификации новых записей. Для повышения точности классификации и обеспечения объективного разделения данных на обучающее и тестовое множества применяется 10-кратная перекрестная проверка. Оценка результатов осуществляется путем расчета общей точности и общей ошибки классификации. Поскольку в ходе проведенных авторами исследований использование алгоритмов нечеткой классификации в анализе функций принадлежности оказалось достаточно перспективным, то именно в этом направлении необходимо расширять разработанную авторами систему. При этом пользователям должна быть предоставлена возможность выбора алгоритма нечеткой классификации, специально приспособленного для конструирования функций принадлежности. В ходе дальнейших исследований необходимо сравнить возможности разработанной системы нечеткой классификации с другими методами получения данных, применяемыми для классификации данных биоинформатики.

Stock Market Structural Changes Discovering Helical Structure of Volatility Wave Fourier Image

Andrejs Puckovs¹, Andrejs Matvejevs², ^{1,2} *Riga Technical University*

Abstract – This paper explores an alternative volatility estimation approach discovering the helical structure of Fourier coefficients of volatility wave. Volatility wave is calculated by using wavelet decomposition with consequent logarithmic variance indicator estimation for each decomposed part of the signal and subsequent volatility matrix transform in a specified way. Further, using discrete Fourier transform the Fourier image of obtained volatility wave is analyzed. The Fourier image coefficients of the transformed volatility indicator have a clear helical (spiral) structure that evolves in time. This brings a new understanding of volatility and its evolution process from signal theory (and wavelet theory) perspective. We have found some regularity in the volatility evolution process. The minimum total distance indicator between Fourier coefficients is proposed as a measure of such regularity. This indicator has a nature of volatility lower bound.

Keywords – Continuous wavelet transform, Fourier transform, helical (spiral) structure, minimum total distance, stock indices, time series, volatility, wavelet analysis.

I. INTRODUCTION

This paper continues the series of papers [2]–[5] devoted to the volatility analysis of financial time series – stock market data. The idea of this research is to discover new properties of financial time series using wavelet decomposition.

The main key stone of this research is a vector representation of functions; this idea is widely used in the Fourier transform, where a main function is represented as a sum of elementary functions. The same idea is used in a wavelet transform, where the compact defined function, shifted and scaled, is used as a mother wavelet function, and any signal can be represented as a sum of mother wavelet functions.

In the financial time series analysis, it is important not only to analyze the financial time series (as the signal or stochastic process), but also the volatility indicator (amount of variation or dispersion from the average). Volatility indicator is evolving in time and volatility forecast is very important in the financial time series analysis, since the price of derivative financial instruments is volatility dependent.

The main approaches used in the current paper are as follows:

1. The signal (financial time series) can be decomposed in parts or components by using wavelet filtration, and each component can be analyzed from a volatility perspective. The sum of volatility of all components is proportional to overall signal volatility.

2. The signal can be considered from a volatility perspective and evolving volatility can be considered a signal,

which can be decomposed in components by using wavelet filtration.

Volatility components, which are evolving in time, discover some specific properties of the signal, which cannot be discovered in an original signal without decomposition. Since wavelet decomposition uses the shifted and scaled mother wavelet function, it acts like a microscope, highlighting certain parts of the signal. These specific properties were described in [1]–[4] with the author term “North-East Volatility Wind Effect”.

The main conclusion made in paper [2] is the following – a slight increase in volatility in the low-frequency components of the signal leads to significant disturbances in high-frequency components that destine the entire signal volatility growth. This effect is called “North-East Volatility Wind”. This conclusion is important for risk management on the stock markets. “North-East Volatility Wind” Effect described in [2] brings out a deeper understanding of volatility evolution and an opportunity to illuminate most dramatically market drawdowns initially. This opportunity is explained by an ability to see very small changes in volatility (logarithmic variance) of the low-frequency components of the signal.

Further research of volatility evolution in components (volatility layers) and volatility transmission between layers is conducted in [3].

However, volatility in the perspective of scaling parameter can also be considered the signal (defined in a space of scaling parameter), which has a regular wave form. This signal is named volatility wave, which is the object of current research. Evolving volatility wave is represented by Fourier image by using discrete Fourier transform. Here and further the research algorithm is described.

II. RESEARCH ALGORITHM

Research algorithm consists of 4 (four) blocks: Data Loading & Preprocessing Block, Signal Decomposition Block, Volatility Analysis Block and Fourier Analysis Block. Research algorithm is described in Fig. 1. In the output of each block, output is illustrated in figures – by 2D and 3D plots. Number of each plot is related to the figure number in the current paper.

A. Data Loading and Preprocessing Block

Data Loading & Preprocessing Block uses the ticker (finance.yahoo.com data depositary ticker as id) and provides the analyzed signal $X(t)$ in output. This block consists of 3 main parts: data loader operator, which downloads daily financial

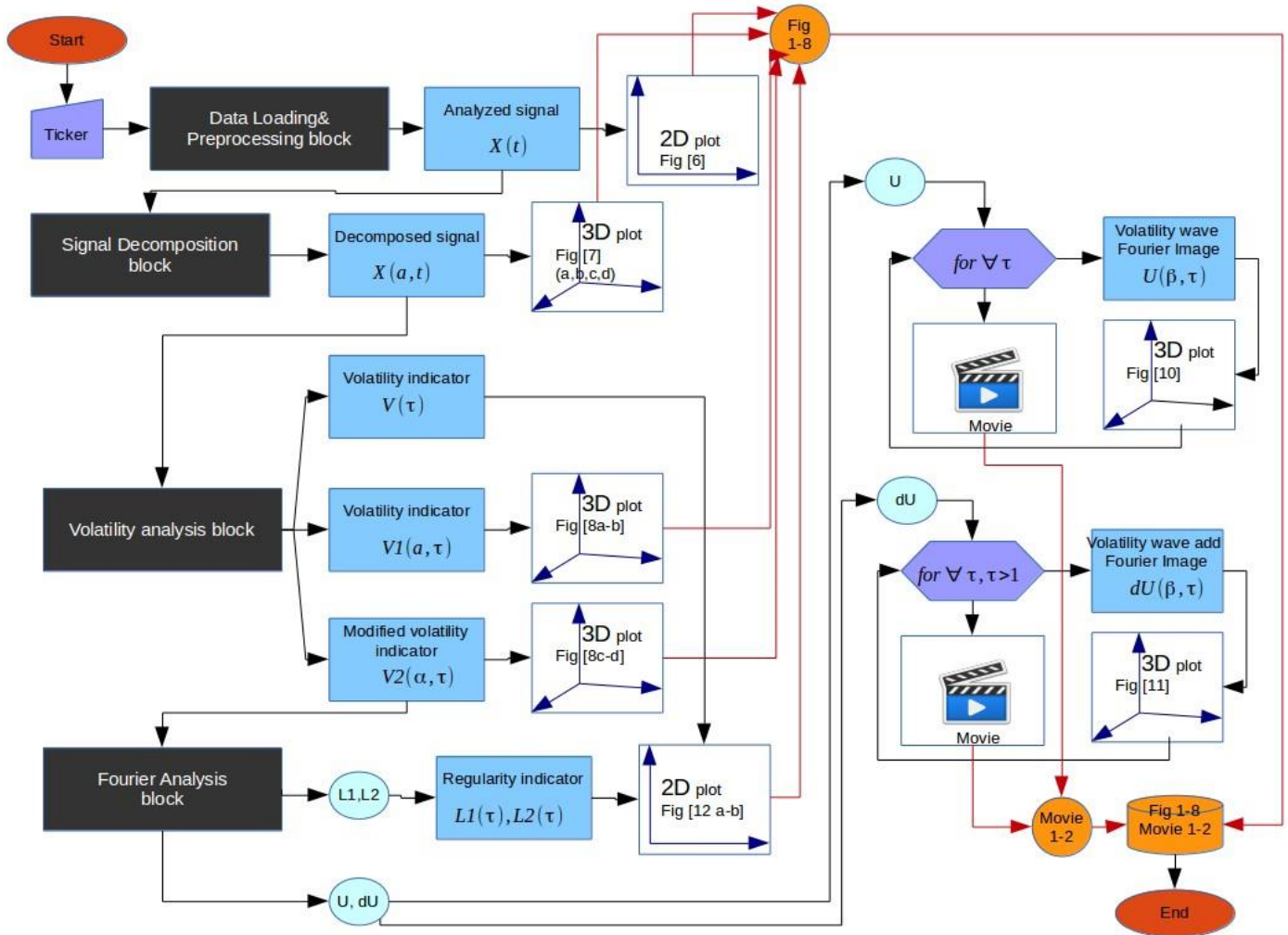


Fig. 1. Research algorithm, block-scheme.

time series (for the whole available period) by a predefined ticker. This operator fits Log-operand, which calculates a natural logarithm of financial time series in order to remove the network effect in financial time series data. Log-operands fit Norm-operand, which is normalizing additions of signal to $N(0,1)$ process, in other words, it is calculating additions to signal $\Delta X(t)$ and normalizing it to the process with zero mathematical expectation $\mu = 0$ and unit standard deviation $\sigma = 1$. Norm-operand provides output of Data Loading & Preprocessing Block, which is called analyzed signal $X(t)$. The algorithm is described in Fig. 2.

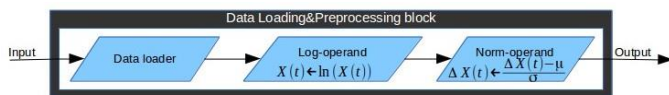


Fig. 2. Data Loading and Preprocessing Block.

Procedure of this block was described in [2]–[4]. Programming codes in Matlab language are shown in [5].

B. Signal Decomposition Block

Signal Decomposition Block is fitted by Data Loading & Preprocessing Block. Signal Decomposition Block is placed

between Data Mining and Preprocessing Block and Signal Analysis Block.

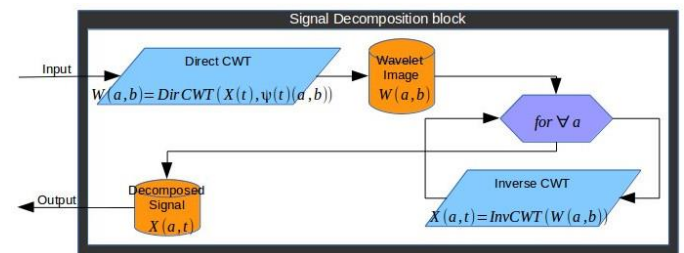


Fig. 3. Signal Decomposition Block.

The block consists of two parts: Direct Continuous Wavelet Transform (Direct CWT) and Inverse Continuous Wavelet Transform (Inverse CWT) operands. Direct CWT operand provides Wavelet image $W(a,b)$ of the analyzed signal $X(t)$. Inverse CWT operand loads Wavelet image $W(a,b)$ for each scaling parameter a and transforms it back to a part of analyzed signal $X(a,t)$. Inverse CWT operand operates in loops ($\forall a$) (for each scaling parameter a), as a result by using Direct CWT and Inverse CWT operands analyzed signal $X(t)$ is decomposed to $X(a,t)$ [2]. Signal decomposition by using wavelet transform keeps the following idea: the number

of signal components only depends on scaling parameter a and by summing all components of decomposed signal original (analyzed) $X(t)$ signal can be reconstructed. Decomposition of analyzed signal $X(t)$ to $X(a,t)$ serves for the subsequent volatility analysis, which has been done for $X(a,t)$. Subsequent volatility analysis is performed for the decomposed signal, which describes volatility evolution since and of volatility transmission. Exploration of Signal Decomposition Block was described in [2]–[4]. Programming codes in Matlab language are shown in [5].

C. Volatility Analysis Block

Volatility Analysis Block is fed by Signal Decomposition Block output – decomposed signal $X(a,t)$. Volatility Analysis Block run in loop for each scaling parameter a . By selecting decomposed signal $X(a,t)$ for defined a the block by using a window function selects 25-trading-day interval τ in time space t (in other words from $X(a,t)$ for certain a it selects $t \in [1:25]$) and calculates logarithmic variance $\ln(D(X(a,\tau)))$ (or volatility indicator $V1(a,\tau)$) [2].

After this window function is shifted in time till the last τ and the procedure is repeated for each a scaling parameter $\forall a$.

Volatility indicator $V1$ is stored in memory and further shown in a 3D plot (in Fig. 8), as well as it is used for modified volatility indicator calculation.

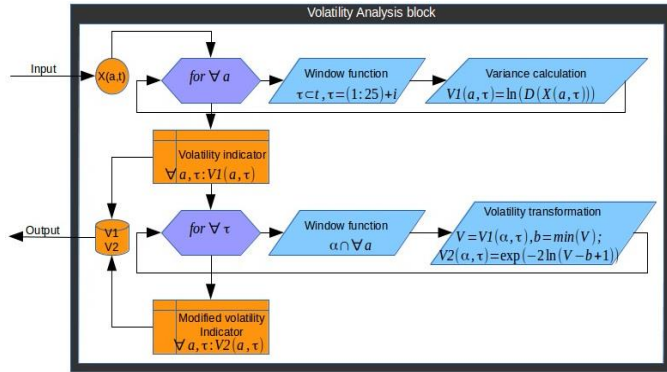


Fig. 4. Volatility Analysis Block.

Modified volatility indicator calculation is based on volatility indicator $V1$ data and starts from running τ index (from one to last τ). For certain τ volatility indicator $V1(a,\tau)$ is selected along a horizon $V1(:,\tau)$ (or in other words $V1(a,\tau), \alpha \in [a_{\min}, a_{\max}]$). Here and further $V1(a,\tau)$ is abbreviated as $V1$ volatility wave. After this, specific transformation with $V1(a,\tau)$ is done in accordance with equation (1). After this the procedure is repeated by changing $\tau \leftarrow \tau + 1$ to the last τ .

$$V2(\alpha, \tau) = \exp(-2 \ln(V - b + 1)), \quad (1)$$

$$V = V1(\alpha, \tau), \alpha \in [a_{\min}, a_{\max}], b = \min(V).$$

Illustration of this transformation is done in Fig. 9. This modified volatility indicator $V2$ by its nature illuminates local maxima lines of $V1$.

Schematic illustration of Volatility Analysis Block is performed in Fig. 4. This block in the output provides $V1$ and $V2$ volatility indicators, which are illustrated in 3D plots in Fig. 9. Modified volatility indicator $V2$ is used for the subsequent analysis in Fourier Analysis Block.

Detailed exploration of volatility indicator $V1$ calculation can be found in [2]–[3]. Programing code in Matlab environment is provided in [5]. Transformation from $V1$ to $V2$ is described in [6].

D. Fourier Analysis Block

Fourier Analysis Block uses modified volatility indicator $V2$ as input and provides several outputs: Fourier image of volatility wave $U(\beta, \tau)$, Fourier image of volatility wave growth $dU(\beta, \tau)$, regularity indicators $L1(\tau)$ (calculated from $U(\beta, \tau)$) and $L2(\tau)$ (calculated from $dU(\beta, \tau)$). Regularity indicators $L1(\tau)$ and $L2(\tau)$ are a measure of minimum total distance between all Fourier image coefficients. Here and further a detailed explanation is provided.

Fourier Analysis Block uses modified volatility indicator $V2$ as input. It runs a running index τ and for certain τ selects volatility wave $V2(:, \tau)$ or $V2(\alpha, \tau), \alpha \in [a_{\min}, a_{\max}]$. This volatility wave is transformed from time to frequency domain by using discrete Fourier transform algorithm, described in (2).

$$U(\beta, \tau) = \langle V2(\alpha, \tau), e^{j\beta\alpha} \rangle. \quad (2)$$

As a result, Fourier image of volatility wave $U(\beta, \tau)$ (where β is a period). Fourier analysis is repeated for each τ . The same procedure is done for volatility wave growth $dV2$. The $dV2$ indicator is calculated by (3).

$$dV2(\alpha, \tau) = V2(\alpha, \tau) - V2(\alpha, \tau - 1). \quad (3)$$

Volatility wave growth representation in frequency domain is done by a fast Fourier transform, as a result Fourier image of volatility wave growth $dU(\beta, \tau)$ is calculated in the block.

Fourier images $U(\beta, \tau)$ and $dU(\beta, \tau)$ have specific properties – Fourier image coefficients demonstrate a clear regular helical (spiral) structure. Fourier images of volatility waves calculated for stock indices are shown in Figs. 10–11.

Analogical results for Fourier analysis of volatility evolution waves for DJIA (The Dow Jones Industrial Index) are shown in papers [6]–[7].

Authors propose Fourier image regularity indicators $L1(\tau)$ and $L2(\tau)$ defined on time space τ according equation (4):

$$L1(\tau) = - \sum_{\forall \beta} |U(\beta, \tau) - U(\beta - 1, \tau)|$$

$$L2(\tau) = - \sum_{\forall \beta} |dU(\beta, \tau) - dU(\beta - 1, \tau)| \quad (4)$$

The block-scheme of Fourier Analysis Block is shown in Fig. 5.

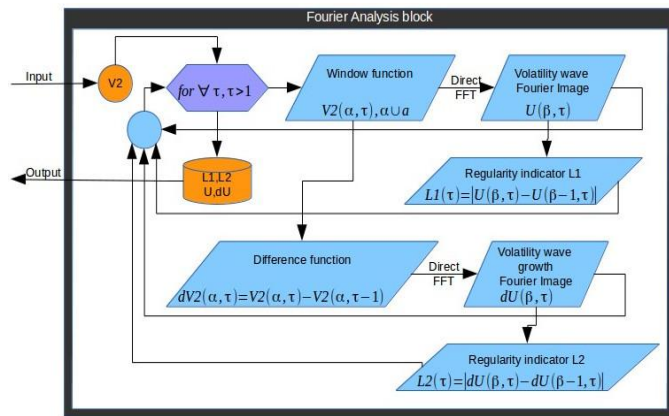


Fig. 5. Fourier Analysis Block.

According to the block-scheme in Fig. 5, the block runs in for loops with running index τ and calculates $U(\beta, \tau), dU(\beta, \tau)$ Fourier images and from them $L1(\tau)$ and $L2(\tau)$ regularity indicators. As a result, the block provides output with $L1, L2$ which are loaded to 2D plots.

By running for loops with running index τ , U, dU Fourier images are loaded to scatter 3D plots and the result image is added to movie, stored in a hard drive. Detailed illustrations of movie making and plotting are shown in Fig. 1.

III. FINANCIAL TIME SERIES ANALYSIS

Here and further, for the subsequent research the Dow Jones Industrial Index financial time series data of the period (1993 – nowadays) are selected.

By feeding Data Loading & Preprocessing Block with ticker (^DJI), the block provides in output the signal shown in Fig. 6.

In the Signal Decomposition Block, the analyzed signal is decomposed in parts by using wavelet filtration – Direct CWT and Inverse CWT. The output is provided from various viewpoints in Fig. 7.

Decomposed signal is used as input for Volatility Analysis Block, which provides a picture of volatility evolution shown in Fig. 8.

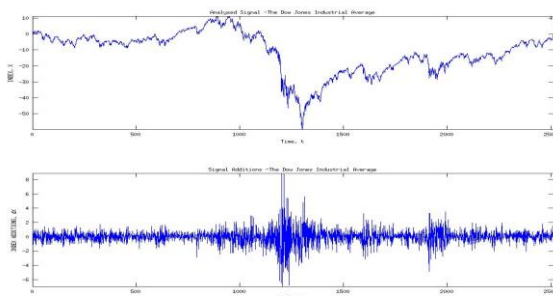


Fig. 6. Analyzed signal of the Dow Jones industrial average (above – the signal, below – signal growth).

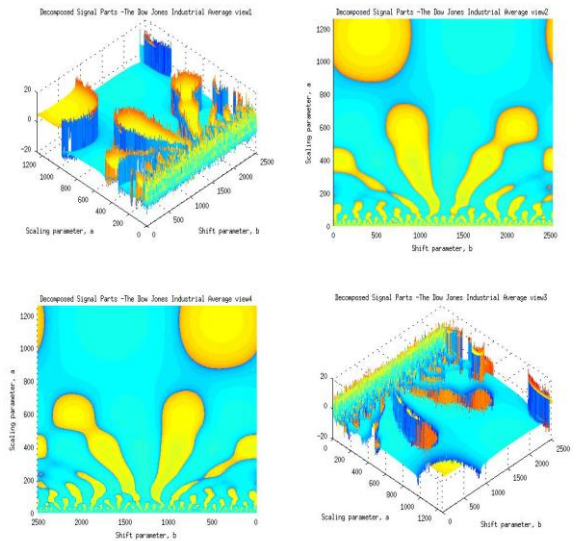


Fig. 7. Decomposed signal of the Dow Jones industrial average.

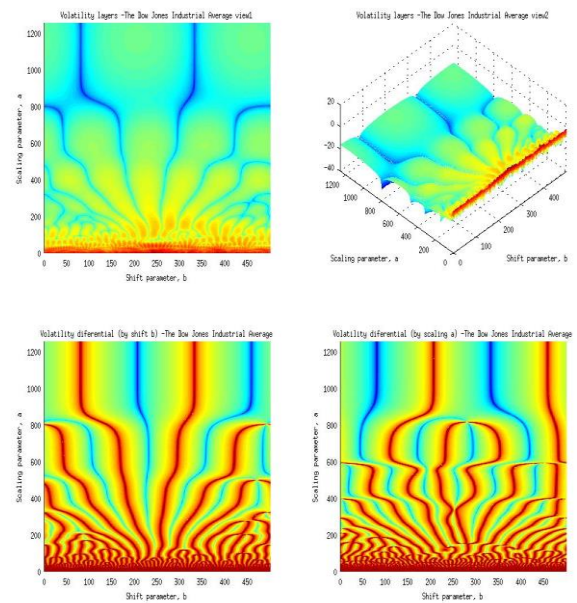


Fig. 8. Volatility analysis of the Dow Jones industrial average.

For the subsequent analysis, the modified volatility data $V2$ are used. The difference between Volatility $V1$ and Modified volatility $V2$ is shown in Fig. 9. Modified Volatility $V2$ illuminates local maxima lines.

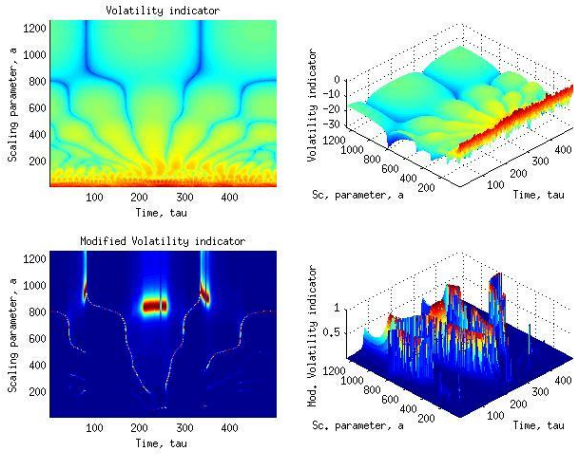


Fig. 9. Volatility (top) and modified volatility (bottom) indicators.

For the subsequent analysis, the modified volatility data V2 fits Fourier Analysis Block, which analyzes volatility waves in a frequency domain. Fourier images of volatility wave $U(\beta, \tau)$ are shown in Fig. 10 (a-c).

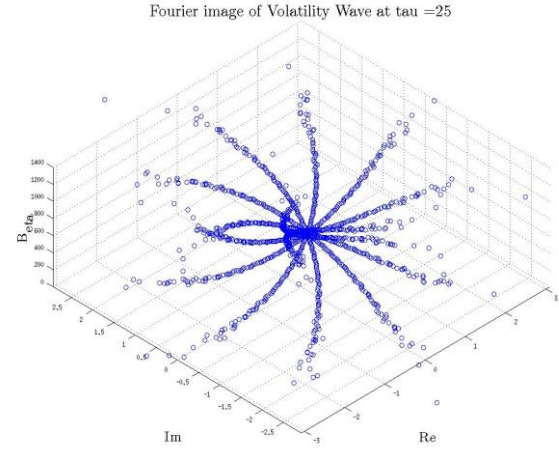
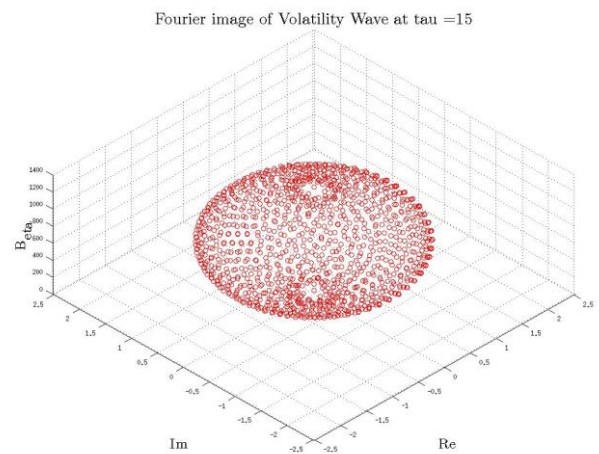
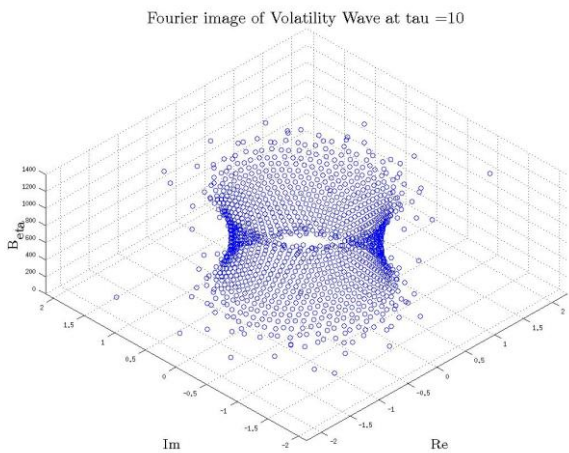
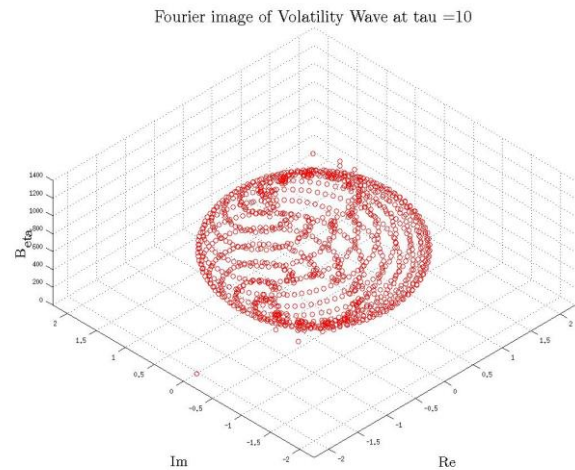
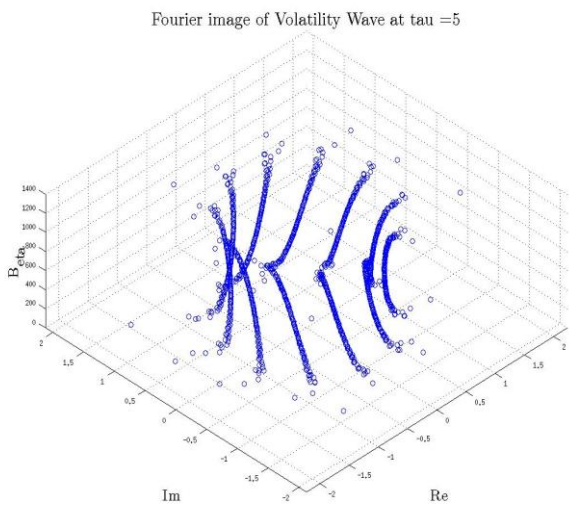


Fig. 10. a-c) Fourier image of volatility wave in time $\tau = \{5, 10, 25\}$.

Fourier images of volatility wave growth $dU(\beta, \tau)$ are shown in Fig. 11 (a-f).



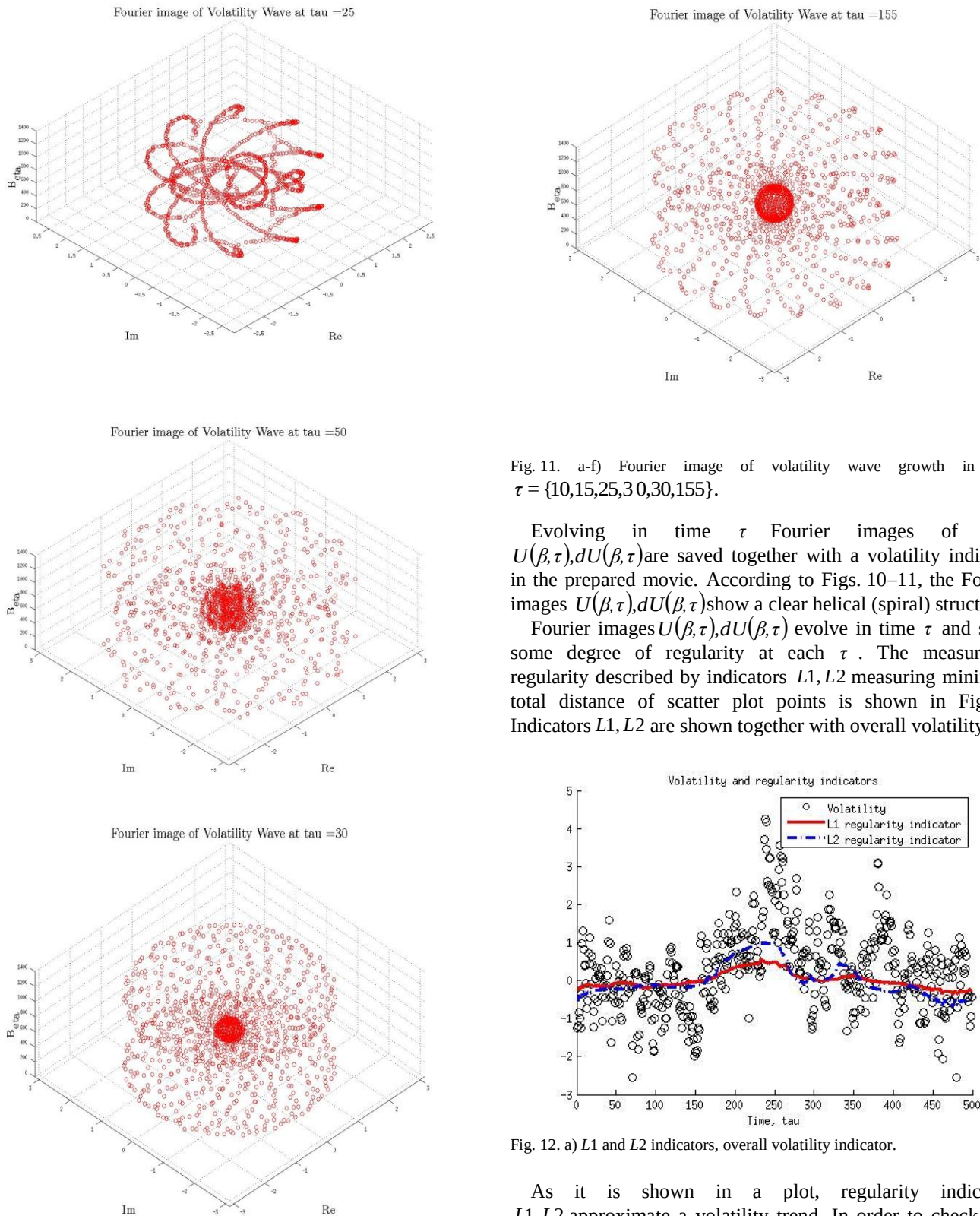


Fig. 11. a-f) Fourier image of volatility wave growth in time $\tau = \{10, 15, 25, 30, 30, 155\}$.

Evolving in time τ Fourier images of both $U(\beta, \tau)$, $dU(\beta, \tau)$ are saved together with a volatility indicator in the prepared movie. According to Figs. 10–11, the Fourier images $U(\beta, \tau)$, $dU(\beta, \tau)$ show a clear helical (spiral) structure.

Fourier images $U(\beta, \tau)$, $dU(\beta, \tau)$ evolve in time τ and show some degree of regularity at each τ . The measure of regularity described by indicators $L1, L2$ measuring minimum total distance of scatter plot points is shown in Fig. 12. Indicators $L1, L2$ are shown together with overall volatility.

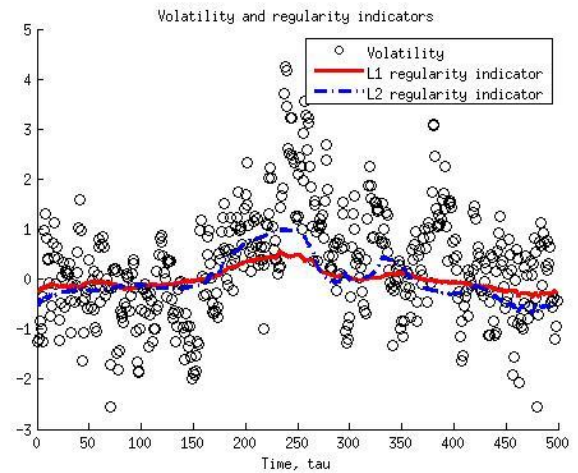


Fig. 12. a) $L1$ and $L2$ indicators, overall volatility indicator.

As it is shown in a plot, regularity indicators $L1, L2$ approximate a volatility trend. In order to check this, by using wavelet filtration (discrete wavelet transform decomposition till the 10th level with Dobeshi wavelet “db7”) high frequency components of volatility indicator were removed. The result is shown in Fig. 12.

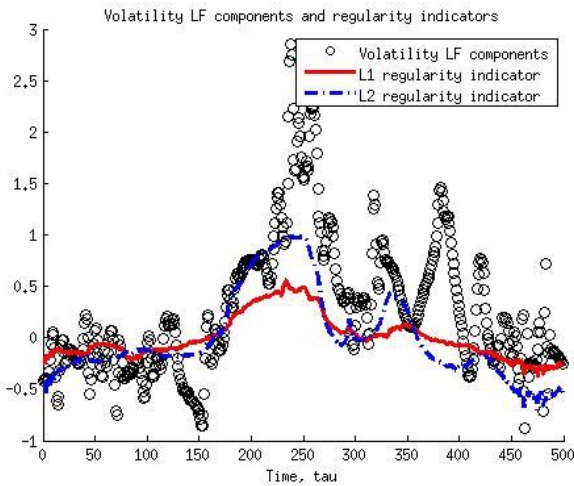


Fig. 12. b) L1 and L2 indicators, LF components of overall volatility indicator.

Regularity indicators $L1, L2$ have high correlation with volatility indicator low frequency components. The Pearson correlation results are shown in Table I.

TABLE I
Pearson Cross-correlation Matrix

	Volatility indicator	L1 indicator	L2 indicator
Volatility indicator	1	0.7443	0.7145
L1 indicator	0.7443	1	0.9665
L2 indicator	0.7145	0.9665	1

In order to bring clearness to the analyzed process – the process of wavelet filtration, evolving volatility measure and Fourier analysis of the so-called volatility wave – the process and output of this process are to be described next.

IV. RESEARCH RESULTS AND DISCUSSION

Suppose that each investor reacts to the events happened on financial markets by making their individual decisions taking into account their own investment horizon. Moreover, this decision is based on filtering the time series of stock prices.

If the situation is stable in the market, individual investors will not change their open positions rapidly, otherwise market instability and price volatility should rise. By itself volatility is a measure of strikes between individual investors. More strikes, more jagged the index is.

Understanding of the so-called volatility wave starts with understanding of interrelations between various investment horizon investors. Local maxima of volatility wave dedicate an investment horizon, where the strike between investors is more noticeable.

Understanding of Fourier image of volatility wave is similar. Maximum magnitude Fourier harmonics should indicate frequency (which is inverse to an investment horizon), which responds to volatility wave local maxima.

The second important understanding of Fourier image of volatility wave is illumination of structure. As it is shown in Fig. 10 and Fig. 11, markets definitely have a structure, since various investment horizon investors should interrelate to each other. This structure is evolving in time, can take different forms. Increase of regularity or emanation of certain harmonics normally means running a volatility wave, which results in the market crisis. This statement becomes clear, understanding $L1, L2$ measures, while the total distance between all Fourier coefficients turns to minimum, a volatility indicator trend (low frequency component) rises.

Discovering the structure of volatility wave Fourier image is very important, currently the minimum total distance $L1, L2$ indicators need modification to describe different regularity structures.

REFERENCES

- [1] C. Chatfield, *The Analysis of Time Series*, NY: Chapman&HALL/CRC, 2006, 329 pp. 255–268.
- [2] A. Pučkovs, A. Matvejevs, “Northeast Volatility Wind,” *Effect. Aplimat Journal*, 2014, vol. 6, pp. 324–339. ISSN 1337-6365.
- [3] A. Pučkovs, “Stock Indexes Volatility Evolution Research in Volatility Layers” *Zhurnal nauchnykh publikatsiy aspirantov i doktorantov*, 2014, no. 5(95) Mai 2014, pp. 32–37. ISSN 1991-3087.
- [4] A. Pučkovs, “Stochastic Process, Based on North-East Volatility Wind Effect,” *Zhurnal nauchnykh publikatsiy aspirantov i doktorantov*, 2014, no. 5(95) Mai 2014, pp. 220–226. ISSN 1991-3087.
- [5] A. Pučkovs, “Computational Aspects of North-East Volatility Wind Effect,” *Zhurnal nauchnykh publikatsiy aspirantov i doktorantov*, 2014, no. 5(95) Mai 2014, pp. 261–274. ISSN 1991-3087.
- [6] A. Pučkovs, “North-East Volatility Wind Effect in Traveling Wave Perspective. Helical Structure of Volatility wave Fourier Coefficients,” *Zhurnal nauchnykh publikatsiy aspirantov i doktorantov*, 2014, July, pp. 5–5. ISSN 1991-3087.
- [7] A. Pučkovs, “Alternative Financial Timeseries Risk Indicator Estimation by using North-East Volatility Wind Effect” *Zhurnal nauchnykh publikatsiy aspirantov i doktorantov*, 2014, July, pp. 10–10. ISSN 1991-3087.
- [8] V.M. Gadre, *Lecture Series in Multirate Digital Signal Processing and Multiresolution Analysis*. Bombay: IIT Bombay [Online] Available: <http://www.youtube.com/watch?v=iNgV2NEyLhQ> [Accessed 15 Oct. 2013].
- [9] T.K. Basu, *Lecture Series on Digital Signal Processing*. Kharagpur: IIT Kharagpur [Online] Available: <http://www.youtube.com/watch?v=8zbBJkw5KsM>.
- [10] Yahoo. Dow Jones Industrial index data. [Online]. Available: <http://finance.yahoo.com/q/hp?s=%5EDJI+Historical+Prices> [Accessed: Oct. 1, 2013].
- [11] Kronover, *Fraktaly i kaos v dinamicheskikh sistemakh*. Osnovy. teorii. M: Postmarket, 2000. p. 352.
- [12] A. N. Shiryayev. *Osnovy stokhasticheskoy finansovoy matematiki*. M: Fazis, 1998. p. 983.
- [13] A. N. Yakovlev, *Vvedeniye v veyvlet preobrazovaniye*. Novosibirsk: MINOBR Novosibirskiy GTU, 2003, p. 102.
- [14] A. V. Novopashina, V. A. Sankov, V. Y. U. Buddo, *Prostranstvenno-vremennyy analiz seysmicheskikh struktur baykalskoy riftovoy sistemy*. Irkutsk: Institut zemnoy kory SO RAN, 2011, pp. 55–64.

Andrejs Puckovs was born in Riga in 1985. He graduated from Riga Technical University (Latvia), where he received his Bachelor Degree in Management in 2007. He has the Qualification of Economist and the Professional Master degree in Economics (Financial Analysis). Andrejs Puckovs is currently completing his research for the Doctoral degree at Riga Technical University in the field of Mathematical Statistics and Its Applications. From 2008 to 2010 he worked as a Fund Administrator in Custody and Correspondent Banking Department, JSC Swedbank. E-mail: andrejs.puckovs@rtu.lv.

Andrejs Matvejevs is the Doctor of Technical Sciences in Information Systems. Until 2009 he was a Chief Actuary at the insurance company "BALVA". For more than 25 years he has taught at Riga Technical University and Riga International College of Business Administration, Latvia. His previous research was devoted to solving of dynamical systems with random

perturbation. His current professional research interests include applications of Markov chains to actuarial technologies: mathematics of finance and security portfolio. He is the author of about 40 scientific publications, two textbooks and numerous conference papers.
E-mail: andrejs.matvejevs@rtu.lv.

Andrejs Pučkovs, Andrejs Matvejevs. Akciju tirgu struktūras izmaiņu atklāšana, izpētot volatilitātes viļņu Furjē attēla spirālveida struktūru

Šajā rakstā tiek aplūkoti alternatīvie volatilitātes rādītāji, kas balstīti uz volatilitātes viļņu Furjē attēla spirālveida regularitātes mērījumiem. Volatilitātes vilnis tiek iegūts ar veivlet filtrāciju (ar tiešo un apgriezto nepārtraukto viļņu pārveidojumiem), veicot analizējamā signāla (finanšu laikrindu) dekompozīciju ar turpmāko volatilitātes (jeb logaritmiskās dispersijas) pētījumu katrā signālu komponentē. Turpmāk volatilitātes rādītājs tiek pārveidots noteiktā veidā, tā kā ir aprakstīts šai rakstā. Šī pārveidojuma jēdziens ir volatilitātes indikatora mērogošana un lokālo maksimumu izšķiršana. Modificēts volatilitātes rādītājs tiek analizēts mērogošanas rādītāja griezumā katrā laika momentā „tau”. Modificēts volatilitātes rādītājs mērogošanas parametra griezumā veido viļņveida formu, dēvētu par volatilitātes vilni. Ar Furjē analīzi volatilitātes vilnis tiek pārveidots Furjē attēlā. Rezultātā reālās un imaginārās Furjē attēla daļas veido regulārās formas spirālveida struktūru. Furjē attēla regularitāte tiek noteikta ar $L1$ un $L2$ rādītājiem, kas tiek aprēķināti, kā minimālā distance starp visiem volatilitātes viļņu Furjē attēla koeficientiem. $L1$ un $L2$ rādītāji pēc savas dabas ir alternatīvi volatilitātes rādītāji. Šis raksts atklāj jaunu skatu uz volatilitāti un to evolūciju Furjē attēla koeficientu spirālveida struktūras griezumā. Šī pieeja ļauj prognozēt jaunas finanšu krīzes rašanos.

Андрей Пучков, Андрей Матвеев. Выявление структурных изменений фондовых рынков, исследующих спиралевидную структуру Фурье образа волн волатильности

В данной статье рассматриваются альтернативные оценки волатильности, основанные на оценке регулярности спиралевидной структуры Фурье образа волн волатильности. Волна волатильности, получаемая с помощью вейвлет-декомпозиции (Прямого Непрерывного вейвлет-преобразования и Обратного Непрерывного вейвлет-преобразования), посредством декомпозиции анализируемого сигнала (финансового временного ряда) с последующим вычислением показателя волатильности (логарифмической дисперсии) для каждого компонента сигнала. Далее показатель волатильности преобразуется согласно преобразованию, описанному в настоящей работе. Суть данного преобразования заключается в масштабировании и выделении линий локальных максимумов для показателя волатильности. В результате преобразования получаем модифицированный показатель волатильности. Модифицированный показатель волатильности анализируется в разрезе показателя масштаба для каждого временного показателя «тау». Модифицированный показатель волатильности в разрезе показателя масштаба образует волнообразную форму, названную волной волатильности. Посредством Фурье анализа волны волатильности выявляется Фурье образ волны волатильности. В результате мнимые и действительные части коэффициентов Фурье образуют четкую спиралевидную структуру регулярной формы. Регулярность или структурность Фурье образа определяется с помощью показателей $L1$, $L2$, рассчитываемых как минимальное расстояние между всеми коэффициентами Фурье образа волны волатильности. Показатели регулярности $L1$, $L2$ являются альтернативными показателями волатильности. Данная статья открывает новое видение волатильности и её эволюции с точки зрения спиралевидной структуры коэффициентов Фурье образа волны волатильности. Данное видение позволяет открыть новые способы предсказания финансовых кризисов.

Modeling VIX Index Based on Semi-parametric Markov Models with Frank Copula

Andrejs Matvejevs¹, Jegors Fjodorovs², ^{1,2} *Riga Technical University*

Abstract – The research studies the estimation of a semi-parametric stationary Markov models based on a Frank copula density function. Described techniques allow us to estimate the parameters of the Frank copula, which has a better fit compared to previously selected regression models (estimators of the marginal distribution and the copula parameters are provided). We show how to apply our technique to the financial index VIX – a market mechanism that measures the 30-day forward implied volatility of the underlying index, the S&P500. Moreover, using MatLab we made VIX option index study – found the best copula fit under our condition, estimated nonlinear parameters and showed evaluation steps for copula based semi-parametric models.

Keywords – Diffusion processes, Frank copula, Markov models, semi parametric regressions, VIX index.

I. INTRODUCTION

Copulas have become popular in the finance and insurance community in the past years, where modeling and estimating the dependence structure between several univariate time series are of great interest; see Frees and Valdez (1998) [1] and Embrechts et al. (2002) [2] for review.

A copula function is a multivariate distribution function with standard uniform marginals. By Sklar's (1959) [3] theorem, one can always model any multivariate distribution by modeling its marginal distributions and its copula function separately, where the copula captures all the scale-free dependence in the multivariate distribution [4]. The central result of this theorem, which states that any continuous N-dimensional cumulative distribution function F, evaluated at a point $X = (X_1, \dots, X_n)$ can be represented as

$$F(x) = C(F_1(x_1), \dots, F_n(x_n)),$$

where C is called a copula function and $F_i(x_i)$, $i = 1, \dots, n$ are the marginal distributions [4]. The use of copulas, therefore, splits a complicated problem (finding a multivariate distribution) into two simpler tasks. The first task is to model the univariate marginal distributions and the second task is finding a copula that summarizes the dependence structure between them.

It is also useful to represent copulas as joint distribution functions of standard uniform random variables:

$$U = F(X_1) \text{ and } V = F(X_2) \\ C(u, v) = P(U \leq u, V \leq v)$$

The outcome of uniform random variables falls into the interval [0, 1]; therefore, the domain of a copula must be the N-dimensional unit cube. Similarly, because the mapping represents a probability, the range of the copula must also be the unit interval. It is also easy to determine the value of a copula on the border of its domain. When one argument equals zero, the probability of any joint event must also be zero. Similarly, when all but one of the inputs is equal to one, the joint probability must be equal to the (marginal) probability of the argument that does not equal one. Finally, the function must be increasing in all its arguments [4].

Apart from the standard distribution functions, copulas have associated densities:

$$c(u, v) = \frac{\partial^2 C(u, v)}{\partial u \partial v}$$

which permit the bivariate density $f(u; v)$ as the product of the copula density and the density functions of the margins

$$f(u, v) = c(F_1(u), F_2(v)) f_1(u) f_2(v)$$

The expression above indicates how the simple product of two marginal distributions will fail to properly measure the joint distribution of two asset prices unless they are in fact independent and the dependence information captured by the copula density [4], $c(F_1(u), F_2(v))$, is normalized to unity.

II. COPULA-BASED SEMI-PARAMETRIC MODELS FOR STOCHASTIC SIMULATIONS

The possibility of identifying nonlinear time series using nonparametric estimates of the conditional mean and conditional variance was studied in many papers (see, for example, [5], and references there). As a rule, analyzing the dependence structure of stationary time series $\{x_t, t \in \mathbb{Z}\}$ regressive models defined by invariant marginal distributions and copula functions that capture the temporal dependence of the processes. As indicated in [5], this allows separating out the temporal dependence (such as tail dependence) from the marginal behavior (such as fat tails) of a time series. One more advantage of this type regressive approach is a possibility to apply probabilistic limit theorems for transition from difference equations to continuous time stochastic differential equations ([5], [6]). In our paper, we also study a class of copula-based semi-parametric stationary Markov models in a form of scalar difference equation

$$t \in Z: X_t = X_{t-1} + \varepsilon f(X_{t-1}, \varepsilon) + \varepsilon g(X_{t-1}, \varepsilon) \xi_t \quad (1)$$

where $\{\xi_t, t \in Z\}$ is i.i.d., $N(0; 1)$, and ε is a small positive parameter, which will be used for diffusion approximation of (1). Regressions (1) are high-usage equations for simulation and parameter estimation of stochastic volatility models ([5]). Unfortunately, defined by (1) Markov chain has incompact phase space that complicates an application of probabilistic limit theorem. Copula approach helps to simplify asymptotic analysis of (1). Due to persistence of small parameter ε after a substitution $U_t = F(X_t)$ in equation (1) for a further diffusion approximation one can write a difference equation in the same form like (1):

$$t \in Z: U_t = U_{t-1} + \mathcal{A}(U_{t-1}, \varepsilon) + \mathcal{B}(U_{t-1}, \varepsilon) \xi_t \quad (2)$$

But now this equation defines a Markov chain on the compact $[0, 1]$. This makes it easier to formulate construction for transition probability and further estimators of functions $f(u)$ and $g(u)$. After diffusion approximation of (2) one can make a inverse substitution and derive a stochastic differential equation as diffusion approximation for (1).

Described algorithm allows evaluating the parameters of copula, which have the best fit to previously selected model. In our copula dependence study we used MatLab, which helps to evaluate copula parameters and choose the best copula class, based on log likelihood estimation, for the selected financial market data. These copula based models are easy to simulate and can be expressed as semi-parametric regression transformation models. Moreover, using this MatLab we made VIX option index simulation – found the best copula fit under our condition and built semi-parametric autoregression.

III. PROPOSED ESTIMATION ALGORITHM AND EVALUATION OF PARAMETERS

Let us take into consideration VIX – Market Volatility Index – daily data from 25 October 2009 to 1 October 2014. The VIX is a market mechanism that measures the 30-day forward implied volatility of the underlying index, the S&P 500. Being able to meaningfully interpret movements in the VIX and its reaction to market events can give investors an edge in managing the risk and profitability of their trading book and in designing portfolio strategies using VIX derivatives to capitalize on the dynamic and time-varying correlation of the VIX with its underlying S&P 500 Index. Let us build for this option index semi-parametric copula based model, using AIC and BIC criteria.

The easiest way of parameter estimating of the semi-regressive model for the VIX index would be to hold the algorithm:

- To simulate U_t points, which are $R[0,1]$ (uniform) or transform the existing sample into $R[0,1]$;
- To build a scatter plot for (U_{t-1}, U_t) ;
- To make several statistical tests to find the suited distribution of data;

- To take into account a scatter plot and distribution of data, try to choose copula from the existing class or build your own copula, if you know marginal distributions;
- To test copula consistency to data (for example AIB and BIC);
- To find regression parameters.

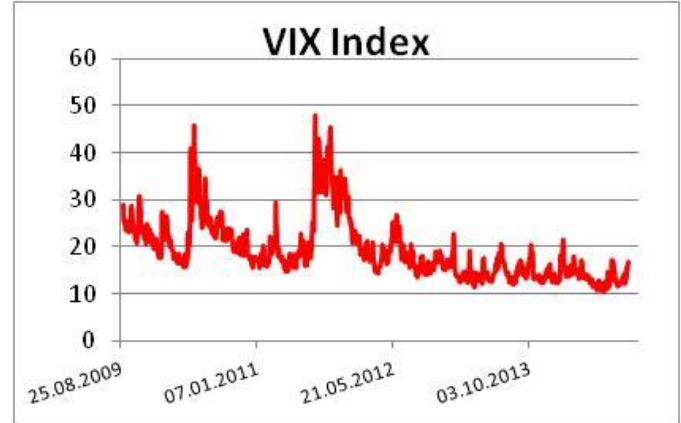


Fig. 1. Historical VIX index levels.

Using Matlab program, we built scatter plots for VIX index transformed into uniform distribution $R[0,1]$ and non-transformed data.

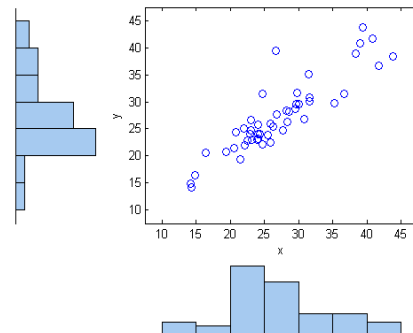


Fig. 2. Scatter plot for non-transformed VIX index data.

An important issue faced by an applied researcher interested in using the class of semi-parametric copula-based time series models is the choice of an appropriate parametric copula [7]. In different papers, Chen et al. (1998) [8] propose two simple tests for the correct specification of a parametric copula in the context of modeling the contemporaneous dependence between several univariate time series and of the innovations of univariate GARCH models used to filter each univariate time series.

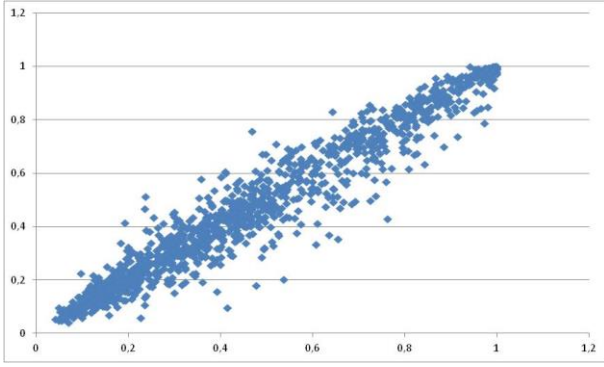


Fig. 3. Scatter plot for transformed into R[0,1] VIX index data.

Chen and Fan (2004b) [9] establish pseudo-likelihood ratio tests for selection of parametric copula models for multivariate i. i. d. observations under copula misspecification [4]. But our suggestion is simpler – we can choose the best copula fit using

AIC and BIC criteria or using χ^2 test for data distribution. We take for different copula comparisons AIC and BIC.

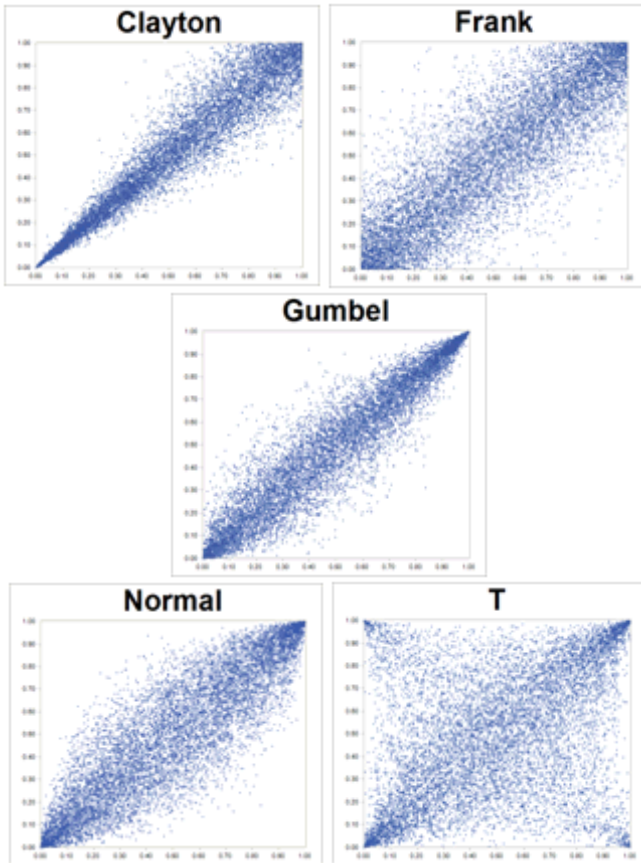


Fig. 4. Most common types of copula in finance (theoretical illustrations).

For the first copula choosing step, it is reasonable to compare graphical parametric copulas with VIX data scatter plot (Fig. 2), (Fig. 3). As we can see, the most suitable copulas for our data are Gumbel, Frank and Normal. For this sample of copulas, is useful to calculate AIC and BIC criteria.

TABLE I

AIC AND BIC CRITERIA FOR VIX INDEX DATA

Copula	AIC	BIC
Gumbel copula	-109.3	-105.7
Frank copula	-255.1	-247.1
Normal copula	-248.9	-242.2

Taking into account AIC and BIC criteria, we should choose the Frank copula for further model estimation. Let us see how to derive semi-parametric regression parameters using Frank copula representation:

$$C(u_{t+1}, u_t) = -a^{-1} \ln(1 + \frac{g_{u_{t+1}} g_{u_t}}{g_1}) \quad (3)$$

$$g_{u_t} = e^{-au_t} - 1$$

Inserting expression (3) into conditional expectation, we get our parameters:

$$\begin{aligned} E(U_{t+1} | U_t = u) &= \int_0^1 u_{t+1} dF_{u_{t+1}|u_t}(u) = \int_0^1 u_{t+1} p(u_{t+1} | u_t) du_{t+1} = \\ &= \int_0^1 u_{t+1} \frac{\partial C(u_{t+1}, u_t)}{\partial u_{t+1} \partial u_t} du_{t+1} = \int_0^1 u_{t+1} c(u_{t+1}, u_t) du_{t+1} = \\ &= -a \int_0^1 u_{t+1} \frac{g_1(1 + g_{u_t+u_{t+1}})}{(g_{u_t} g_{u_{t+1}} + g_1)^2} du_{t+1} = \\ &= \frac{e^{au_t} ((e^a - 1)(\ln(e^{au_t}(e^a - 1)) - \ln(e^a(e^a - 1))) + a(e^a - e^{au_t}))}{a(e^{au_t} - 1)(e^a - e^{au_t})} \end{aligned} \quad (4)$$

$$g((U_{t+1} | U_t = u)) = E((U_{t+1} - f(U_t))^2 / U_t = u) =$$

$$\begin{aligned} &= \int_0^1 (U_{t+1} - f(U_t))^2 c(U_{t+1}, U_t) dU_{t+1} = \\ &= \int_0^1 \frac{1}{t} (\alpha \ln(t) - \frac{\alpha \ln(t)(\exp(\alpha) - 1)}{(1 - \frac{1}{t})(\exp(\alpha) - t)} + \frac{t}{\exp(\alpha) - t})^2 x \\ &\quad (1 - \exp(-\alpha)) \exp(-\alpha u_t) \frac{1}{t} \\ &\quad x \frac{dt}{(1 - \exp(-\alpha) - (1 - \frac{1}{t})(1 - \exp(-\alpha u_t)))^2} \end{aligned} \quad (5)$$

where $t = \exp(U_{t+1})$

It is impossible to solve analytically (4) and (5) expressions. But numerically it is doable, for example, in the Matlab or Mathematica. For the Frank copula, we can use an inverse

function with the aim to return to our base equation (1). Besides integral (5) diverges at 0 point, as a result it is impossible to find solution to (5).

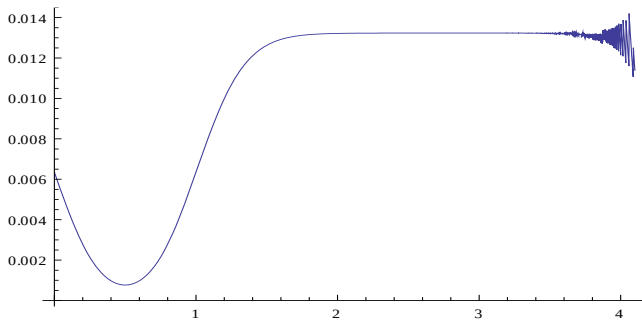


Fig. 5. Second moment of U_t in an interval $[0.001; 4]$ with a parameter $\alpha = 8.68$.

However, using Mathematica software it is possible to draw values of equation (5). As it is seen in Fig. 4, equation (5) in the interval $(0;1)$ is nonlinear. This second moment adds a nonlinear effect to regression (1). This second moment starts to diverge from point 1, but visual divergence is seen from point 2.7.

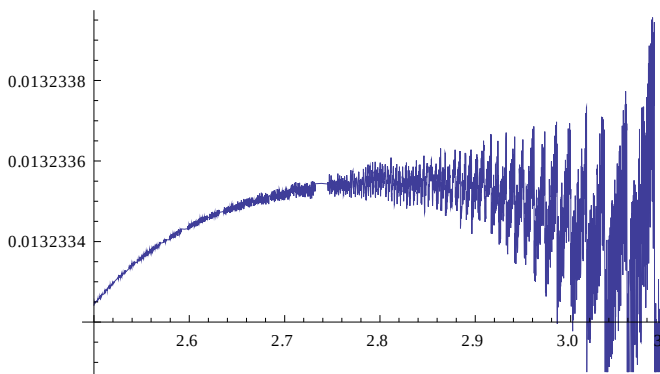


Fig. 6. Second moment of U_t in an interval $[0.001; 4]$ with a parameter $\alpha = 8.68$.

The final Markov copula based semi-parametric regression model in a form of (2) is equal to the sum of (4) and (6):

$$\begin{aligned}
 t \in Z : U_t &= U_{t-1} + f(U_{t-1}, \varepsilon) + g(U_{t-1}, \varepsilon) \xi_t = \\
 &= U_{t-1} + \\
 &+ \frac{e^{au_t} ((e^a - 1) (\ln(e^{au_t} (e^a - 1)) - \ln(e^a (e^a - 1))) + a(e^a - e^{au_t}))}{a(e^{au_t} - 1)(e^a - e^{au_t})} + \\
 &+ \int_0^1 \frac{1}{t} (\alpha \ln(t) - \frac{\alpha \ln(t)(\exp(\alpha) - 1)}{(1 - \frac{1}{t})(\exp(\alpha) - t)} + \frac{t}{\exp(\alpha) - t})^2 x
 \end{aligned}$$

$$x \frac{(1 - \exp(-\alpha)) \exp(-\alpha u_t) \frac{1}{t}}{(1 - \exp(-\alpha) - (1 - \frac{1}{t})(1 - \exp(-\alpha u_t)))^2} dt \xi_t$$

where $\{\xi_t, t \in Z\}$ is i. i. d., $N(0; 1)$.

But if we deal with copulas, we should not skip some facts. For example, it is not easy to say which parametric copula best fits a given dataset, since some copulas may fit better near the center and other near the tails and many copulas do not have moments that are directly related to the Pearson correlation; it is difficult to compare financial models based on correlation. Of course, if we want to use this model in practice, it is crucial to compare different class models which could be suitable for these data. This can give an applied added value for this method.

IV. CONCLUSIONS AND FURTHER WORK

The algorithm for copula simulation and semi-parametric regression coefficient finding using a Markov chain has been presented. For Option VIX index data using MatLab the best fitted copula model, which is the Frank copula, has been found. According to this copula, principals have been shown and parameters of the semi-parametric regression model coefficients have been evaluated in a copula space. The future research may be devoted to developing numerical algorithms for simulation of trajectories of various random processes, evaluating the characteristic of semi-parametric models based on copula, and finding continuous stochastic models using limit theorems [12].

REFERENCES

- [1] E. W. Frees, E. A. Valde, "Understanding relationships using copulas". *North American Actuarial Journal* 2, 1998, pp. 1–25. <http://dx.doi.org/10.1080/10920277.1998.10595667>
- [2] P. Embrechts, A. McNeil, D. Straumann, "Correlation and dependence properties in risk management: properties and pitfalls," in: Dempster, M. (Ed.), *Risk Management: Value at Risk and Beyond*. Cambridge University Press, Cambridge, pp. 176–223, 2002. <http://dx.doi.org/10.1017/CBO9780511615337.008>
- [3] A. Sklar, "Fonctions de repartition a dimensionset leurs marges," *Publ. Inst. Statist. Univ. Paris* 8, 229–231, 1959.
- [4] W. Darsow, B. Nguyen, E. Olsen, "Copulas and Markov processes," *Illinois Journal of Mathematics* 36, 600–642, 1992..
- [5] X. Chen, Y. Fan, "Estimation of copula-based semiparametric time series models," *Journal of Econometrics*, 2006. <http://dx.doi.org/10.1016/j.jeconom.2005.03.004>
- [6] D. B. Nelson, "ARCH models as diffusion approximations," *Journal of Econometrics*, vol. 45, no. 1–2, pp. 7–38, 1990.
- [7] Y. Ait-Sahalia, R. Kimmel, "Maximum likelihood estimation of stochastic volatility models," *Journal of Financial Economics*, 2007. <http://dx.doi.org/10.1016/j.jfineco.2005.10.006>
- [8] X. Chen, L. P. Hansen, M. Carrasco, "Nonlinearity and temporal dependence," Working Paper, University of Chicago, 1998.
- [9] X. Chen, Y. Fan, "Pseudo-likelihood ratio tests for model selection in semiparametric multivariate copula models," *Canadian Journal of Statistics*, 2004.
- [10] H. R. Stoll, R. E. Whaley, "The Dynamics of Stock Index and Stock Index Futures Returns," *Journal of Financial and Quantitative Analysis*, no. 25, 1990, pp. 441–468. <http://dx.doi.org/10.2307/2331010>
- [11] H. Joe, *Multivariate Models and Dependence Concepts*. Chapman & Hall/CRC, 1997. <http://dx.doi.org/10.1201/b13150>

- [12] J. Carkovs, "On Diffusion Approximation of Discrete Markov Dynamical Systems. In Computational Geometry," *Proceedings of World Academy of Science, Engineering and Technology*, vol. 30 pp. 1–6, 2008.

Jegors Fjodorovs is a Doctoral student. He received the Master Degree in Mathematics from the University of Latvia. He received the Qualification of Mathematician–Statistician from the same university. Since 2011 he has been studying at the doctoral study program at Riga Technical University, the Chair of Probability Theory and Mathematical Statistics. The main fields of research are stochastic differential equations, Markov processes and copulas with application in finance mathematics and econometrics.

Presently he is a Lecturer with Riga Technical University in the course "Random Processes" and a Risk Engineer with JSC Swedbank in the Wealth Management Division. He has also worked for seven years at the State Treasury of the Republic of Latvia, where he was able to manage market risks of government assets/liabilities.

E-mail: Jegors.Fjodorovs@rtu.lv.

Andrejs Matvejevs received the Doctoral Degree in 1989. He graduated from the Faculty of Automation and Computing Technique, Riga Technical University, and became an Associate Professor with Riga Technical University in 2000 and a Full Professor in 2005. He has made the most significant contribution in the field of actuarial mathematics. Andrejs Matvejevs is a Doctor of Technical Sciences in Information Systems. Until 2009 he was a Chief Actuary at the insurance company "BALVA". For more than 25 years he has taught at Riga Technical University and Riga International College of Business Administration, Latvia. His previous research was devoted to solving dynamical systems with random perturbation. His current professional research interests include applications of Markov chains to actuarial technologies: mathematics of finance and security portfolio. He is the author of about 40 scientific publications, two study books and numerous conference papers.

E-mail: Andrejs.Matvejevs@rtu.lv.

Andrejs Matvejevs, Jegors Fjodorovs. VIX indeksa modelēšana, izmantojot neparametriskos Markova modeļus ar Franka kopulu

Šajā rakstā tiek aprakstīts neparametriskās Markova modeļa novērtēšanas algoritms, izmantojot Franka kopulas blīvumu. Uz kopulām bāzētās neparametriskās regresijas atšķiras ar to, ka pētnieks varētu sadalīt dažāda veida efektus (risku avotus) un katru modelēt atsevišķi (neparametriskais marginālais sadalījums un parametriskā kopulas funkcija), bet kopula savieno visu no mēroga neatkarīgo laika sakarību. Rakstā tika izmantots finanšu indekss VIX – mēra 30 dienas uz priekšu netiešu svārstīgumu, bāzes indeksa S & P 500. Šis indekss aprēķināms no akciju opciju cenām. Aprakstītā pieeja ļauj mums novērtēt parametrus Franka kopulai, kura pēc noteiktajiem statistiskajiem kritērijiem ir labākais variants VIX indeksa datu vēsturiskās atkarības aprakstīšanai. Turpmāk, balstoties uz Franka kopulas blīvuma funkciju, tika aprakstīts neparametriskās Markova regresijas koeficientu atrašanās mehānisms. Šī modeļa parametru noteikšana ir sarežģīta – analītiski nav atrisinājuma (parametriskais integrāls diverģē 0 punktā). Tādējādi tika izmantotas MatLab un Mathematica skaitliskās metodes, kas ļauj pārliecināties par pareizo metodoloģisko pieeju – no ilustrētajiem grafikiem var redzēt, ka otrais moments pievieno nelinearitāti aprakstītajam vienādojumam. Rezultātā, izmantojot aprakstīto metodoloģiju, var imitēt VIX indeksa dinamiku dažādiem laika intervāliem un iegūtos novērtējumus izmantot finanšu riska vadībai (riska ierobežošanas operācijas, izmantojot ar opcijām) vai pieņemot spekulatīvas tirdzniecības pozīcijas ar opcijām.

Андрей Матвеев, Егор Фёдоров. Моделирование VIX индекса посредством непараметрических Марковских моделей с копулой Франка

Данная статья описывает алгоритм оценки непараметрической Марковской модели с помощью плотности копулы Франка. Копульные непараметрические регрессии отличаются тем, что исследователь может разделить различные виды (источники) риска, каждый смоделировать отдельно (непараметрические маргинальные распределения и параметрическая копульная функция) и соединить копулой, свободной от масштаба временной зависимости. В статье был использован финансовый индекс VIX, измеряющий 30-дневную будущую внутреннюю волатильность на основе индекса акций S & P 500. Этот индекс рассчитывается, исходя из цен опционов. Описанный подход позволяет оценить параметры копулы Франка, правильность выбора которой устанавливается с помощью статистических критериев и является лучшим для данных индекса VIX. То есть эта копула лучше остальных копул описывает историческую зависимость. Далее, на основе функции плотности Франка копулы, был описан механизм оценки коэффициентов непараметрической Марковской регрессии. Такая оценка параметров трудоёмка – нет аналитического решения (параметрический интеграл расходится в точке 0). Таким образом, вычисление параметров происходит с использованием численных методов в пакетах Matlab и Mathematica. Проверить правильность подхода позволяют графические иллюстрации, где можно видеть, что второй момент, добавленный к уравнению, является нелинейным. В результате, используя описанную методологию, можно имитировать индекс VIX в разные промежутки времени и полученные результаты использовать в управлении финансовыми рисками (операции хеджирования через опционы) или принятии спекулятивных торговых позиций с опционами.

Alternative Methods for Combining Probability Boxes

Oleg Uzhga-Rebrov¹, Galina Kuleshova², ¹Rezekne Higher Educational Institution, ²Riga Technical University

Abstract – Probability boxes (p-boxes) are used as a tool for modeling uncertainty regarding probability distributions in the sets of relevant elements (random events, values of the random variable etc.). To combine information produced by two or more p-boxes, Dempster's rule for belief combination is commonly used. However, there are plenty of other rules for belief combination developed within the theory of evidence. The purpose of this paper is to present and analyze some widespread rules of that kind as well as examine their potentialities regarding combining the information provided by probability boxes.

Keywords – Basic probability assignments, belief combination rules, p-boxes, theory of evidence.

I. INTRODUCTION

Probabilistic evaluations represent occurrence chances of random events. Whenever initial data are missing or insufficient, the evaluation of relevant probabilities becomes quite difficult. If the evaluation is made by experts, the validity of the estimates obtained in principle cannot be evaluated a priori. To enable consideration of uncertainties related to probability evaluation, different techniques can be used. Common idea behind all those techniques is that instead of a single probability function, boundary probabilities are specified. It is supposed that a true probability function is between those boundary probability functions. However, situations are frequently possible when the accumulated distribution functions can only be constructed on different sets

of relevant elements. Situations of this kind occur when analyzing risks and/or safety of technical system operation. Among the reasons causing this kind of uncertainties, the following can be mentioned [2]:

- imprecisely defined probability distributions;
- ill-identifiable or even unknown correlations;
- essential measurement errors;
- impact of unrecognized factors on the model output;
- small sizes of samples;
- uncertainty of a model;
- non-stationarity (inconstant distributions).

To model this kind of uncertainties, the p-boxes technique was suggested. Theoretical foundations of the technique were first described in [1], [2]. The use of representation exploiting probability boxes in the structural analysis based on the method of finite elements is discussed in [8], whereas p-box application in the general context of risk assessment is examined in [3].

In what follows, to make the presentation simple and visual, it will be assumed that X is a set of real numbers, $\mathbf{R}$. This assumption in no extent will affect the commonness of material presentation.

The main idea of the p-box techniques is as follows. If, due to some reasons, it is not possible to construct a single probability distribution function on $\mathbf{R}$, boundary distribution functions similar to those depicted in Fig. 1 are constructed.

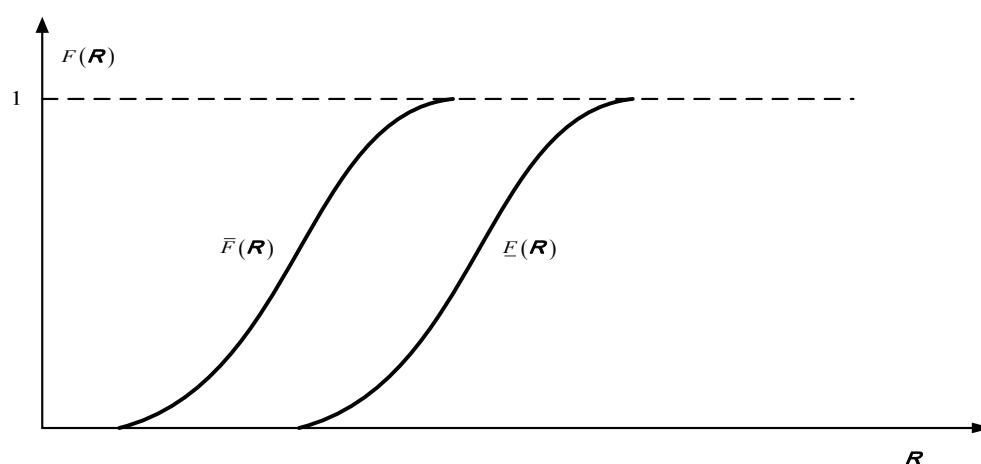


Fig. 1. Boundary probability distribution functions in the set of relevant values, $\mathbf{R}$.

It is stated that a real unknown distribution function $F(\mathbf{R})$ is in between those boundary functions. $\underline{F}(\mathbf{R})$ is the left boundary of the set of possible distribution functions,

which represents the values of the accumulated distribution function for $\mathbf{R} < r$. $\bar{F}(\mathbf{R})$ is the upper boundary of possible distribution functions. If a set of the lower probability values

on $\mathbf{R}$, $\underline{P}(R)$ is known, the boundary distribution functions can be expressed as follows:

$$\bar{F}(r) = 1 - \underline{P}(R > r); \quad (1)$$

$$\underline{F}(r) = \underline{P}(R \leq r). \quad (2)$$

From Fig. 1 it follows that the left boundary $\bar{F}(R)$ is the upper boundary for distribution function values and the lower boundary for value R . Instead, the right-hand boundary $\underline{F}(R)$ is the lower boundary for the values of distribution function and the upper boundary for value R .

The construction of probability boxes can be both parametric and non-parametric. The non-parametric construction is based on the assumption that the form and parameters of underlying probability distribution are unknown. The boundaries of the constructed p-box can have a deliberate form.

How could information provided by two or more probability boxes be aggregated? If boundary functions of distributions for combined probability boxes have a continuous form, such boxes should be discretized in advance. There are two techniques of probability box discretization. The essence of the first one, the so-called boundary discretization, is depicted in Fig. 2.

The second technique of discretization called discretization over mean points is illustrated in Fig. 3.

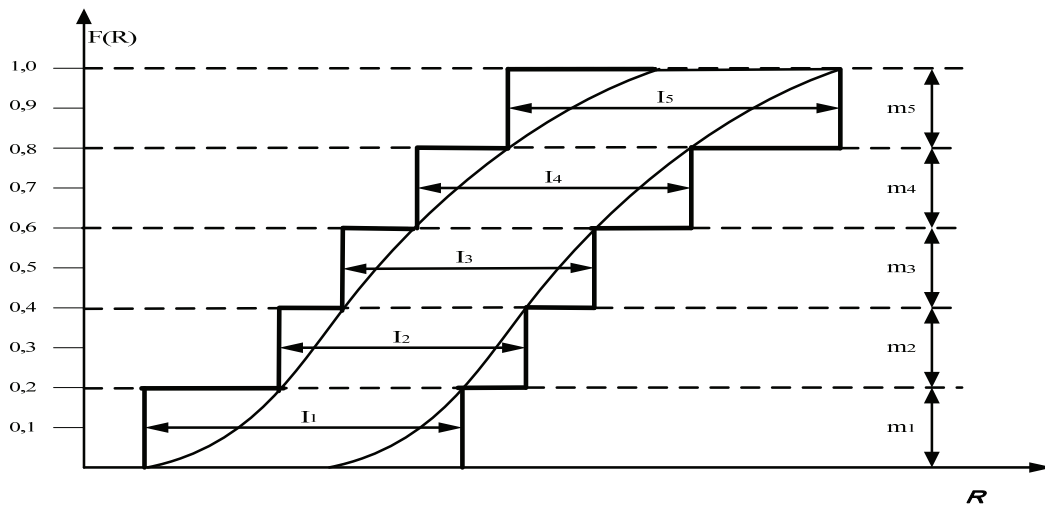


Fig. 2. Schematic representation of the method of boundary discretization of probability boxes.

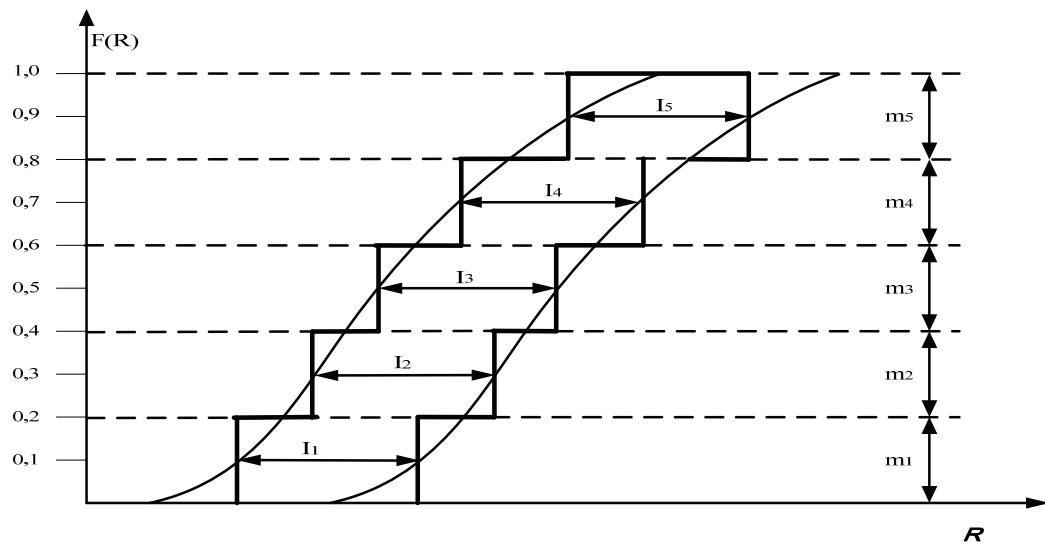


Fig. 3. Schematic representation of probability box discretization method based on mean points.

Let us assume that we have two stepwise or discretized probability boxes constructed on the basis of two independent information sources. The task is to aggregate information from

those sources. The next section considers basics of the theory of evidence and Dempster's belief combination rule that makes it possible to solve the task in a standard way.

II. BASICS OF THE THEORY OF EVIDENCE

The basics of the theory of evidence (Dempster-Shafer theory) were first described in [4]. A more detailed description of the theory can be found in Chapters 2 and 3 of [8]. Below, short data about the basic concepts and definitions of this theory are outlined, which are necessary for the explication of procedures for combining probability boxes.

One of the fundamental concepts of the theory is the concept of **frame of discernment** $\Omega = \{\omega_i / i = 1, \dots, n\}$. It is formed of the elements under consideration. The notion of elements can be treated quite widely depending on the context. Elements can be possible values of unknown variable, individuals who are suspected of making a crime, certain events etc. Only one element is true. Let us denote it as ω_0 and call it a real world.

The essence of Dempster-Shafer's theory is as follows. Based on the evidences available, subsets of elements $A \subseteq \Omega$ are determined; the subsets may contain a real world ω_0 . If a real value of some unknown variable is under consideration, the role of subsets will be played by respective intervals. Relevant subsets or intervals are called **focal elements**. Function $m: 2^\Omega \rightarrow [0, 1]$ that is called **basic probability assignment** can be correlated with the frame of discernment; the function satisfies these requirements:

$$m(\emptyset) = 0;$$

$$\sum_{A \subseteq \Omega} m(A) = 1 \text{ for all } A \subseteq \Omega.$$

The kernel of the theory of evidence is the concept of **belief function** $bel: 2^\Omega \rightarrow [0, 1]$ that meets the following requirements:

$$bel(\emptyset) = 0; \quad bel(\Omega) = 1;$$

For every integer n and every set $A_1, \dots, A_n$ of subsets Ω we have

$$bel(A_1 \cup \dots \cup A_n) = bel(A_1) - \sum_{i < j} bel(A_i \cap A_j) + \dots + (-1)^{n+1} bel(A_1 \cap \dots \cap A_n).$$

Belief functions are correlated with basic probability assignments in this way:

$$bel(A) = \sum_{B_i \subset A} m(B_i). \quad (3)$$

An important question in the Dempster-Shafer theory is the combination of beliefs produced by different evidences. Most widespread rule of belief combination is Dempster's rule [5]. Let on the basis of the first group of evidences there be assigned basic probability masses to certain subsets (focal elements) on the frame of discernment $A_i \subset \Omega, i = 1, \dots, m$. Assume that there is another group of evidences on whose basis there are assigned basic masses of probability to subsets $B_j \subset \Omega, j = 1, \dots, n$. The combined mass of probability corresponding to the overlapping of focal elements $A_i \cap B_j$ is expressed as follows:

$$m_{12}(A_i \cap B_j) = m_1\{A_i\} m_2\{B_j\}. \quad (4)$$

For any subset C , comprising any number of subsets $A_i \cap B_j$, the combined mass of probability can be calculated as a sum

$$\sum_{\substack{i,j \\ A_i \cap B_j \subset C}} m_1\{A_i\} m_2\{B_j\}. \quad (5)$$

Dempster's rule of combination foresees the normalization of the combined masses of probability corresponding to non-empty intersections of marginal focal elements. The value of the normalizing constant K is calculated by this expression:

$$K = \left[1 - \sum_{\substack{i,j \\ A_i \cap B_j \neq \emptyset}} m_1\{A_i\} m_2\{B_j\} \right]^{-1}. \quad (6)$$

Let us consider an example illustrating the combination of probability boxes on the basis of Dempster's rule (the idea is borrowed from [8]).

Example 1. Fig. 4 shows three probability boxes constructed on the basis of evaluations of experts A, B and C in the set of values of random variable π .

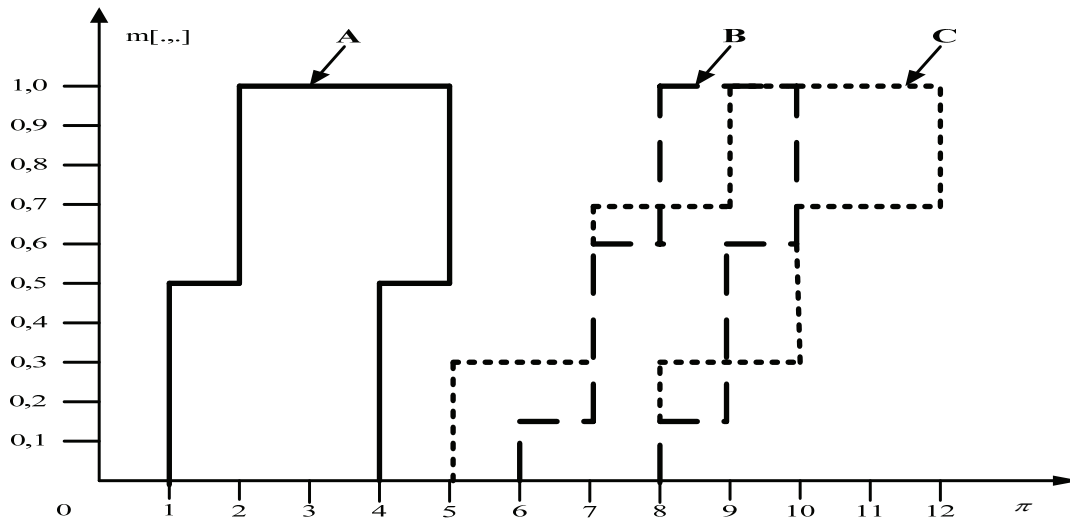


Fig. 4. Probability boxes constructed by three experts A, B and C.

Let us express the information represented by these boxes in the numeric form.

Expert A:	Expert B:	Expert C:
$[1, 4] \quad m_1 = 0.5;$	$[6, 8] \quad m_2 = 0.2;$	$[5, 8] \quad m_3 = 0.3;$
$[2, 5] \quad m_1 = 0.5.$	$[7, 9] \quad m_2 = 0.4;$	$[7, 10] \quad m_3 = 0.4;$
	$[8, 10] \quad m_2 = 0.4.$	$[9, 12] \quad m_3 = 0.3.$

The figures in square brackets represent intervals of probability boxes but the values $m(\cdot)$ are basic probability assignments related to the corresponding intervals.

It is obvious that the information provided by probability box A cannot be combined with the information provided by probability boxes B and C because Dempster's rule can only be used for overlapping probability boxes.

By combining basic probability assignments for the overlapping intervals of probability boxes B and C according to expression (5) and normalizing the results by expression (6) we get:

$$m_{BC}[6, 8] = 0.0857; m_{BC}[7, 8] = 0.2857; m_{BC}[7, 9] = 0.2286; m_{BC}[8, 10] = 0.2286; \\ m_{BC}[9, 10] = 0.1714.$$

The combined probability box is shown in Fig. 5. As can be seen in Fig. 5, the resulting probability box is narrower than the initial boxes. The reason for that is that Dempster's rule of combination takes into account only the overlapping parts of intervals of the initial probability boxes.

Dempster's rule of combination has a very strong underlying logical basis and is characterized by high conservatism. The result of combination only depends on the information, which is provided by both sources. The information provided by a separate source is not included in the result of combination and serves only for combination result normalization.

The shortcoming of Dempster's rule of combination is that it can produce unnatural results in certain specific conditions [6], [7].

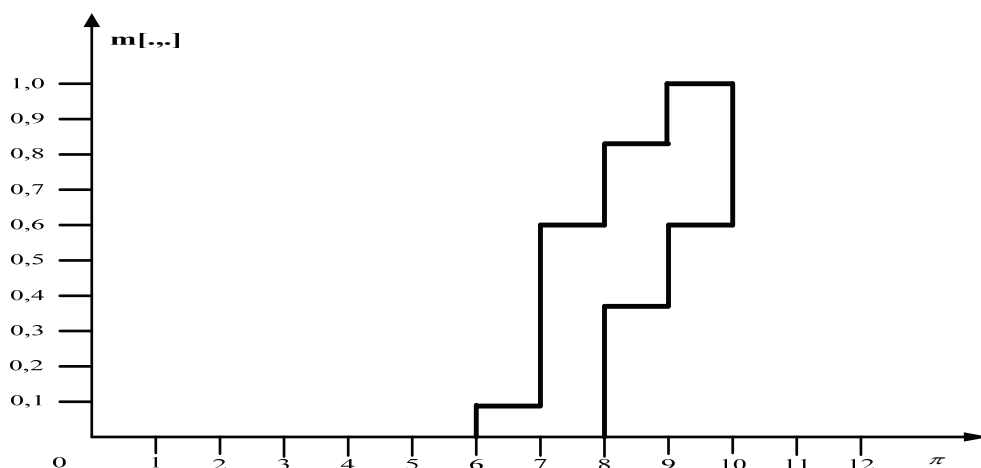


Fig. 5. Probability box obtained through the combination of probability boxes B and C according to Dempster's rule (see Fig. 4).

III. ALTERNATIVE METHODS FOR THE COMBINATION OF PROBABILITY BOXES

Nowadays, a large number of alternative rules for belief combination exist. More details on this kind of techniques can be found in [9]). This section considers most famous of them. One of the rules of this kind is Yager's rule of combination [10]. Yager has refused from the idea to ascribe all combined masses of probability related to empty overlapping of marginal focal elements to an empty set. Like in Dempster's rule, the value of each combined probability mass is calculated here as an orthogonal sum of the corresponding basic masses of probability (5). The normalization of the results is not foreseen. Though Yager's rule allows one to get rid of the main shortcoming of Dempster's rule; it possesses another shortcoming. It is impossible to construct a probability box on the basis of the results of combination because the sum of resulting masses is not equal to 1. Due to that reason, Yager's rule cannot be employed for combining probability boxes.

Dempster's rule of combination assumes as a basis overlapping marginal focal elements that are determined on the basis of different groups of evidences. This rule, however, does not consider the extent of such overlapping, whereas an alternative rule of combination, Zhang's rule [12], does take into account the extents of overlapping of the corresponding focal elements. If a subset C is the result of overlapping of subsets A and B , $C = A \cap B$, Zhang introduces the evaluation of the extent of overlapping of these subsets in this manner:

$$r(A, B) = \frac{|C|}{|A||B|} = \frac{|A \cap B|}{|A||B|}, \quad (7)$$

where $|A|$, $|B|$, $|A \cap B|$ are cardinalities of subsets A , B , $A \cap B$, respectively.

Evaluations (7) are calculated for each pair of the overlapping marginal focal elements. The value of the combined mass of probability related to a certain subset $C \subset \Omega$ is calculated as follows:

$$m(C) = K \sum_{A \cap B = C} \left[\frac{|C|}{|A||B|} m_1(A) m_2(B) \right], \quad (8)$$

where K is a normalizing constant.

Zhang's rule of combination also takes into account only the overlapping parts of intervals of probability boxes. The results of combination according to that rule slightly differ from those obtained using Dempster's rule. That difference is due to the principles of combination underlying both rules.

Zhang's rule can be recommended for the combination of probability boxes in cases when due to some reasons it is necessary to take into account the extent of overlapping of initial boxes.

There are also rules of combination developed that are based on one or another kind of averaging of the initial masses of probability. One of widespread rules of this kind is called **ρ -averaging**. The idea of that averaging is quite simple. If, based on the n groups of evidences basic probability masses $m_i(A)$, $i = 1, \dots, n$, are assigned to a subset $A \subset \Omega$, then the combined mass of probability ascribed to that subset is calculated by expression

$$m_{1,\dots,n}(A) = \frac{1}{n} \sum_{i=1}^n w_i m_i(A), \quad (9)$$

where w_i is a coefficient (weight) characterizing the extent of reliability of the i -th group of evidences. In particular case the averaging can be made without introducing coefficients (assuming the values of all w_i equal 1).

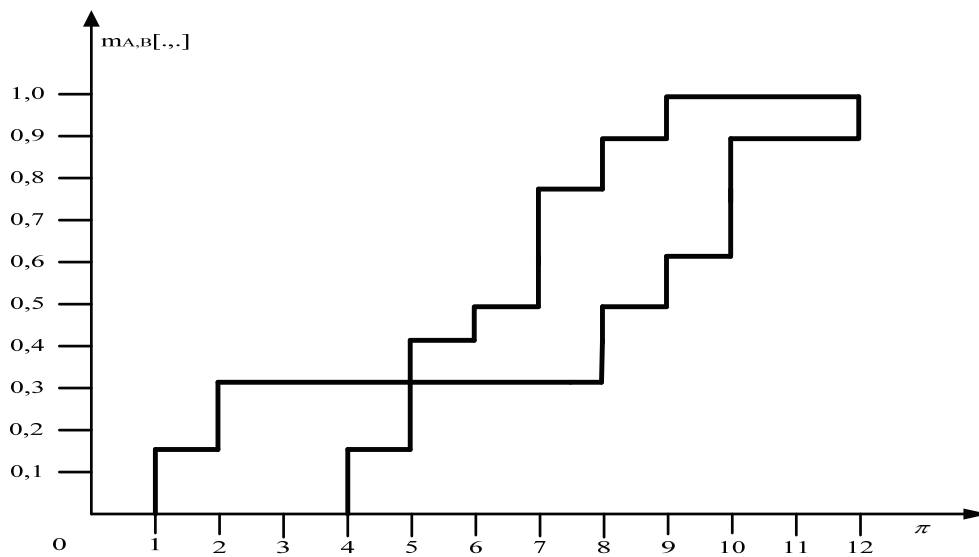


Fig. 6. A probability box obtained as a result of combination of three initial probability boxes in Example 1 on the basis of the rule of ρ -averaging.

Applying the rule of ρ -averaging to all initial probability boxes in Example 1, we get the following result of combination:

$$m_{A,B,C}[1, 4] = 1/3 * 0.5 = 0.167; \quad m_{A,B,C}[2, 5] = 1/3 * 0.5 = 0.167; \quad m_{A,B,C}[5, 8] = 1/3 * 0.3 = 0.100;$$

$$m_{A,B,C}[6, 8] = 1/3 * 0.2 = 0.067; \quad m_{A,B,C}[7, 9] = 1/3 * 0.4 = 0.133; \quad m_{A,B,C}[7, 10] = 1/3 * 0.4 = 0.133;$$

$$m_{A,B,C}[8, 10] = 1/3 * 0.4 = 0.133;$$

$$m_{A,B,C}[9, 12] = 1/3 * 0.3 = 0.100.$$

The resulting probability box is shown in Fig. 6.

It is clear that the rule of ρ -averaging allows combining non-overlapping probability boxes, which might be quite important in certain specific cases.

As opposed to the rule of ρ -averaging that averages probability masses by initial intervals, under ***x-convolving averaging*** specific combination of the initial relevant intervals is performed. The boundaries of each resulting interval are

determined as the mean values of the boundaries of the initial intervals.

The combined values of probability masses ascribed to each resulting interval are calculated in a standard manner as the multiplications of the marginal probability masses assigned to the initial intervals.

An essential feature of that rule of combination and also of the rule of ρ -averaging is that it forms and takes into account the resulting intervals even in the cases when the initial intervals do not overlap.

Applying the rule of ***x-convolving averaging*** to probability boxes A and B in Example 1, we get the following results of combination:

$$m_{AB}[3,5, 6] = 0.10; \quad m_{AB}[4, 6,5] = 0.30;$$

$$m_{AB}[4,5, 7] = 0.40; \quad m_{AB}[5, 7,5] = 0.20.$$

The resulting probability box is shown in Fig. 7.

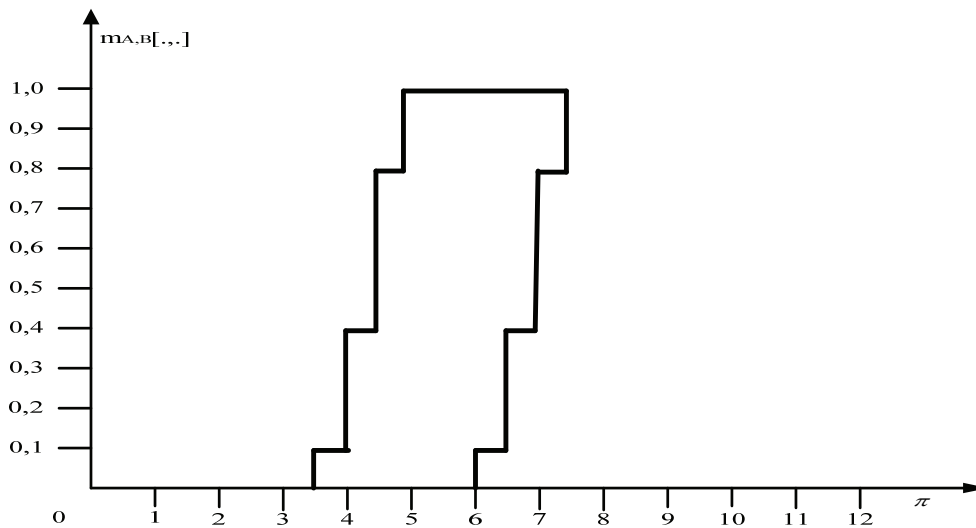


Fig. 7. Probability box obtained as a result of combining initial probability boxes A and B in Example 1 on the basis of the rule of *x-convolving averaging*.

A specific rule of belief combination is the disjunctive rule of Dubois and Prade. That rule is similar to the rule of ***x-convolving averaging***, the only difference being in the determination of the boundaries of the resulting intervals. In the rule of ***x-convolving averaging***, each of the boundaries is determined as the mean of the corresponding boundaries of the initial intervals. In the given rule, the lower boundary of the resulting interval is equal to the minimal value of the lower boundaries of the initial intervals while the upper boundary is

equal to the maximum value of the upper boundaries of the initial intervals.

Let us combine the initial probability boxes A and B in Example 1 by the disjunctive rule of Dubois and Prade. As a result, we have

$$m_{AB}[1,8] = 0.10; \quad m_{AB}[1,9] = 0.20; \quad m_{AB}[1,10] = 0.20;$$

$$m_{AB}[2,8] = 0.10; \quad m_{AB}[2,9] = 0.20; \quad m_{AB}[2,10] = 0.20.$$

The resulting probability box is depicted in Fig. 8.

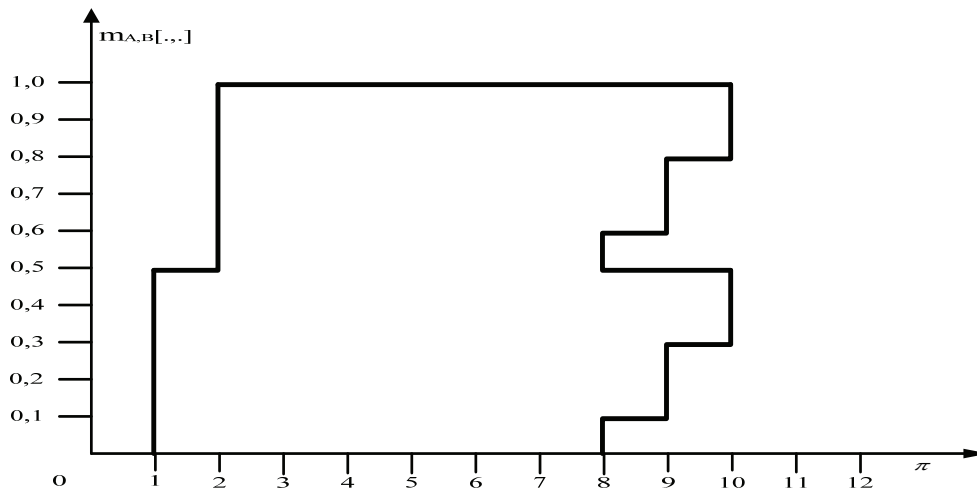


Fig. 8. A probability box obtained as a result of combining initial probability boxes A and B in Example 1 on the basis of the Dubois and Prade disjunctive rule of combination.

IV. CONCLUSION

When developing the theory of probability boxes for combining information provided by two or more probability boxes, the use of Dempster's rule of belief combination was suggested. This suggestion is probably based on the wide reputation and applicability of rule. However, plenty of other rules for combining beliefs are proposed. Most widespread rules of this kind are briefly described in this paper.

All the above-mentioned rules of belief combination except for Yager's rule can be in principle used for combining probability boxes.

It should be noted that all those rules can be divided into three groups: (1) rules working only on the overlapping of intervals that form probability boxes; (2) rules based on the averaging of initial probability masses for initial intervals and (3) rules based on the specific formation of resulting intervals on the basis of the initial intervals. Combination rules of Dempster and Zhang fall to the first group. Zhang's rule only differs from Dempster's rule in that it takes into account the extent of overlapping of the initial intervals when calculating resulting probability masses. Both these rules can be used in the cases when a high confidence of the resulting evaluations is required. In general case, preference has to be given to Dempster's rule as it has a more general character and is simpler from the computational point of view.

The rule of p -averaging is ascribed to the second group and has a universal character. It takes into account all initial probability masses for relevant intervals. The advantage of the method is the possibility of accounting the extent of confidence for different groups of evidences (initial probability boxes).

The next two methods are labeled to the third group. An essential feature of the rule of x -convolving averaging is the specific creation of the resulting intervals, which are the result of averaging of the initial intervals. However, under such a combination, the uncertainty of the resulting evaluations may be higher than that of the initial evaluations, which essentially impedes the interpretation of the results obtained and makes the deduction of validated conclusions quite problematic. That is why the given rule can be recommended for practical use

only in those cases, when it is desired to use considerable amount of the initial information not taking into account its potentially contradictory character.

These conclusions are even more relevant for the Dubois and Prade disjunctive rule of combination. The advantage of the rule is that it makes use of the whole initial information. The shortcoming of the method is a high uncertainty of the results. Due to that, the rule can only be used in exceptional cases.

A general conclusion that can be made on the basis of this paper is – when combining probability boxes researchers should not restrict themselves to Dempster's rule of combination; instead, alternative rules of combination have to be widely used. The choice of proper rule of combination has to be dictated by conditions of a specific task and specific requirements to the results.

REFERENCES

- [1] K. Sentz and S. Ferson, "Combination of evidence in Dempster-Shafer theory," *Sandia National Laboratories*, vol. 4015, 2002. <http://dx.doi.org/10.2172/800792>
- [2] S. Ferson, V. Kreinovich, L. Ginzburg, D. Myers and U. Sentz, "Constructing Probability Boxes and Dempster-Shafer Structures", SAND Report, Sandia National Laboratories, Albuquerque, New Mexico, California, 2003.
- [3] V. Montgomery, "New Statistical Methods in Risk Assessment by Probability Bounds", Ph. D. thesis, Durham University, UK, 2009.
- [4] G. Shafer, *A mathematical theory of evidence*, Princeton University Press, 1976.
- [5] A. P. Dempster, "Upper and lower probabilities induced by a multivalued mapping", *Annals of Mathematical Statistics*, 38, pp. 325–339, 1967. <http://dx.doi.org/10.1214/aoms/1177698950>
- [6] W. Liu and J. Hong, "Re-investigating Dempster's Idea on Evidence Combination," *Knowledge and Information Systems*, vol. 2, no. 2. [Online]. Available: <http://www.cs.qub.ac.uk/~Liu/kis00>, 2000. [Accessed May 6, 2014].
- [7] D. Dubois and H. Prade, "Focusing versus updating in belief function theory," *Advances in the Dempster-Shafer Theory of Evidence*. R. R. Yager, J. Kacprzyk, M. Fidirizzi, Eds., New York, John Wiley&Sons, Inc., pp. 71–95, 1994.
- [8] O. I. Uzhga-Rebrov, *Upravlenije neopredelennostjami. Chastj 3. Sovremennije verovatnostnije metodi*. Rēzekne, RA Izdevniecība, 2010. (In Russian).
- [9] O. I. Uzhga-Rebrov, *Upravlenije neopredelennostjami. Chastj 4. Kombinirovanije neopredelennostej*, Rēzekne, RA Izdevniecība, 2014. (In Russian).

- [10] R. Yager, "On the Dempster-Shafer Framework and New Combination Rules", *Information Sciences*, vol. 41, no. 2, pp.93–137, 1987. [http://dx.doi.org/10.1016/0020-0255\(87\)90007-7](http://dx.doi.org/10.1016/0020-0255(87)90007-7)
- [11] H. Zhang, R. L. Mullen and R. L. Muhanna, "Finite Element Structural Analysis Using Imprecise Probabilities Based on P-Box Representation," *4th International Workshop on Reliable Engineering Computing*, REC 2010, Singapore, pp. 211–225, 2010.
- [12] L. Zhang, "Representation, independence and combination of evidence in the Dempster-Shafer Theory of Evidence", *Advances infig*
- [13] *the Dempster-Shafer Theory of Evidence*. R. R. Yager, J. Kacprzyk, M. Fedrizzi, Eds, New York, John Wiley&Sons, Inc., pp. 51–69, 1994.

Oleg Uzhga-Rebrov is Professor at the Faculty of Economics, Rezekne Higher Education Institution (Latvia). He received his Doctor's degree in Information Systems from Riga Technical University in 1994. His research interests include different approaches to processing incomplete, uncertain and fuzzy information, in particular, fuzzy set theory, rough set theory as well as

fuzzy classification and fuzzy clustering techniques and their applications in bioinformatics.

Address: Rezekne Higher Education Institution, 90 Atbrivoshanas Av., Rezekne LV-4600, Latvia.

E-mail: ushga@ru.lv.

Galina Kuleshova is a Researcher with the Faculty of Computer Science and Information Technology, Riga Technical University (Latvia). She received the M.Sc. degree in *Decision Support Systems* from Riga Technical University. Current research interests include artificial neural networks, data mining, ontology engineering, classification methods and bioinformatics.

Address: Institute of Information Technology, Riga Technical University, 1 Kalku Str., Riga LV-1658, Latvia.

E-mail: galina.kulesova@cs.rtu.lv.

Olegs Užga-Rebrovs, Galina Kuļešova. Alternatīvās metodes varbūtības kastu kombinēšanai

Liela daļa lietišķo uzdevumu varbūtības lielumu novērtēšanai tiek veikta uz ekspertu subjektīvo spriedumu pamata. Tā kā ekspertiem bieži ir grūtības ar vienoziģu relevanto varbūtību novērtējumu, tiek izstrādātas pieejas, kuras ļauj modelēt papildu nenoteiktības, kas ir saistītas ar ekspertu varbūtības novērtējumiem. Tāda veida zināmās pieejas ir nenoteiktās varbūtības un izplūdušās varbūtības. Ar varbūtību kastu palīdzību tiek modelētas nenoteiktības, kas ir saistītas ar varbūtību sadalījuma ekspertu novērtējumu. Ja eksperts nav pārliecināts, ka novērtējama sadalījums ir ticams, viena varbūtību sadalījuma vietā viņš uzdod divus varbūtību sadalījumus, kuri veido varbūtību kasti. Varbūtību sadalījumiem, kuri veido varbūtību kasti, var būt dažāda forma. Sadalījumi var būt arī nesimetriski. Problēmas rodas tad, kad jākombinē informācija, kas tiek piedāvāta ar divām vai vairākām varbūtību kastēm, kuras konstruē neatkarīgi eksperti. Oriģinālos darbos par varbūtību kastēm šim mērķim tiek piedāvāts izmantot Dempstera pārliecību kombinēšanas likumu. Metodes trūkums – nepieciešamība normēt rezultējošās pārliecības. Šī raksta mērķis ir parādīt iespējas izmantot alternatīvas metodes informācijas kombinēšanai, kura tiek iegūta no divām vai vairākām varbūtības kastēm. Šīs metodes var pielietot tieši, ja varbūtību koki tiek ierobežoti ar diskrētiem varbūtību sadalījumiem. Nepārtraukto varbūtību sadalījumu gadījumā sākotnējiem sadalījumiem jābūt diskretizētiem pirms kombinēšanas operāciju izpildes. Darbā ir minētas divas varbūtības koku diskretizācijas metodes: robeždiskretizācija un diskretizācija vidējos punktos. Varbūtības koku kombinēšanas alternatīvo metožu esamība ļauj izvēlēties piemērotāko metodi konkrēta uzdevuma kontekstā.

Олег Ужга-Ребров, Галина Кулешова. Альтернативные методы комбинирования вероятностных ящиков

Назначение вероятностных оценок в большом числе прикладных задач производится на основе субъективных мнений экспертов. Поскольку эксперты часто затрудняются дать однозначные оценки для релевантных вероятностей, разработаны подходы, позволяющие моделировать дополнительные неопределённости, связанные с экспертным назначением вероятностей. Наиболее известными подходами такого рода являются неопределённые вероятности и нечёткие вероятности. С помощью вероятностных ящиков моделируются неопределённости, связанные с экспертным конструированием вероятностных распределений. Если эксперт не уверен в достоверности оцениваемого вероятностного распределения, вместо одного вероятностного распределения он задаёт два граничных вероятностных распределения, которые и образуют вероятностный ящик. Вероятностные распределения, образующие ящик, могут иметь самую разнообразную форму и не быть симметричными. Проблемы возникают в том случае, когда нужно скомбинировать информацию, даваемую двумя или более вероятностными ящиками, сконструированными независимыми экспертами. В оригинальных работах по вероятностным ящикам для этой цели предлагается использовать правило комбинирования уверенностей Dempstera. Недостатком этого метода следует признать необходимость нормирования результирующих уверенностей. Цель настоящей статьи – показать возможности применения альтернативных методов комбинирования уверенностей для комбинирования информации, даваемой двумя или более вероятностными ящиками. Эти методы непосредственно применимы в случаях, когда вероятностные ящики ограничены дискретными вероятностными распределениями. В случае непрерывных вероятностных распределений эти распределения должны быть дискретизированы до выполнения операций комбинирования. В работе представлены два метода дискретизации вероятностных ящиков: метод граничной дискретизации и метод дискретизации на средних точках. Наличие альтернативных методов комбинирования вероятностных ящиков позволяет выбрать наиболее подходящий метод в контексте конкретной задачи.

Markov Chains in the Task of Author's Writing Style Profile Construction

Pavels Osipovs¹, Andrejs Rinkevics², Galina Kuleshova³, Arkady Borisov⁴,¹⁻⁴ *Riga Technical University*

Abstract – This paper examines the possibility of using Markov chains when constructing a profile of author's writing style. Thus, the constructed profile can be then used to analyze other texts and calculate their level of similarity. The extraction of the unique profile of text writing style that is characteristic of a specific human can be a topical task in many spheres of human activity. As an example, the task of detecting authorship for scientific and fiction texts can be mentioned. The paper describes a basic theoretical apparatus used for profile construction, software implementation of the experimental system as well as the experiments made and provides experimental results and their analysis.

Keywords – Formalization of author's writing style, level of texts similarity, Markov chain.

I. INTRODUCTION

Nowadays the task of accurate determination of text authorship has not been solved with the accuracy required to guarantee correct identification yet. Studies can be mentioned that show up to 80 % correct results. Since the topicality of this kind of task is constantly growing, comparatively many studies in the field are conducted [3].

Different specialists are interested in finding a solution to the task above: literary critics, mathematicians, philologists, lawyers, criminalists, historians etc. [10]. To identify the authorship of a text, inquiries are frequently sent to experts, literary critics or historians who are able to identify the author of the unknown text or determine authorship by characteristic language peculiarities and stylistic techniques. Previously, to identify plagiarism and determine authorship of a literary work, handwriting expertise was used, which is completely unaccepted for the printed text. Automatic identification of the authorship makes it possible to get rid of some essential drawbacks observed in identifying authorship by experts such as time consumption and biases of expert point of view.

Various researchers use different approaches to solve the task of author style detection, e. g.,

- fuzzy area based methods have shown good results [4] on test data sets;
- top-k elements approach [5];
- the ability to use classical statistical approaches in the area under consideration has been studied in [6];
- support vector machine approach has been used in text classification for authorship attribution analysis [7];
- probabilistic approach has been used by Microsoft researchers to unite text classifiers under common meta-class for better effectiveness [8].

This research is focused on studying the possibility of using Markov chains in the task of text authorship identification [9]. The technique used in the study is based on the ideas suggested in paper [1] that discusses the system of anomalous action detection using Markov chains. By using this kind of approach, a model of author's writing style is created, which represents a Markov chain. Chain nodes are an unchanged sequence of words of the specified quantity that is picked up from the text. The weight of the edge characterizes the probability of prolonging the sequence of words of the node, in other words, the connected node of the given arc.

II. DESCRIPTION OF THE TECHNIQUE

This study is aimed at checking the possibility of constructing a model of author's writing style using Markov chains.

General scheme of the implemented technique is shown in Fig. 1.



Fig. 1. General scheme of the technique.

A model of author's writing style is constructed using the texts for which the authorship of the targeted author is known for sure. The model is then used to analyze another text, whose authorship is unknown.

The final result will be Boolean value: **True** or **False**, corresponding, respectively, whether the text under consideration belongs to the target author or not.

A. Peculiarities of Subject Domain

It is worth noting that in some cases an author changes his style as a specific literary technique. Also modern authors make use of assistance of other people, who write some pieces of their works.

The volume of the analyzed text and texts used to train the model is also important. It has to be large enough to contain plenty of information about the specifics of style of its author. It must be stated that as of today texts of small volume are difficult to classify. There is empirically discovered lower limit of text size that contains enough information – it is

26000 words. It is clear that the size differs depending on the author.

Another special feature of the domain of the task under consideration is translated texts. In general case, it is highly not recommendable to analyze a text in the original language. Maximum that can be done is to compare texts translated by a single translator because literary translation is a complicated task and in the process of translation the author contributes to the text parts of his own author's style.

B. Model Training

Model training is conducted on the basis of all the texts collected for which there is a confidence in their authorship. Over each text an analyzer is passed, and by applying to it the iterative procedure a summary graph of the author's style is constructed. In what follows, the procedure is described in more detail.

C. Model Use

The use of the model consists in applying the already trained model to the analysis of the target text. In addition, an iterative procedure is explored that calculates a consequence of values of level metrics of the authorship for a large number of text pieces. The overall result will be the summary mean value of metrics for the whole text. It is also possible to use characteristics that differ from a simple mean value.

The procedure of using the existing model to analyze a new text is described in more detail below.

III. MODEL TRAINING

The construction of the model of author's style includes the following steps:

1. Collection of original texts used to train the model;
2. Pre-treatment of the collected texts;
3. Pre-treatment setting configuration:
 - a. Should punctuation be removed?
 - b. Specify the size of words that will be named as "short".
 - c. Should "short" words be deleted?
 - d. Should the word form be normalized?
4. Specification of the parameter of model learning, w – window size (number of words per node of the graph).
5. Directing the learning process based on the application requirements.

When choosing a window size w , there is a dilemma: its small values contain too little information, but a big size too precisely adapts to the training set (the classic problem of over-education).

The process of navigating a text consists in an iterative application of several operations.

At the beginning of the creation of the model, a window is initialized to an empty set of symbols – $\emptyset$. The initial value of the window w is specified.

Two operations are defined on paths:

- $shift(\sigma, x)$ which shifts the trace σ to the left and adds atomic element x at the end of the track, for example:

$shift(aba, c) = bac$.

- $next(\sigma)$ returns the first character of a trace σ and moves one position to the left, for example, $next(abcd) = a$ and updates path to the state bcd .

The initial state of the Markov chain is defined as the trace with length = w , consisting of null characters. For example, if the window size is $w = 3$, the initial state will have the form $[\emptyset, \emptyset, \emptyset]$.

The process of building the Markov chain consists of the following steps that are iteratively repeated. To each track of the current set these operations are applied:

- let $c = next(\sigma)$;
- setting $next\ state = shift(current\ state, c)$;
- increase counter for state $current\ state$ and arc between $current\ state$ and $next\ state$;
- update $current\ state$ to having value $next\ state$.

That is, at each iteration two counter values are formed: the value for the $current\ state$ and the value for the transition from the $current\ state$ to the $next\ state$. There is also the set or updated value of the transition probability for the transition between the $current\ state$ and $next\ state$ nodes.

As a result, when the operations described above are applied to all of the analyzed text, it will create a graph containing all present in it combination of words of a given length, their frequency of use and communication with other combinations.

IV. MODEL USE

By the term "model using" we mean its application in the analysis of a new text. The final result will be the use of the vector identity values for each level of the anomalous part of the target text size w .

The process of calculating each metric anomaly is to perform a sequence of operations described below.

At the beginning of the analysis, additional global variables X and Y are introduced, which are used throughout all analysis iterations.

Initially, the variables X and Y are equal to 0, the next step changes the current state by adding an atomic element to the end and removing the initial description of the current state.

On the basis of the previously developed Markov chain that contains an author's style template for each transition between the atomic elements of the text written by an unknown author, metric $\mu(a)$ is calculated, which denotes status of the current value of the metric and new values for variables X and Y .

At each step, there are two ways of calculating the value of the metric by the following algorithm. If the current model graph contains transition arc from the $previous\ state$ to the $new\ state$ $\beta_i \rightarrow \beta_{i+1}$, then X and Y are updated using the following function-parameters:

$$Y = Y + F(s, (s, s'));$$

$$X = X + G(s, (s, s'));$$

Else, if in the current model graph an arc from the $previous\ state$ to the $current\ state$ is not presented, then these functions-parameters are used:

$$Y = Y + Z;$$

$X = X + 1$.

And finally, the value of metric $\mu(a)$ is calculated, which is equal to Y/X .

Metric $\mu(a)$ shows how well the Markov chain predicts trace a , that is, the smaller its value is the better the model predicts the author's style.

Since μ is parameterized by functions F , G and the number Z , a different selection of F and G will affect the final value of the classifier, which adds the ability to customize the fine specifics of the problem domain.

A. Metrics of the Difference

Functions F and G can be implemented in different ways, depending on the characteristics of the analyzed text. Nowadays, there may be used the following approaches:

- Probabilistic metric;
- Local minimal entropy metric;
- Frequency-based metric.

The summary results do not differ very strongly, but depending on the characteristics of the analyzed text, the best results can be shown by different metrics.

V. EXPERIMENTAL SOFTWARE

For the experiments, a software package was designed that allows proceeding of the analysis of the texts with different kinds of settings.

As the main programming language in the development of experimental software system Python [11], [12] was used, as a

universal tool having readable syntax, non-strict typing and often used in scientific research.

Since for large texts the final graph of the author's style may have a very large number of nodes and interconnection arcs, a 64-bit platform was used to store such big data structures in RAM. The established platform of experiments consists of several main software modules.

For easy experiment management and evaluation, a graphic user interface has been created that allows managing all experiment settings, saving/loading and visualizing experiments results (see Fig. 2).

A. A Module for Text Pre-processing

The main task of this module is to pre-process the text.

Global constants defined in the module:

- `SHORT_WORD_LENGTH` – the number that characterizes the length of the word, which will be declared as a small or insignificant word. It is assumed that most interjections, unions and other insignificant words fall under this characterization.
- `PUNCTUATION` – a character set defined as punctuation.
- `REGEX_PUNCTUATION` – a compiled set of punctuation marks for use in regular expressions.

There are two classes presented in the module:

- `SplitterFlags` – a class containing all the flags that define details and methods for text pre-processing;
- `Splitter` – base class for text pre-processing.

The overall structure of the module is shown in Fig. 3.

Fig. 2. Graphic user interface for experiment management.

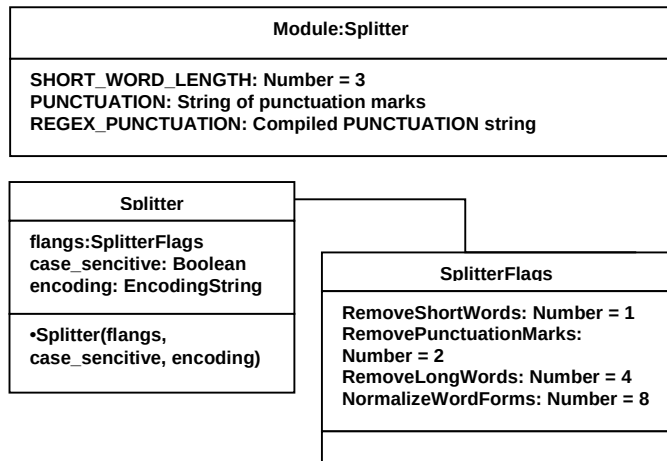


Fig. 3. The structure of the text pre-processing module.

B. Module to Create and Use Models of the Author's Style

The main module that combines all the functionality is “Similarity” module; it includes mechanisms for the construction and use of models of the author's style, based on a directed graph represented as a Markov chain.

The overall structure of the module is shown in Fig. 4.

The module includes the following classes:

- SimilarityMatcherFactory – a base class, which includes all necessary for work functions;

- ProbabilitySimilarity – a class that extends the base class and calculates the difference in author's style of the unknown author's text and current style model, on the basis of probability metric;

- LocalEntropyReductionSimilarity – a class that extends the base class and calculates the difference in author's style of the unknown author's text and current style model on the basis of local entropy metric.

- MissRateSimilarity – a class that extends the base class and calculates the difference in author's style of the unknown author's text and current style model on the basis of frequency-based metric.

- GraphSimilarity – a class that extends the base class and calculates the difference in author's style of the unknown author's text and current style model on the basis of graph similarity based metric.

C. Module for Directed Graph Creation and Manipulation

The module for graph manipulation is used to create and store the model of the author's style. Additionally, the module has an ability to save the graph in GraphML [2] file format.

The module contains two classes:

- WindowNode – this class is a basic data type representing the node of the graph, which includes all the necessary information in order to represent the graph as a Markov chain;
- WindowGraph – this class is used to store all the nodes of the graph and adding new ones.

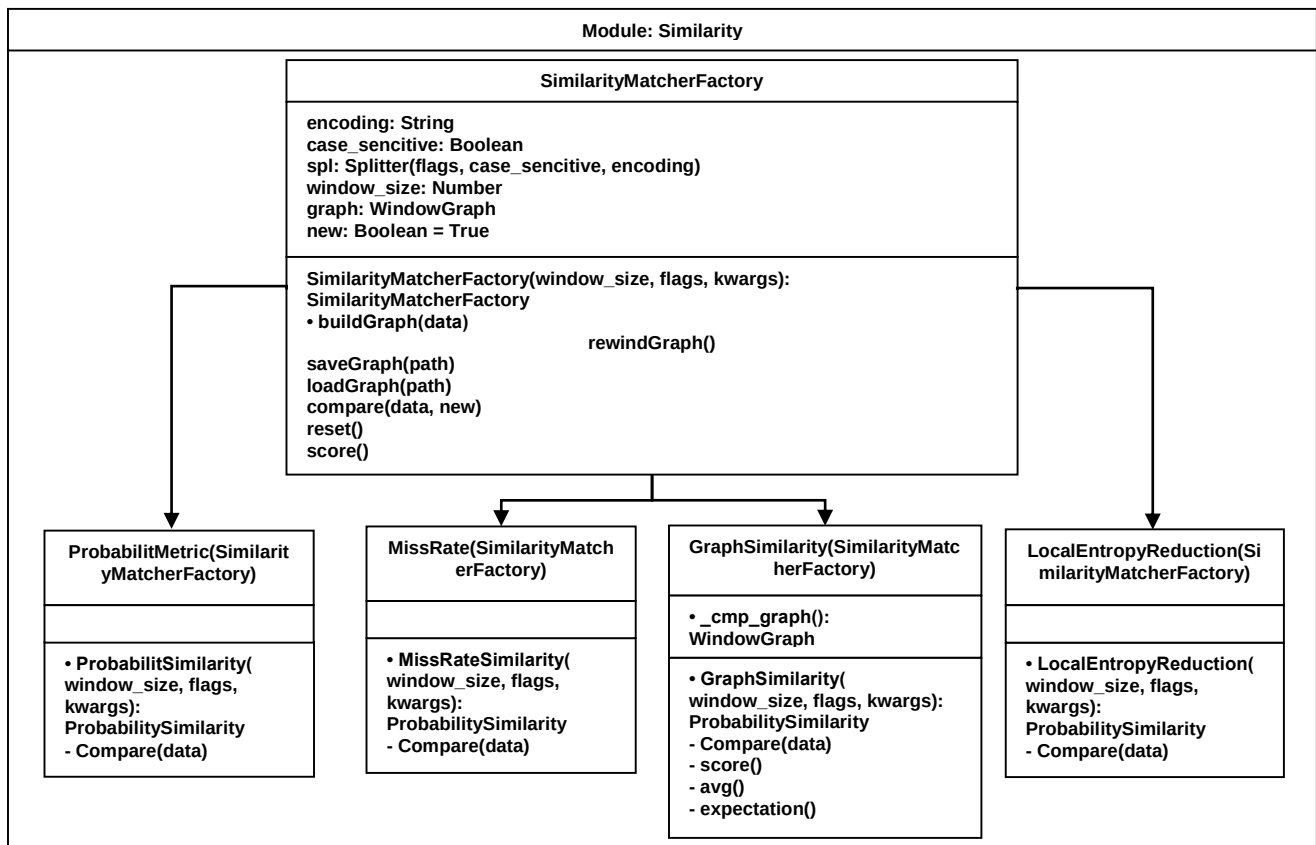


Fig. 4. The program structure of the module to create and use models of the author's style.

The overall structure of the module is shown in Fig. 6.

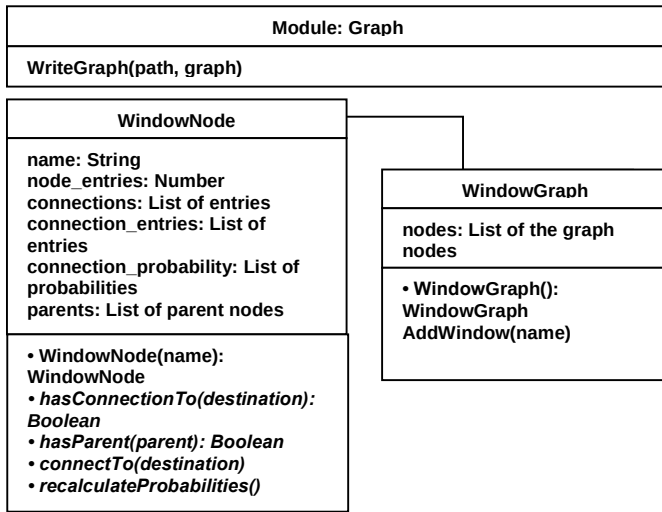


Fig. 5. The program structure of the module creation of a directed graph.

D. The General Scheme of the Component Relationship

The main component that connects all the modules is “Similarity”. By importing this class, child classes get access to all the necessary functionality to create a model of the author’s style.

The block diagram that shows the interconnection of components is shown in Fig. 6.

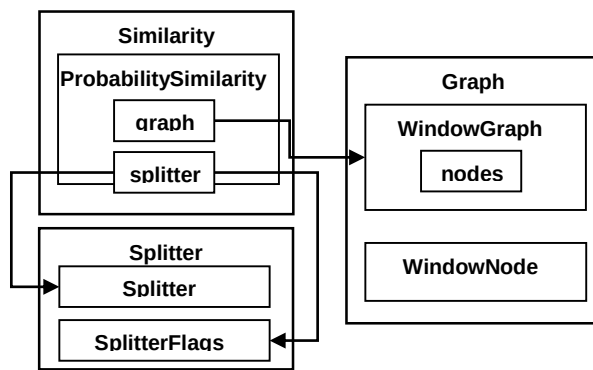


Fig. 6. Interconnections of the main components.

VI. EXPERIMENTS AND RESULTS

Experiments were carried out on the basis of the works of 10 authors, each of which presented from 10 to 40 texts. The number of words in each of the texts used was more than 26000.

The basic process of the experiments was accomplished as follows:

1. Cross-comparisons to establish the authorship were performed. It was assumed that there were sufficiently long fragments belonging to a number of authors that use

phonological writing. For the works of each author, the model of the author’s style was built.

2. At the next step of the analysis, a text with a known author was selected, but the author was marked as unknown.
3. Upon receiving the result, it was compared with the expected one that characterized the accuracy of the determination of authorship.

A. Comparison of the Effectiveness for Different Values of the Window Size

Here, the results of text classification using different values of the window size are compared.

As a result of measurement, it was ascertained that the length of the windows affects the performance. If the lowest possible window length equal to one atomic element is applied, modeling takes a lot of time. When the dimension of the window is increased, there is a growth rate of construction of models of author’s style. This is due to the lessening of linkages between nodes, as the same windows in the construction are rarely observed.

The parameters of one of the conducted experiments are given in Table I.

TABLE I
EXPERIMENT PLAN

Parameter	Value/State
Window size (W)	1, 2, 3, 5, 10, 25, 50, 100
Z	1.5
G	1.0
Pre-processing	No
Case normalization	Yes
Metric used	Probability metric

According to the results of the experiment, the method of step-by-step metric was able to determine the authorship with different sizes of windows in 60–80 % of the total work. The accuracy of author detection with different values of window size is illustrated in Fig. 7.

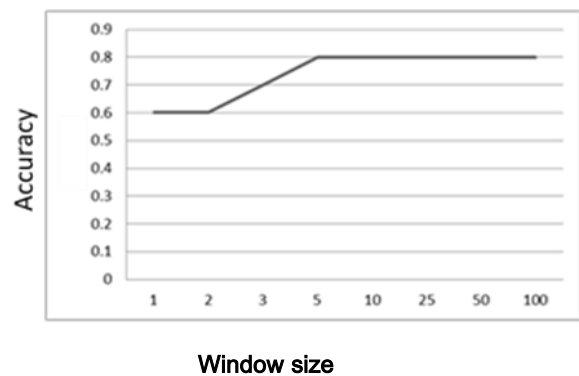


Fig. 7. The success rate for different window sizes.

The impact of different types of pre-processing on the effectiveness of text classification has been studied

experimentally. This experiment is based on the results of an experiment investigating the dependence of recognition on the dimension of the window size of the author's style model. The continuation of the study foresees to elucidate the effect of pre-processing of text on final classification results. The main objective of this experiment is to increase the detection of stylistic factors, accumulating only the information required in the model of the author's style.

The applied variants of text pre-processing are as follows:

1. remove short (irrelevant) words;
2. remove punctuation;
3. remove short (insignificant), the words and punctuation marks;
4. remove long words;
5. normalize word forms;
6. remove long words and punctuation marks;
7. case-sensitivity, remove punctuation;
8. case-sensitivity, remove short (irrelevant) words;
9. case-sensitivity, remove short (insignificant), the words and punctuation marks;
10. case-sensitivity, remove long words.

The chart with experimental results is shown in Fig. 8.

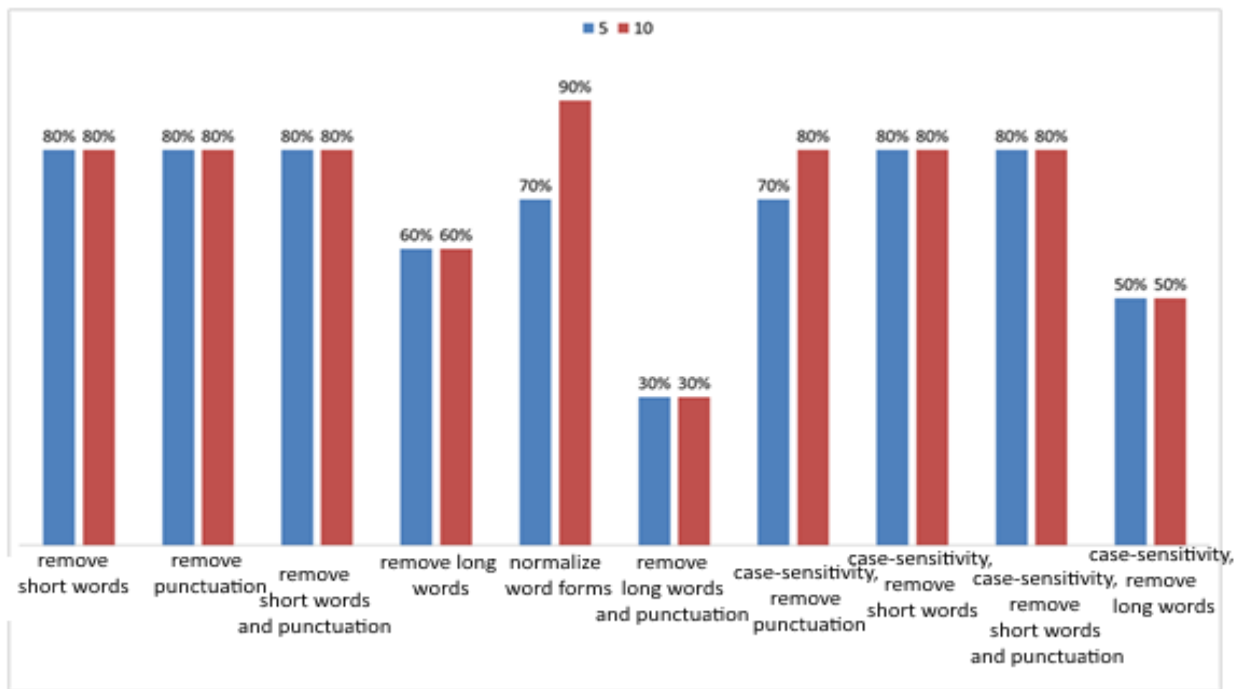


Fig. 8. Results for different types of text normalization.

According to the results of the experiment, it can be assumed that short words and punctuation marks do not affect the final classification effectiveness. By using the prior removal of short (insignificant) words, together with the removal of punctuation mark, or separately, the total recognition of the author's works does not change and is held at the level of 8 recognized works out of 10.

In turn, in the case of prior removal of long words, a lot of information about the author's style is lost. The accuracy of

determination falls by 20 – 30%. This observation confirms the results of experiments 4, 6, and 10. Together with the removal of punctuation, the recognizability drops to 3 words out of 10, which is the worst indicator of all the experimentally verifiable changes. Presumably, long words and punctuation accumulate most information about the author's style.

As a result, we can point to the possibility of a pre-removal of short words and punctuation marks, as they do not affect the accuracy of the determination. This modification will increase the processing speed and reduce memory consumption by the construction of the graph. Supposedly, one can use the normalization of the word forms to improve recognizability. But this approach has a weakness in the transformation of the forms of words. As of today, the conversion algorithm is based on the grammar rules of the studied language, which can distort the author's invented speech turns, losing the important elements of the author's style.

VII. CONCLUSION

According to the results of the conducted research, it can be stated that the use of this method makes it possible to formalize the specifics of the author's style as a software object. The main source of data for this object is objects that are multi-dimensional in their attributes, such as literary texts written by professional writers. It is assumed that the writer in his work adheres to a certain manner of writing, which makes it possible to use different methods for determining the authorship of his texts.

Summary results of the made experiments show the possibility of determining the authorship of the text with a probability of 60–80 %. In the process of setting up a pilot system, some new ideas how to increase the effectiveness of the developed approach, were obtained. These ideas require software implementation and new experiments to assess their impact on the effectiveness of text classification.

REFERENCES

- [1] P. A. Osipov and A. N. Borisov, "Abnormal action detection based on Markov models", in *Automatic Control and Computer Sciences*, vol. 45, no. 2. 2011, pp. 94–105. <http://dx.doi.org/10.3103/S0146411611020052>
- [2] The GraphML File Format. [Online]. Available: <http://graphml.graphdrawing.org>. [Accessed 05 July, 2014].
- [3] M. S. Elayidom, C. Jose et al., "Text classification for authorship attribution analysis", in *Advanced Computing: An International Journal*, ACIJ, vol. 4, no. 5, Sep. 2013, 10 p.
- [4] N. Homem and J. P. Carvalho, "Authorship Identification and Author Fuzzy Fingerprints" in *Fuzzy Information Processing Society (NAFIPS), 2011 Annual Meeting of the North American*, 978-1-61284-968-3/11/2011 IEEE, 2011, pp. 1–6.
- [5] A. Metwally, D. Agrawal and A. Abbadi "Efficient Computation of Frequent and Top-k Elements in Data Streams", University of California, Santa Barbara, USA, Tech. Rep. 2005–23, September, 2005.
- [6] R. M. Dabagh "Authorship attribution and statistical text analysis", in *Metodološki zvezki*, vol. 4, no. 2, 2007, pp. 149–163.
- [7] R. Zheng, Yi Qin, Z. Huang, H. Chen, "Authorship analysis in cybercrime investigation", H. Chen et al. (Eds.): *ISI 2003*, LNCS 2665, Springer-Verlag Berlin Heidelberg, 2003, pp. 59–73.
- [8] P. N. Bennett, S. T. Dumais and E. Horvitz. "The combination of text classifiers using reliability indicators", *Information Retrieval*, vol. 8, no. 1, pp. 67–100, 2005.
- [9] C. Sanderson and S. Guenter, "On Authorship Attribution via Markov Chains and Sequence Kernels," *18th International Conference on Pattern Recognition*, ICPR 2006, Aug. 20–24, 2006, Hong Kong, China. <http://dx.doi.org/10.1109/ICPR.2006.899>
- [10] E. Stamatatos, W. Daelemans et al., "Overview of the Author Identification Task at PAN 2014", *CLEF Conference, PAN part*, Sheffield, UK, Sep. 15–18, 2014.
- [11] H. P. Langtangen, "A Primer on Scientific Programming with Python", in *Texts in Computational Science and Engineering*, vol. 6. 4th ed. 2014, XXXI, 872 p. ISBN 978-3-642-54959-5.
- [12] J. R. Johansson, P.D. Nation and F. Nori, "QuTiP: An open-source Python framework for the dynamics of open quantum systems", in *Computer Physics Communications*, vol. 183, Issue 8, 2012, pp. 1760–1772. <http://dx.doi.org/10.1016/j.cpc.2012.02.021>

Pavel Osipov, Mg. sc. comp. He is a Doctoral Student of the Institute of Information Technology, Riga Technical University. He received his Master Degree from the Transport and Telecommunications Institute, Latvia. His research interests include web data mining, machine learning and knowledge extraction. A large part of research is focused on different aspects of user behavior modeling. Another new area of interests is to explore using Python programming language at all steps of scientific research, from initial idea brainstorm support, through all steps, till final article preparation in Text format, to provide for young researchers the common uniform research envelopment.
E-mail: pavel.osipovs@rtu.lv.

Andrejs Rinkevics, B. Sc. (Eng.), Institute of Information Technology, Riga Technical University. His interests include computer vision, machine learning, business process improvement with modern technologies, implementation of solutions, research of wireless technologies and data security issues.
E-mail: andrejs.rinkevics@rtu.lv.

Galina Kuleshova is a Researcher with the Faculty of Computer Science and Information Technology, Riga Technical University (Latvia). She received her M. Sc. degree in *Decision Support Systems* from Riga Technical University. Current research interests include artificial neural networks, data mining, classification methods and bioinformatics.
Address: Institute of Information Technology, Riga Technical University, 1 Kalku Str., Riga LV-1658, Latvia.
E-mail: galina.kulesova@cs.rtu.lv.

Arkady Borisov holds the Doctoral degree of Technical Sciences in Control of Technical Systems and the *Dr. habil. sc. comp.* degree. He is a Professor of Computer Science with the Faculty of Computer Science and Information Technology, Riga Technical University (Latvia). His research interests include fuzzy sets, fuzzy logic and computational intelligence. He has 205 publications in the area. He has supervised a number of national research grants and participated in the European research project ECLIPS.
E-mail: arkadijs.borisovs@cs.rtu.lv

Pāvels Osipovs, Andrejs Rinkevičs, Galina Kuļešova, Arkādijs Borisovs. Markova ķēžu pielietojšanas iespēju izpēti autora stila identifikācijai

Rakstā aprakstīts pētījums par Markova ķēžu pielietojšanu autora stila modeļu būvēšanai. Autora stila modelēšana ir aktuāls uzdevums. Pielietojot šādu modeli, to var salīdzināt ar dažādiem citiem tekstiem, kuriem autorība nav zināma. Šī salīdzinājuma rezultāts ir līdzības līmenis starp diviem tekstiem; ja tas ir pietiekami augsts, tad mēs varam teikt, ka abus tekstus rakstījis viens un tas pats cilvēks. Autora stila modeļa izmantošanas process ir iedalīts divās daļās: modeļa apmācība un tā izmantošana teksta analīzē. Modeļa apmācība balstīta uz tekstiem, kuriem autorība ir zināma. Tā rezultātā apmācīts modelis saglabā individuālu teikumu un frāzes būvniecības iezīmes. Svarīga pieejas iezīme šajā stadijā ir prasība izmantot liela apjoma tekstu apmācības procesā. Izmantošanas posmā, apmācīts modelis tiek izmantots, lai aprēķinātu stila līdzības līmeni ar analizēto tekstu. Apskatīts teorētiskais pamats, kā būvēt Markova ķēdes grafu, balstoties uz autora tekstu, kā arī apskatīta iespēja turpmākai teksta attīrīšanai, pirms tas tiek izmantots, lai apmācītu modeli, un tā ietekme uz gala klasifikācijas rezultātu. Veikti dažādi eksperimenti, lai novērtētu parametru ietekmi uz izmantoto algoritmu klasifikācijas efektivitāti. Galīgais līmenis ir tad, ja pareizie rezultāti sasniedz 60–80 %, kas ir samērā labi. Tālākai izpētei vajadzētu palielināt klasifikācijas precizitāti.

Павел Осипов, Андрей Ринкевич, Галина Кулешова, Аркадий Борисов. Исследование возможностей применения Марковских цепей для идентификации авторского стиля

В статье описано исследование возможностей применения Марковских цепей в задаче построения модели авторского стиля. Моделирование особенностей стилистики человека является актуальной задачей. Имея такую модель, возможно сравнивать с ней различные тексты, авторство которых не установлено. Итогом такого сравнения будет уровень сходства авторских стилей двух текстов. Если он достаточно высок, то можно говорить о том, что оба текста написал один и тот же человек. Использование модели авторского стиля делится на две условные части: обучение модели и непосредственно ее использование для анализа текста. Построение модели происходит на наборах текстов, для которых заведомо известно авторство. В итоге созданная модель хранит в себе особенности построения фраз и словосочетаний конкретного человека. Важной особенностью подхода на данном этапе является требование использовать для обучения тексты большого объема. На этапе использования обученная модель применяется для вычисления уровня сходства стиля с анализируемым текстом. Рассмотрена основная теоретическая база построения графа Марковской цепи, основываясь на авторском тексте. Рассматривается возможность дополнительной очистки текста перед его использованием для обучения модели и влияние этого на итоговый результат классификации. Произведены различные эксперименты для оценки влияния параметров используемого алгоритма на эффективность классификации. Итоговый уровень корректных результатов находится в районе 60–80 %, что сравнительно неплохо. Дальнейшие исследования должны увеличить уровень распознавания.

State of the Art in the Healthcare Cyber-physical Systems

Alyona Skorobogatjko¹, Andrejs Romanovs², Nadezhda Kunicina³, ¹⁻³ *Riga Technical University*

Abstract – For several years, experts from different fields are interested in virtual and physical world interaction opportunities. Cyber-physical systems (CPSs) are developed to integrate real physical processes and virtual computational processes. CPSs are used in multiple areas such as medicine, traffic management and security, automotive engineering, industrial and process control, energy saving, ecological monitoring and management, avionics and space equipment, industrial robots, technical infrastructure management, distributed robotic systems, protection target systems, nanotechnology and biological systems technology. CPS sensors generate the sensitive data that must be protected and shared only in a secure manner. This paper provides an overview of CPSs, their application in medicine and compliance with existing standards and possibilities for further advanced system development.

Keywords – Embedded systems, healthcare cyber-physical system, sensor networks.

I. INTRODUCTION

New competitive approach to the physical and virtual world integration with cyber-physical systems (CPSs) is one of the European Union research priorities. Cyber-physical systems will change the way people interface with systems, the same way as the Internet has transformed the way people interface with information.

The aim of the current research is to develop application strategies for cyber-physical system use in medical organizations and telemedicine. Topicality of the theme is based on the fact that cyber-physical systems are a promising area, development of which is critical for medical progress. One of the main challenges of this research is to design strategy for CPS application to be safe, secure, resistant to abnormalities and resilient in a variety of unforeseen and rapidly changing environments.

There are currently no available standards for healthcare cyber-physical systems. Therefore, it is necessary to conduct an in-depth research to identify medical systems' best practices and standards, which can be applied to healthcare cyber-physical systems.

CPS usage analysis will be performed within the framework of the research, and CPS possible application in disease prevention, diagnosis, treatment and patients' rehabilitation will be investigated. Particular attention will be paid to such essential components as usage, architecture, sensitivity, data management, data security and information exchange between multiple systems. As a result, healthcare cyber-physical system's overall requirements will be formulated and an application strategy will be developed.

This paper is organized as follows. In Section II, the definition and concept of a cyber-physical system are

introduced, and the embedded system development history is overviewed. In Section III, goals and activities of this area are observed in Latvian organizations. In Section IV, examples of CPS application in different fields, including healthcare, are given. In Section V, CPS design process features are discussed. Section VI includes a brief summary and conclusion.

II. CYBER-PHYSICAL SYSTEM OVERVIEW

Cyber-physical systems are developed to integrate real physical processes and virtual computational processes. Many objects used in a modern daily life are cyber-physical systems. Concept of CPS is complicated, but it can be illustrated with a concept map (see Fig. 1).

The definition of cyber-physical system from Cyber-Physical Systems Week [1] is as follows: "Cyber-physical systems (CPSs) are complex engineering systems that rely on the integration of physical, computation, and communication processes to function".

Cyber-physical systems have not appeared from anywhere, they have a long history of development, which continues. This chapter is an introduction to cyber-physical systems, their history and overview of the main components and characteristics.

A. From Embedded Systems to Cyber-physical Systems

Always a growing need for different purpose information management systems leads to the optimization of computing tool design techniques. Most of the world's currently used information management systems are embedded systems and networks. They are closely related to the control or management objects.

From certain common computing system classifications, the best suited to the modern situation is the classification proposed by David Patterson and John Hennessy [2]. Their classification was guided by the use of the system. They divided computing system into 3 categories: desktop computers, servers and embedded systems. Embedded systems by the area of use are separated into:

- Automatic control systems;
- Measuring systems and systems that read information from sensors;
- Real-time "question – answer" type information systems;
- Digital data transmission systems;
- Complex real-time systems;
- Moving objects management systems;
- A general purpose computer system subsystems;
- Multimedia systems.

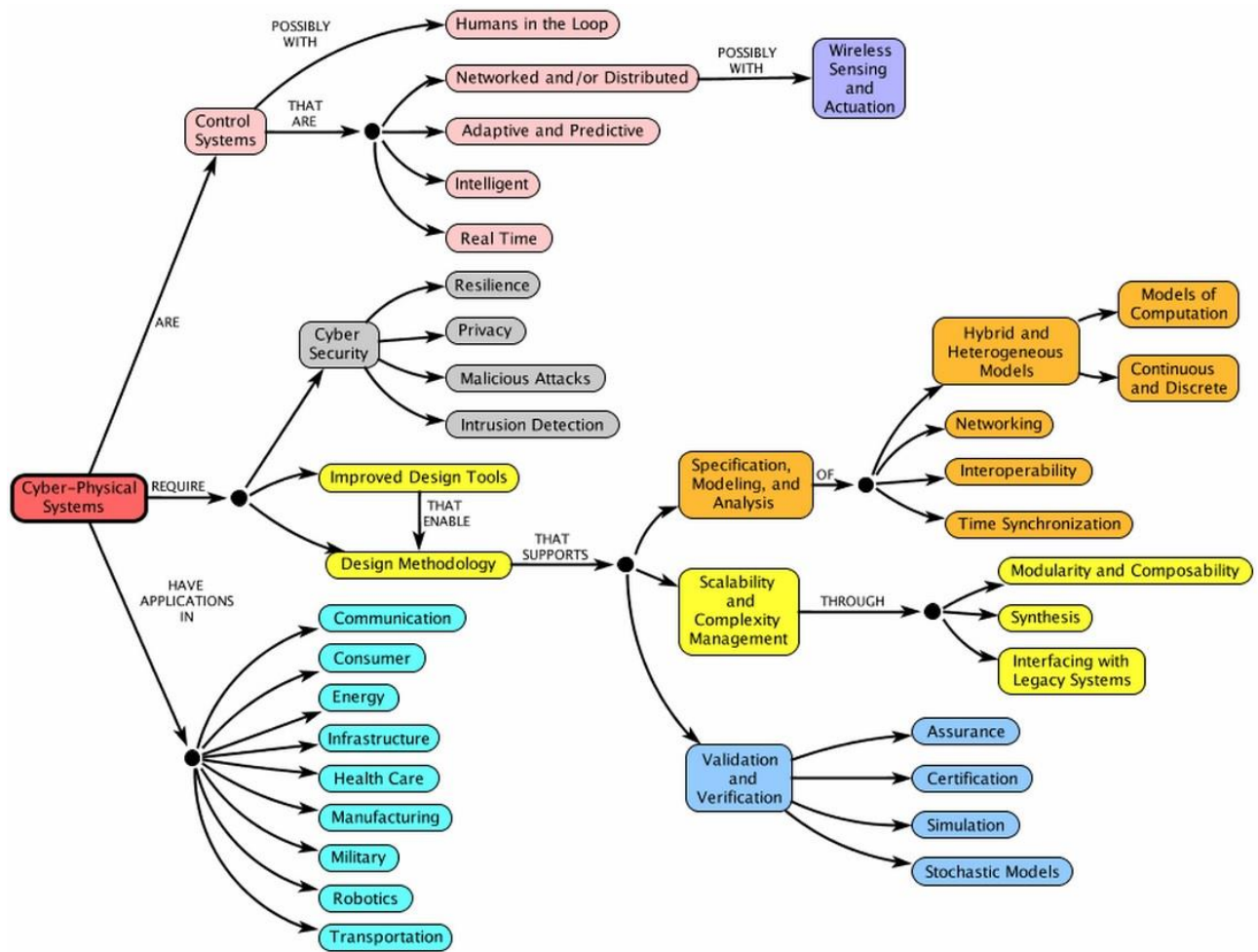


Fig. 1. A concept map of cyber-physical systems [3].

The concept of embedded systems appeared in the early 50s and it is in rapid development even today. It is interesting to view the evolution of embedded systems:

- Information management systems, 1960s;
- Embedded computing systems, 1970s;
- Embedded distributed systems, 1990s;
- Cyber-physical systems, since 2006.

Information management system is a computing system designed for management purposes, but it is the most alienated from the control object. Integrated micro-scheme and microprocessor development led to information management system bringing directly to the management object. The world has entered the era of embedded systems. System elements are gradually becoming cheaper and their integration increases, as well as the security level and the opportunity to combine them in controlled networks.

Downturn in embedded systems' elements prices and increasing connection with physical management objects led to the appearance of cyber-physical systems.

Cyber-physical systems are specialized computing systems that interact with control or management objects. Cyber-physical systems integrate computing, communication, data storage with real world's objects and physical processes. All

the above-mentioned processes must occur in real-time, in a safe, secure and efficient manner. Cyber-physical systems must be scalable, cost-effective and adaptive. Cyber-physical systems are in use in various areas, such as smart medical technologies, environmental monitoring and traffic management.

Wireless sensor networks can become an important part of cyber-physical systems, because of high sensitivity capability that is one of the main driving factors of CPS application distribution. The rapid development of WSN, medical sensors and cloud computing systems makes cyber-physical systems impressive candidates for use in inpatient and outpatient health care improvement [4]. Cloud computing maturity is a direct result of a few technologies, such as distributed computing, internet technology, system management and hardware development [5].

B. Concept of Cyber-physical Systems

Helen Gill, which at that time was the United States National Science Foundation Director of the embedded and hybrid systems sector, coined the term "cyber-physical system" in 2006. The new term highlighted NSF CPS Workshop originality. The seminar organizers sought to

review the role of embedded systems, and they succeeded. They saw the overall industry trend: after a few years cyber-physical system elaboration suddenly started in the world. Many countries recognized CPS as a priority research direction.

Cyber-physical systems integrate computing and physical processes. Compared with embedded systems, much more physical components are involved in CPSs. In embedded systems, the key focus is on the computing element, but in cyber-physical systems, it is on the link between computational and physical elements. Cyber-physical system parts exchange information with each other that is why the third component – communication (see Fig. 2) is added there. For this reason, cyber-physical system is denoted by the symbol C3 (Computation, Communication and Control).

Link improvement between computational and physical elements extends CPS usage possibilities.

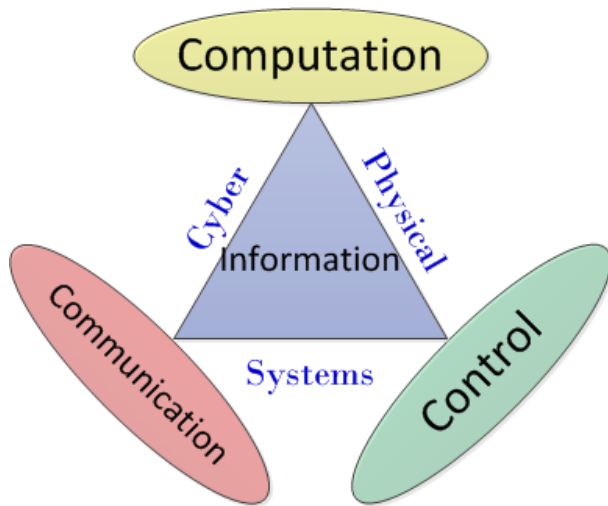


Fig. 2. Three main components of cyber-physical system [6].

III. SITUATION IN LATVIA

One of the main National Research Program aims in the period of 2014–2017 is: “to develop the scientific expertise of the next generation of ICT systems, creating a new competitive approach to the physical and virtual world integration cyber-physical systems, the development of competitive smart sensor networks and innovative hardware and software platforms, exploring and further developing competitive models based on the new information and communication technologies and their use in today’s environment”.

In the discussion within the Working Group of Smart Specialization Strategies, Modris Greitans (Director of the Electronics and Computer Science Institute) expressed the view that cyber-physical system is a large area and Latvia, as a small country, cannot overlay all aspects; however, certain aspects of this field have been studied and implemented in several institutions in Latvia with a certain specialization. Below, Latvian organizations are listed that develop cyber-physical systems:

- Electronics and Computer Science Institute: signal processing, intelligent embedded systems and their networks, testing and profiling environments;
- University of Latvia: different types of software for CFS;
- Riga Technical University: robotics and artificial intelligence, electrical and optical systems;
- Latvian University of Agriculture: biological principle-based computer-controlled systems;
- Ventspils University: smart sensor systems;
- Transport and Telecommunication Institute: smart transport systems.

These institutions do not compete among themselves but complement each other with different aspects. Together they fail to cover all the challenges in this field of research.

IV. CYBER-PHYSICAL SYSTEM APPLICATION

Cyber-physical systems are used in multiple areas, such as medicine, traffic management and security, automotive engineering, industrial and process control, energy saving, ecological monitoring and management, avionics and space equipment, industrial robots, technical infrastructure management, distributed robotic systems, protection target systems, nanotechnology and biological systems technology.

A. Smart Technologies

Modern construction technology enables the creation of intelligent building designed with minimum energy consumption or even without it. However, they need constant monitoring. Engineers must attach smart buildings to smart grids, and add control mechanism – cyber-physical systems [7].

Smart transport is equipped with various types of computerized embedded control systems. Cyber-physical solution use in this field makes it possible to create a full-fledged single system that will link the vehicles with other vehicles, environment and infrastructure [8]. An example of cyber-physical systems known to the general public is Google car, which does not require a driver.

B. Internet of Things and Industry 4.0

In several sources [5], [9], [10] it is predicted that within ten years almost half of the electronic devices will be connected to the World Wide Web. This network is termed as the Internet of Things. It connects not only household appliances such as refrigerators, thermostats, but also sophisticated production equipment.

Industry 4.0 concept aims at the comprehensive cyber-physical system use in manufacturing, customer relationship management and supply chain management processes, combining it all into one system. Smart manufacturing lines communicate with each other in order to optimize the production process.

Comprehensive use of cyber-physical systems for commerce, industry and public health, military and civilian purposes, makes the protection of these systems a matter of national significance. That is why embedded system security systems, mainly anomaly detection system that allows

resisting spoofing and service failure type attacks, are currently actively developed [11].

C. Healthcare Cyber-physical Systems

There are hospitals, where robots already bring dishes to patients, sort mail, change bed linen and collect waste. Robotic beds transport patients to the surgery room. However, a fully automated healthcare system has not been implemented yet. Currently, a number of hospitals in the world remote operations are carried out with the help of a robotic hand and high-resolution cameras [12]; however, there is still a long way to autonomous surgery when a cyber-physical system itself, without human management, performs the operation.

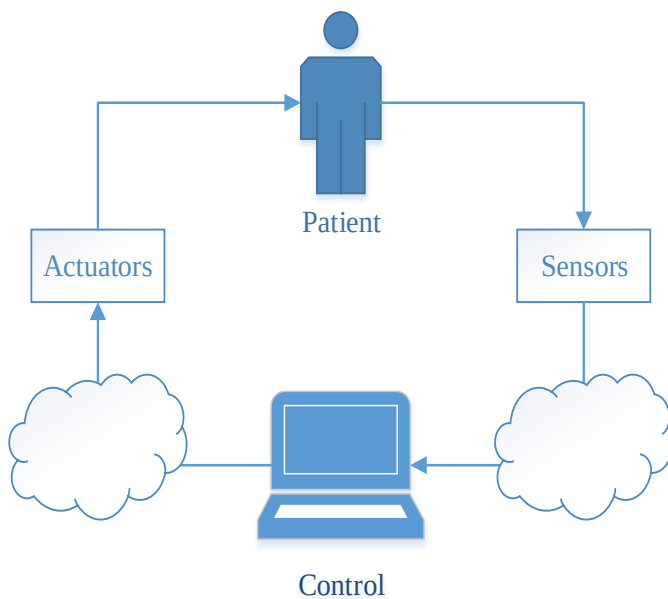


Fig. 3. An example of healthcare cyber-physical system.

Human-in-the-loop cyber-physical systems can greatly improve lives of people with special needs. Human-in-the-loop cyber-physical systems formulate opinions about the user's intentions based on his cognitive performance by analyzing data from sensors attached to the body or head. Embedded system converts these findings to robot control signals, which, thanks to robotic management mechanisms, allow users to interact with the surrounding natural environment. Example of human-in-the-loop CPS is robotic assistance systems and intelligent prosthesis [13].

From the examples above, it is clear that the cyber-physical systems are widely used all over the world in various industries, including medicine and healthcare. However, only one research was found [14], which summarizes the current situation related to the CPS application in healthcare, and it offers a detailed taxonomy.

Healthcare cyber-physical systems were categorized by application (assisted, controlled), architecture (infrastructure, data requirement, composition system), sensing (sensor type, method, parameter), data management (data integration, data storage, data processing), computation (modeling, monitoring), communication (scheduling, protocol), security

(privacy, encryption), control/actuation (decision-making, mechanism).

Existing healthcare cyber-physical systems were mapped to this taxonomy, and a number of healthcare CPS groups (with several examples) were allocated: notable CPS applications (*Electronic Medical Records, Medical CPS and Big Data Platform, Smart Checklist*), daily living applications (*LiveNet, Fall-Detecting System, HipGuard*), medical status monitoring applications (*MobiHealth, CodeBlue, AlarmNet*), medication intake applications (*iCabiNET, iPackage.*).

Using this taxonomy and mapping, different cyber-physical systems can be categorized. In addition, it is quite convenient to use this taxonomy for formulation of new cyber-physical system requirements.

V. DESIGN PROCESS

Cyber-physical system, as well as any embedded system design process includes requirements management and project management, as well as test and safety plan. However, cyber-physical system development process is not possible without the involvement of a different sector professional. CPS development requires the knowledge of computers, software, networks, and physical processes, which will be integrated into the system.

Real-time operating systems manage embedded systems. There software correctness is associated not only with the correct ending, but also, for a certain task implementation compliance of spent time with the planned time. It is quite difficult to transfer these properties to the cyber-physical systems. Different types of processors and memory architectures, as well as a variety of operation systems and programming environments are used in CPSs.

The development of cyber-physical systems, which closely interact with physical processes, requires a technically sophisticated low-level design. CPS developers are forced to fight with break controllers, memory architecture, processes planning, assembler level programming, network interface design and installation of drivers. However, it would be better to focus on system functionality and behavior definition.

The great complexity is not determinate due to the real environment, in which the cyber-physical system is used. Cyber-physical system may change by the time and no one of the system's component is completely safe.

All the above-described deviations from the final task solving encouraged looking for new approaches to cyber-physical system design. Experts from the University of Berkeley [15] believe that cyber-physical system's design should be based on the common system behavior modeling and hardware, software, network and physical process single design.

CPS design process is an iterative process, which consists of three phases: modeling, design and analysis (see Fig. 4). In order to develop systems, which are able to work in the real world, a new approach – model designing is used.

Firstly, the real world analysis should be performed, and physical processes identified, which will interact with cyber-physical system. Then it is possible to start development of an

abstraction of the real world processes. Model design allows overcoming some of the fundamental problems in the cyber-physical system design, such as security, determinism and time.

Every system is designed to address a specific problem or complete a task. System requirements can be described in a simple language, in formal way or with models. The formal description reduces the number of possible errors, but modeling serves to illustrate the system dynamics.

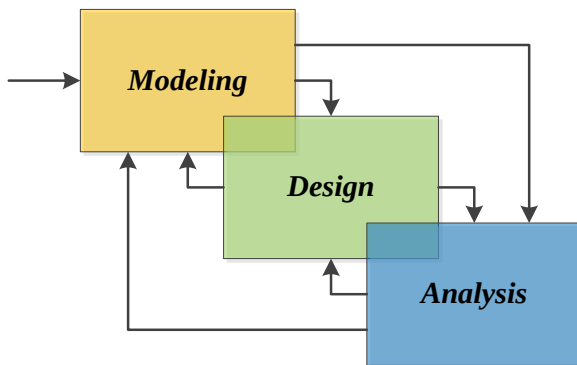


Fig. 4. Cyber-physical system design process.

During the process of analysis, the extent to which the system has met the requirements specification can be determined. CPSs are either commissioned or altered in accordance with the previous and the new requirements. New desires, needs and requirements to a cyber-physical system will also arise in the use process.

In real conditions, for the research purposes only a verified and validated cyber-physical system that successfully passed safety tests must be used. It also applies to the new version of the system or a supplement. Nevertheless, in healthcare facilities only certified cyber-physical systems must be in use.

A set of regulations and standards is applied to all medical systems, which indicates that the system is electrically, physically, biologically and chemically safe for the end user. Cyber-physical system sensors generate the sensitive data that must be protected and shared only in a secure manner. Currently, there are no standards and regulations, which relate directly to healthcare cyber-physical systems [8], [14]. However, it is worthwhile to be acquainted with medical equipment standards, regulations and clinical standards.

According to the European Union Glossary of Terms [16]:

- Standard is national or international level detailed engineering characteristics that determine the optimal requirements – performance, safety and reliability.
- Regulation is a set of rules with the force of law issued by the executive authority.

As shown, the regulations are mandatory, while standards – desirable. Although the regulations of medical equipment are similar in several countries, they are not identical. This means that the same medical device can be recognized as safe for use in medical institutions of the European Union, but not in the United States.

The CE marking certifies that a product has been approved for use and distribution in Europe – a direct translation of the French term “Conformité Européen” or “Approved in Europe”. By contrast, in the United States U.S. FDA (United States Food and Drug Administration) shall indicate the medical equipment compliance with their regulations. Countries, which have not developed or adopted any medical equipment regulations primarily, apply the WHO (World Health Organization) regulations.

While standards of healthcare cyber-physical systems are not yet developed, a number of industrial, communications (for example, Bluetooth – IEEE 802.15.1, Wi-Fi – IEEE 802.11), quality (ISO 9001, ISO 13845, ISO 14971), medical equipment (ISO/IEEE 11073, HL7 – Health Level 7, DICOM – Digital Imaging and Communications in Medicine) and clinical standards can be applied to them.

ISO/IEEE 11073 is a standard set for personal healthcare devices. The general context of it is data communication and interface between the agent and the manager.

Health Level 7 is a group of standards for the exchange, integration, sharing, and retrieval of electronic health information. HL7 determines many adaptable standards, guidelines, and methodologies that enable communication of various computer systems in hospitals and different healthcare providers.

Generalizing, medical equipment standards have been drawn up to improve the quality and number of facilities, and equipment interoperability.

VI. CONCLUSION

Healthcare cyber-physical systems can reach the level of functionality, adaptability, and effectiveness previously unattainable for medical devices.

During the first part of research, the authors reviewed healthcare cyber-physical system application and compliance with existing standards and possibilities for further system development.

After identification of main tasks of the healthcare cyber-physical system and looking at the above-mentioned regulations and standards, a requirement definition can be initiated so that the CPSs comply with appropriate regulations and standards.

ACKNOWLEDGEMENT

This research has been supported by the Latvian state research program 2014–2017 project “The Next Generation of Information and Communication Technologies (NexIT)”.

REFERENCES

- [1] “Cyber-Physical Systems Week” [Online]. Available: <http://www.cpsweek.org/> [Accessed: Sept. 1, 2014].
- [2] D. Patterson, J. Hennessy, *Computer Organization and Design: The Hardware Software Interface*. 5th edition. Morgan Kaufmann, 2013.
- [3] “Cyber-Physical Systems” [Online]. Available: <http://cyberphysicalsystems.org/> [Accessed: Sept. 1, 2014].
- [4] A. Milenkovic, C. Otto, E. Jovanov, “Wireless sensor networks for personal health monitoring: issues and an implementation,” *Computer Communications*, vol. 29, no. 13–14, 2006, pp. 2521–2533. <http://dx.doi.org/10.1016/j.comcom.2006.02.011>

- [5] R. Buyya, J. Broberg, A. Goscinski, *Cloud Computing: Principles and Paradigms*, John Wiley & Sons, Hoboken, NJ, USA, 2011. <http://dx.doi.org/10.1002/9780470940105>
- [6] N. Wu and X. Li, *RFID Applications in Cyber-Physical System, Deploying RFID – Challenges, Solutions, and Open Issues*. InTech, 2011.
- [7] Cyphers, *Cyber-Physical European Roadmap & Strategy*. Deliverable D5.1: CPS: State of the Art, 2014.
- [8] Steering Committee for Foundations in Innovation for Cyber-Physical Systems, *Foundations for Innovation: Strategic R&D Opportunities for 21st Century*, 2013.
- [9] AENEAS Industry Association, *Part B of the 2014 ECSEL MARSIA*, 2013.
- [10] A. Koubaa, B. Andersson, "A Vision of Cyber-Physical Internet," *Proceedings of the 8th International Workshop on Real-Time Networks*, Dublin, Ireland, 2009.
- [11] S. Amin, G. Schwartcz, A. Hussain, "In Quest of Benchmarking Security Risks to Cyber-Physical Systems," *IEEE Network*. Jan./Feb. 2013, pp. 19–24. <http://dx.doi.org/10.1109/MNET.2013.6423187>
- [12] "Networking and Information Technology Research and Development Program," *High-Confidence Medical Devices: Cyber-Physical Systems for 21st Century Health Care*. NTRD, 2009.
- [13] G. Schirmer, D. Erdogmus, K. Chowdhury, T. Padir, "The Future of Human-in-the-Loop Cyber-Physical Systems," *Computer*, IEEE computer Society Digital Library, 2013, pp. 36–45. <http://doi.ieeecomputersociety.org/10.1109/MC.2012.428>
- [14] S. A. Haque, S. M. Aziz, M. Rahman, "Review of Cyber-Physical System in Healthcare," *International Journal of Distributed Sensor Networks*, Hindawi Publishing Corporation, vol. 2014, p.20, 2014. <http://dx.doi.org/10.1155/2014/217415>
- [15] E. Lee, S. Seshia, *Introduction to Embedded Systems, A Cyber-Physical Systems Approach*. 2014.
- [16] Eiropas Savienības terminu vārdnīca. – R., UNDP, 2004.

Alyona Skorobogatjko, B. sc. ing., is a Master's Student at the Department of Information Technology Management, Riga Technical University (Latvia).

She received the Bachelor Degree in Information Technology from Riga Technical University in 2012. Her Bachelor Thesis on SCADA network security received special support from the Latvian local energy production company and was among laureates of the best Latvian theses in the field of Information Technology in 2012. She has experience in software development, and currently she is a Project Manager at an international company. Her professional interests include project management, software and systems, particularly healthcare systems, design, modeling and development. She has participated in a number of international scientific conferences and multiple projects.
E-mail: alona.skorobogatjko@rtu.lv.

Andrejs Romanovs, Dr. sc. ing., is an Associate Professor and Lead Researcher at the Information Technology Institute, Riga Technical University. He has 25 years of professional experience teaching post-graduate courses at RTU and developing more than 50 industrial and management information systems. His professional interests include modeling and design of management information systems, IT governance, IT security and risk management, integrated information technologies, as well as education in these areas. A. Romanovs is a senior member of the IEEE, LSS; the author of more than 50 papers in scientific journals and conference proceedings in the field of Information Technology.
E-mail: andrejs.romanovs@rtu.lv,

Nadezhda Kunicina, Dr. sc. ing. Since 2005 she has been an Associate Professor with RTU. Her fields of scientific interest are electrical engineering, embedded systems, sustainable transport systems. The research interest is related to the improvement of electric energy effectiveness in industrial electronics and electric transport. She is a Scientific Project Manager of FP7 – ARTEMIS – 2012 called Arrowhead project. The aim of project is to address the technical and applicative challenges associated with cooperative automation.
E-mail: nadezda.kunicina@rtu.lv.

Aļona Skorobogatjko, Andrejs Romanovs, Nadežda Kunicina. Veselības aizsardzības kiberfizikālo sistēmu apskats

Jau vairākus gadus dažādu nozaru speciālisti visā pasaulē interesējas par virtuālās un fiziskās pasaules mijiedarbības iespēju izmantošanu. Kiberfizikālo sistēmu pielietošanas mērķis ir integrēt reālus fiziskus procesus ar virtuāliem skaitļošanas procesiem. Pētījuma tēmas aktualitāte ir neapšaubāma un balstās uz to, ka kiberfizikālās sistēmas ir daudzsoļīga joma, kuras attīstība virza uz priekšu medicīnu. Kiberfizikālās sistēmas ir specializētas skaitļošanas sistēmas, kurām ir fiziskas mijiedarbības līdzekļi ar kontroles un vadības objektu. Iegultās sistēmās galvenais fokuss ir skaitļošanas elements, bet kiberfizikālās sistēmās – saite starp skaitļošanas un fizisko elementu. Kiberfizikālās sistēmas tiek pielietotas tādās jomās kā medicīna, transporta plūsmu vadība, automobiļu būvniecība, industriālo un tehnoloģisko procesu vadība, enerģijas ekonomija, ekoloģijas monitorings un pārvaldība, avionika un kosmiskā tehnika, industriālie roboti, tehniskās infrastruktūras pārvaldība, sadalītas robottehniskās sistēmas, sistēmas aizsardzības mērķiem. Ir slimnīcas, kur jau tagad roboti pienes pacientiem ēdienu, šķiro pastu, maina gultasveļu, kā arī transportē pacientus uz operāciju zāli ar robotizētām gultām. Taču pilnībā automatizēta veselības aizsardzības sistēma vēl nevienā medicīnas iestādē nav ieviesta. Kiberfizikālo sistēmu izstrādei ir nepieciešamas gan zināšanas par datoriem, programmatūru, tīkliem, gan arī par fiziskiem procesiem, ar kuriem tiks integrēta izstrādājamā sistēma. Kiberfizikālo sistēmu izstrāde ir iteratīvs process, kas sastāv no 3 fāzēm: modelēšanas, izstrādes un analīzes. Pirms veselības aizsardzības kiberfizikālās sistēmas izstrādes ir vērts iepazīties ar medicīnas iekārtu standartiem un nolikumiem, kā arī klīnisko pētījumu standartiem, kurus apskatot un, identificējot izstrādājamās sistēmas galvenos uzdevumus, var uzsākt prasību definēšanu atbilstoši attiecīgajiem nolikumiem un standartiem.

Алёна Скоробогатько, Андрей Романов, Надежда Кунницына. Современное состояние медицинских кибер-физических систем

На протяжении последних десятилетий специалисты из разных областей заинтересованы в исследовании возможностей взаимодействия реального мира с виртуальным. Кибер-физические системы – это специализированные вычислительные системы, которые включают в себя объекты взаимодействия с реальным миром, вычислительные объекты и систему коммуникации. Тема исследования весьма актуальна, так как кибер-физические системы являются многообещающей отраслью, развитие которой способствует развитию медицины в целом. Во встроенных системах главной является вычислительная часть, а в кибер-физических системах – связь между вычислительными и физическими элементами. В современном мире кибер-физические системы используются в медицине, в том числе в телемедицине, управлении транспортом и обеспечении безопасности, автомобилестроении, в распределённых роботизированных системах, управлении производственными и технологическими процессами. Существуют больницы, в которых уже сейчас роботы разносят еду пациентам, меняют постельное бельё, с помощью роботизированной кровати пациенты транспортируются в операционные залы, но полностью автоматизированная система здравоохранения ещё не введена ни в одном медицинском учреждении. Разработка кибер-физических систем невозможна без знаний о вычислительных системах, сетях, но ещё более важными являются знания о физических процессах, интегрируемых с виртуальными процессами. Процесс разработки кибер-физических систем является итеративным процессом и состоит из трех фаз: моделирования, разработки и анализа. На данный момент не существует стандартов и регламентов, относящихся именно к медицинским кибер-физическим системам, поэтому стоит ознакомиться со стандартами и регламентами разработки медицинских систем, а также со стандартами клинических исследований. Рассмотрев вышеупомянутые регламенты и стандарты и определив главные задачи разрабатываемой системы, можно начать процесс формулирования требований к медицинской кибер-физической системе.

Artificial Neural Network Generalization and Simplification via Pruning

Andrey Bondarenko¹, Arkady Borisov², ¹CTCo Ltd., ²Riga Technical University

Abstract – Artificial neural networks (ANNs) are well known for their classification abilities. Although choosing hyper-parameters such as neuron layer count and size can be a quite tedious task. Pruning approaches assume that a sufficiently large ANN has already been trained and can be simplified with acceptable classification accuracy loss. The current paper presents a node pruning algorithm and gives experimental results for pruned network accuracy rates versus their non-pruned counterparts.

Keywords – Artificial neural networks, generalization, overfitting, pruning.

I. INTRODUCTION

Artificial neural networks have been successfully applied in many different areas to solve the problems of classification and regression. ANNs, specifically multi-layer feed-forward artificial neural networks, trained using error back-propagation give good results, and the recent advances in deep learning have proved to give unprecedented classification accuracy [1]–[2]. Main problem with regular (non-deep) ANNs is choosing hyper-parameters, which can severely influence model performance. Choosing architecture with insufficient number of neurons can give unsatisfying classification rates, while choosing too many neurons will badly influence training time, but more than that it will cause overfitting. To overcome these problems, two approaches exist: to grow neural networks [3], [4] and to train an excessively large network with subsequent pruning. In the current paper, we use the second approach to overcome overfitting of trained ANNs, thus enhancing their generalization abilities, as well as preparing the previously trained artificial neural networks for rule extraction.

The current paper is structured as follows: Section II gives an overview of pruning methods; Section III describes the used algorithm; Section IV presents results of experiments, and Sections V and VI hold discussion and conclusion.

II. PRUNING METHOD OVERVIEW

First of all, we should note that some algorithms prune weights between neuron layers, others prune neurons themselves, and some combine both approaches. Paper [5] gives a nice overview; we briefly present main ideas mentioned there along with other methods not listed in the source paper. We can divide all pruning algorithms into two main categories: node/weight removal based on sensitivity analysis and penalty term based methods that utilize penalty term to remove “unused” / least important weights. Some algorithms combine both approaches, while some cannot be easily added to one or the other family of methods. Sensitivity

analysis relies on the calculation of influence of specific node or weight. Some algorithms just calculate pairs of weights coming into/out of a single neuron. Finding the smallest weight pairs assumes that their influence on classification result is minimal.

A. Sensitivity Analysis Based Methods

Sensitivity method from [6] by Mozer and Smolensky calculates an error with unit removed and without it being removed, thus deleting least important units. Instead of calculating an error directly, they use derivative calculated during error back-propagation to approximate it. Segee and Carter in [7] have found that small variance in weights coming into a neuron is signaling that the subject neuron can be safely removed.

Karnin in [8] describes the method for weight pruning, which does not require specific sensitivity calculation phase. All necessary data about weight updates are stored during training. This makes this approach unusable in case one wants to prune the already trained ANN. Nevertheless, such an approach is perfectly usable in case one has control over training of ANN. The main idea is the sensitivity analysis of weights and their removal if they have too small sensitivity.

$$S_{ij} = - \frac{E(w^f) - E(0)}{w^f - 0} w^f \quad (1)$$

Here, w^f is a specific weight value, 0 is its value after its pruning, $E(w^f)$ is an error with given weight enabled and $E(0)$ is an error when this weight is pruned. Finally, S_{ij} is sensitivity of weight between nodes i, j . Instead of computing sensitivity value directly (which would lead to a sensitivity estimation phase), the authors propose to estimate it using a sum of all changes in weight during training:

$$\hat{S}_{ij} = - \sum_{n=0}^{N-1} \frac{\partial E}{\partial w_{ij}} \Delta w_{ij}(n) \frac{w_{ij}^f}{w_{ij}^f - w_{ij}^i} \quad (2)$$

Here, N – is a number of weight updates.

Rudy Setiono [9] describes a rather simple pruning algorithm, which uses simple heuristics to find weights to be pruned. It assumes that one has ANN with a single hidden layer (although an approach can be generalized to multiple hidden layers) and removes input-hidden and hidden-output weights if their values are not satisfying a specific constant. Actually, there are two constants used, both of them should be set-up manually.

Rudy Setiono also described a rule extraction algorithm [10] called N2FPA, which uses simple estimations of effect of removal of neurons in the network. Neurons are removed one by one, in case an error worsens significantly pruning stops.

This is the method, which was used and slightly modified in the current paper.

Le Cun et al. in [11] describe the method called *Optimal Brain Damage* (OBD), which measures “saliency” of a weight by estimating the second derivative of the error with respect to the weight. They made a couple of assumptions after which computed such derivatives during the modified error back-propagation. One drawback of such a method is the necessity of storage of Hessian matrix. After one weight is pruned, retraining is done to find another weight to prune.

Optimal Brain Surgery from [12] (OBS) goes one step further in comparison with OBD, it utilizes an inverse Hessian matrix to calculate optimal weight to be deleted, but at the same time it solves an optimization problem, which gives remaining weight updates necessary to lower a network error. Such an approach allows for simultaneous update of all remaining weights; thus, retraining is not required. OBS is one of the best methods for pruning. Likewise OBD, it should hold the Hessian matrix, thus requiring additional memory.

B. Penalty Based Methods

Penalty term methods use weight decay/penalty term in one way or another to force a neural network during its training to get rid of unnecessary weights.

Chauvin [13] uses a cost function with a specific term, which poses average energy expended by weights, as well there is a modification with additional magnitude of weight term, which penalizes large weights and large amount of weights.

Weigend et al. [14]–[16] minimize a specific cost function with additional term penalizing network complexity as a function of the weight magnitudes relative to the defined constant w_0 . Choosing such a constant should be done via trials/errors.

Ji et al. [17] propose another penalty term pruning approach based on a modified error function, which tries to minimize a number of hidden nodes and weight magnitudes. The limitation of the proposed method is that it assumes a single hidden layer ANN with one input and *linear* output node. The method assumes retraining after each removed weight.

C. Weight Decay Methods

Plaut et al. [18] propose a simple cost function, which decays weights. Cost function built specifically to allow an algorithm to favor nodes with many small weights in contrast to a node with single large connection. Nowlan and Hinton [19] describe a more complex cost function with a penalty term, which models the probability distributions of weights as a mixture of Gaussians.

D. Interactive Pruning

Sietsma and Dow [19] describe an interactive method in which a designer inspects a network and marks nodes to be pruned. Algorithm provides several heuristics to determine candidates for removal. The authors have demonstrated on training problems that their method is capable of finding relatively small networks with good accuracy in comparison

with large trained networks, which were not able to find a solution.

E. Auto-pruning Methods

Next discussed approach is an auto-pruning method called *lprune* [20] by Lutz Prechelt. It proposes pruning at each step all weights not satisfying a specific formula controlled by parameter λ . Experiments showed that this parameter should be adaptive; the algorithm to support dynamic adjustment was proposed. According to Prechelt, the proposed methods overcome OBD and OBS in terms of accuracy and simplicity of pruned ANN.

Another auto-pruning method [21] by William Finoff et al. applies a modified cost function, does not require full training of ANN and uses dynamic adjustment of penalty term. Similar to OBS, this method performs dynamic topology adjustments.

F. Other Methods

Kruchke [22] describes a local bottleneck method, in which neurons “compete” with each other to survive. Magnitudes of vectors determine a degree to which a neuron participates in modeling target function; this is treated as a neuron gain. In case gain is zero, a neuron does not participate in a task and can be removed. In case, two neurons have parallel or anti-parallel weight vectors, they are redundant and can be removed as well. The method uses a specific parameter, which should be tuned carefully.

The same author proposes another method called Distributed Bottlenecks [22]–[23], which puts constraints on weights rather than deletes them. This serves as sort of dimensionality reduction. Such an approach makes weight vectors that are further apart than average to become further from each other and vectors that are closer than average to become closer. Again the method uses a special constant, which should be chosen manually

III. PROPOSED ALGORITHM

We used the algorithm described by Rudy Setiono in [10] (part of N2FPA rule extraction method). In essence we used node pruning. We operated on the trained ANN, on each pruning we tried to determine a neuron, which needs to be removed. For all neurons in hidden layers we calculated a classification error for network operating without them. (This essentially means we set activations of pruned neurons to zero.) Neuron, which upon removal produced a network with the smallest cost function, was removed. Afterwards, the network was retrained. If classification accuracy raised above a specific threshold (we used tolerance to error equal to 2.5 % out of the smallest encountered classification error, i. e., we accepted removal of a neuron in case the classification worsened for no more than 2.5 %), then a neuron was accepted to as pruned. If an error significantly, then a candidate neuron was left intact and new search for a pruning candidate was initiated on the already retrained network (remember we retrained it without a “pruned” neuron.) Retraining gives chances to get a simpler network with higher generalization and good classification rates, which can be observed in Table I.

TABLE I
TYPE SIZES, SPACES AND INTERVALS

	MLP train avg.	MLP train std. dev	MLP test avg.	MLP test std. dev	Pruned train avg.	Pruned train std. dev	Pruned test avg.	Pruned test std. dev	Architecture before/after pruning (hidden nodes in 2 hidden layers)
Ionosphere (10-fold X-validation)	10.83%	0.0013	10.83%	0.0115	5.39%	0.0169	10.44%	0.0326	15-15 / 5.4-3.8
Monks-1 (train/test)	20.16%	0.0120	29.68%	0.0166	18.47%	0.1067	24.35%	0.0952	15-15 / 5-3
Monks-2 (train/test)	36.82%	0.0135	36.55%	0.0087	31.83%	0.2005	32.58%	0.2170	15-10 / 5-3.1
Monks-3 (train/test)	6.64%	0.0026	2.80%	0.0007	5.98%	0.0078	2.85%	0.0105	15-15 / 1.7-1.1
WPBC (10-fold X-validation)	0%	0.0	0%	0.0	0%	0.0	0%	0.0	10-10 / 1-1
WDBC (10-fold X-validation)	3.89%	0.0040	4.04%	0.0278	3.03%	0.0098	3.69%	0.0271	30-30 / 23.2-17.1
Pima (10-fold X-validation)	23.02%	0.0073	23.56%	0.0498	25.94%	0.0412	26.81%	0.0598	10-10 / 2.8-3.1
Haberman (10-fold X-validation)	26.13%	0.0055	26.57%	0.0255	28.5%	0.0762	28.2%	0.0981	15-15 / 2.3-3.7
Parkinsons (10-fold X-validation)	24.62%	0.0021	24.61%	0.0197	16.29%	0.0241	15.82%	0.0848	30-30 / 26.8-28

Below you can find a pseudo code of the pruning algorithm:

Inputs:

maxIter - determines maximum number of pruning iterations
maxPrunedNodes - maximum number of nodes to be pruned
errorRiseTol - determines acceptable error rise
maxFallbacks - in case neurons are pruned and then reverted - how many times before we quit?

Program:

```

iter = 1
while (iter < maxIter || prunedNodes < maxPrunedNodes)
  for all not pruned neurons in all hidden layers
    if (lastNeuronInLayer()) continue;
    removeNeuron(n)
    cost = testNetwork()
    if (cost > largestKnownCost)
      largestKnownCost = cost
      indexOfPrunedNeuron = getIndexOfPrunedNeuron
      prunedNodes = prunedNodes + 1
    end
  end
end

retrainNetwork()

classError = testNetwork()
errRatio = classError/smallestClassError
if (errRatio > 1 + errorRiseTol)
  revert pruned neuron
  fallbacksCounter = fallbacksCounter + 1
  prunedNodes = prunedNodes - 1
else
  //leave pruned neuron as is
  fallbacksCounter = 0
end
if (fallbacksCounter >= maxFallbacks)
  break
end
iter = iter + 1
end

```

Here one can notice hyper-parameters listed in the beginning, *maxIter* – controls a maximum possible number of pruning iterations, *maxPrunedNeurons* – controls a maximum number of neurons to be pruned. We need both parameters, as neurons after pruning can be restored; thus, some iterations will not result in network pruning. However, they will leave network with weights adjusted during retraining. Apart from that, there are two other hyper-parameters: *errorRiseTol* – which controls maximum rise of error (we used a classification error) after removal of a neuron which will not cause pruned neuron reversal/restore. Thus, let us say an error has risen by more than 5 % in case of neuron removal in comparison with the best/lowest known error rate and our parameter is 0.075. We will leave the network intact, but if it is below 0.05 the network will get back the pruned neuron. Finally, *maxFallbacks* controls how many attempts an algorithm makes in pruning neurons and reverting them back consequently before termination. Thus, if this parameter is equal to 10, then in case of ten subsequent iterations a neuron is pruned, but then restored due to a high rise in an error, the algorithm terminates.

ANN itself is trained using the cross-entropy cost function equipped with a penalty term (weight decay). Below there is a cost function:

$$F(w, v) = - \left(\sum_{i=1}^k \sum_{p=1}^C l_p^i \log S_p^i + (1 - l_p^i)(1 - \log S_p^i) \right), \quad (3)$$

where k is the number of patterns, $l_p^i = 0$ or 1 is the target value for pattern x^i at an output unit p , $p = 1, 2, \dots, C$. C is the number of output units/neurons. S_p^i is the output of the network at unit p :

$$S_p^i = \sigma \left(\sum_{m=1}^h \sigma \left((x^i)^T w^m \right) v_p^m \right). \quad (4)$$

Here, to simplify things a bit we provide formulas for a single hidden layered neural network, but in reality for our experiments we used two hidden layers. x^i is an n -dimensional input pattern, $i = 1, 2, \dots, k$. w^m is an n -dimensional

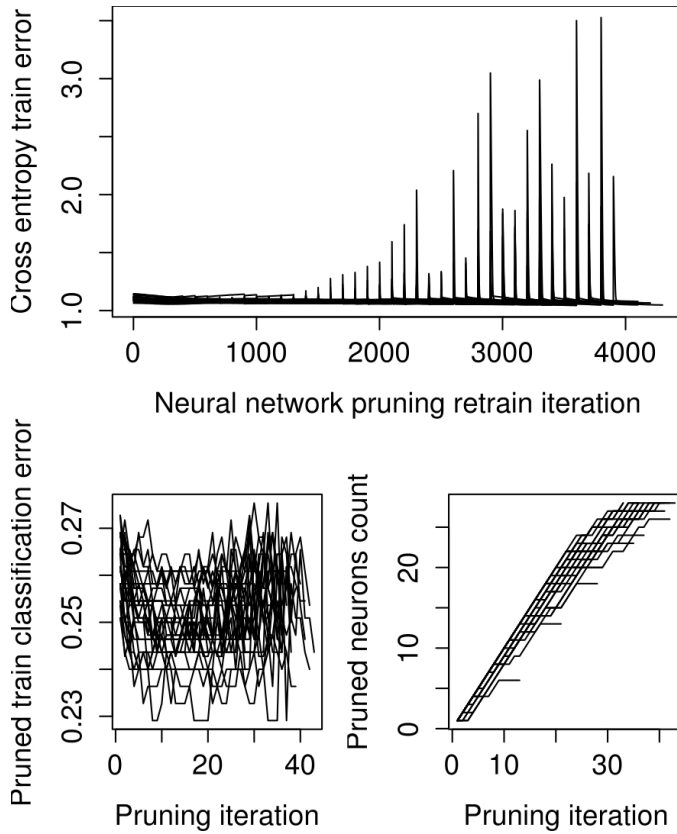


Fig. 1. Haberman data set trained neural network pruning.

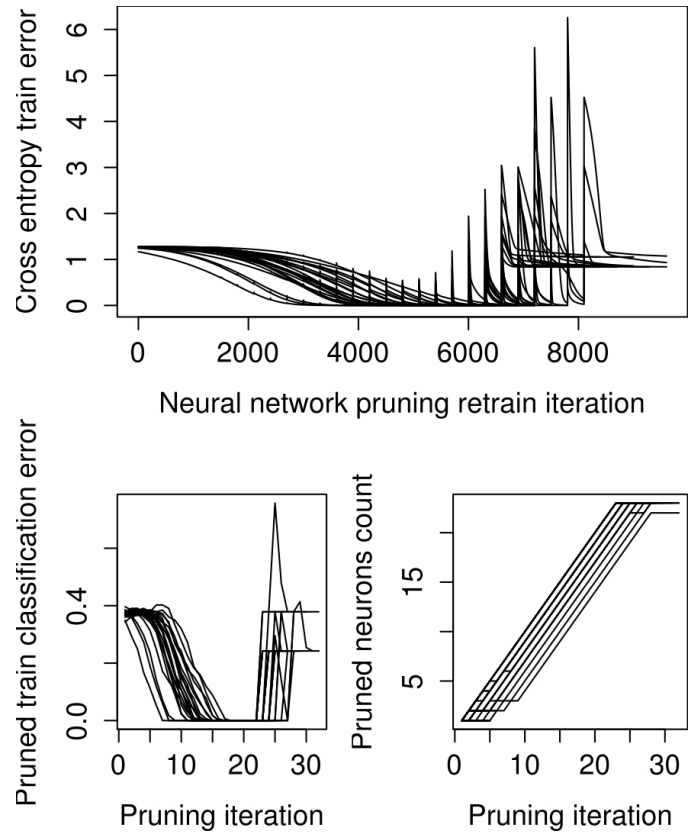


Fig. 3. WDBC data set trained neural network pruning.

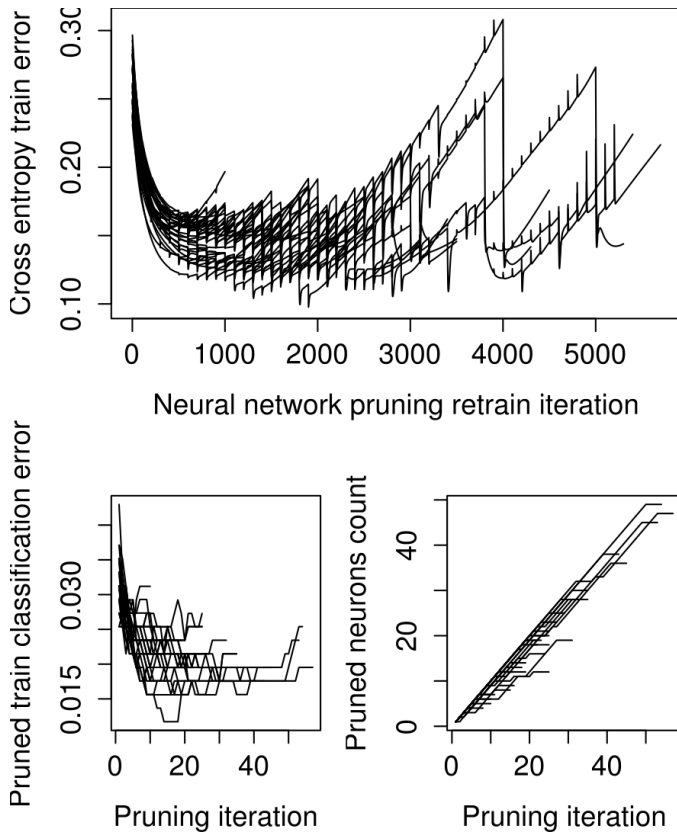


Fig. 2. Monks2 data set trained neural network pruning.

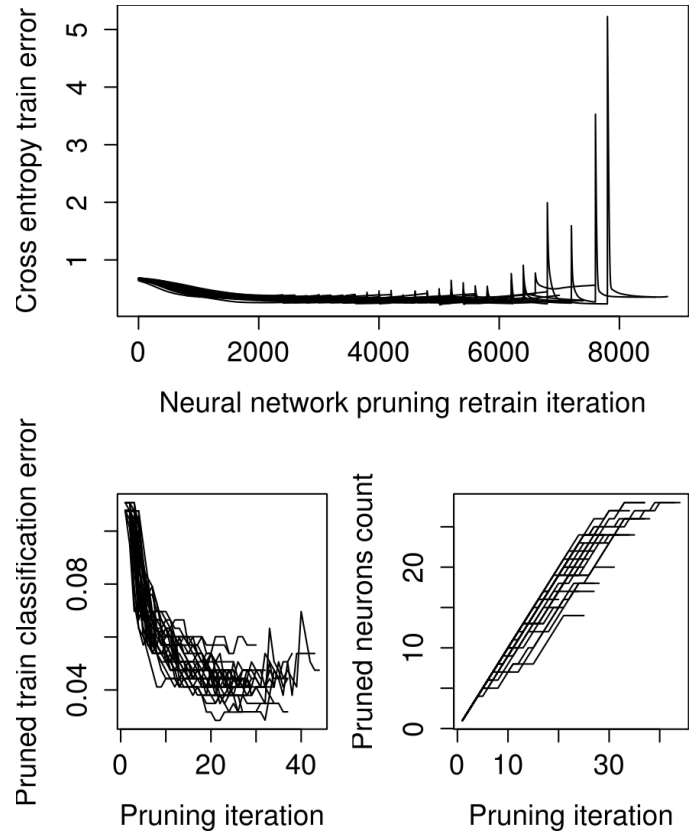


Fig. 4. Ionosphere data set trained neural network pruning.

vector of weights for the arcs connecting the input layer and the m -th hidden unit, $m = 1, 2, \dots, h$. v^m is a C -dimensional vector for the weight connecting the m -th hidden unit and the output layer. The activation function is a sigmoid function with domain $(-1, +1)$:

$$\sigma(y) = \frac{1}{1 + e^{-y}} \quad (5)$$

Finally, for all our weights we applied weight decay factor 0.0001. This is a quite simple approach in comparison with others described in the theoretical part, but still it does the job. Cross-entropy has been chosen as it is capable of dealing with problems of error derivative plateau better than a standard round mean square error (RMSE) [25]. Apart from this, we used Stochastic Gradient Descent batch training. Batch size was chosen to be 20.

IV. EXPERIMENTS

In our experiments, we used three 10-fold cross-validation, but for some test-sets like monk's train and test data are already provided; thus, there we utilized thirty runs to get averaged results. We decided to use two hidden layer neural networks, so that some networks would be able to use an additional layer, in case one of the layers was not needed we would be able to see this after pruning would be finished – such layers should have small number of intact neurons in one of the layers. For our experiments, we used well-known UCI [24] data sets: Monks-1, Monks-2, Monks-3, Ionosphere, Haberman, Pima diabetes, WDBC, WPBC and Parkinsons.

Some of the mentioned problems use only categorical variables – like monks. In such cases, we transformed input data into binary format, thus instead of 5 inputs we used 17. In other cases, the only transformation applied was rescaling of data into $[-1, 1]$ region. Data sets are binary classification problems; we utilized two output neurons to represent a solution to the network. Data sets themselves are pretty small ranging from about 150 to 500 entries. Table I holds classification accuracy for all data sets. It contains average classification rate with standard deviation for both train and test cases along with 10 x-validation folds or runs for non-pruned and pruned neural networks. Last table column holds network hidden layers structures before and after pruning.

For the pruning algorithm we used 0.025 as an *errorRiseTol* tolerance level, usually around 50 (depends on the total number of neurons, it should be around 60–70 % of that) as *maxIter* iterations. MaxIter count should be larger than the maximum number of neurons to be pruned (which was always equal to neuron count in hidden layers minus 2 – we could not prune all neurons from all layers). In all cases, we decided to apply 2 layer hidden neuron networks (with error-backpropagation) trained using cross-entropy cost (error) function and stochastic gradient descent as a learning algorithm. All cases were executed using 10-fold cross-validation except monk's data sets – they had already been divided into training and testing sets.

Figs. 1–4 represent a pruning process. Looking at a cross-entropy train error one can notice spikes upon neuron removal. In majority of cases, a network is rather quickly adapting to

effects of pruning. In cases when a train error rises significantly – a neuron is not pruned. All in all, training set classification graphs show decrease of train set classification error.

V. DISCUSSION

As one can notice in many cases the acquired ANN models are significantly smaller than initial networks. Smaller networks are less prone to overfitting, thus have higher generalization abilities. In some cases, networks were not significantly simplified; these were: WDBC and Parkinsons data sets, where we could see only ~50 % drop in neuron counts. Both of these data sets have a rather complex structure and thus require bigger models (in comparison with other data sets). As we have already noted, the algorithm has auto-stopping criteria allowing it to perform several trials before deciding to stop. The used algorithm assumes training of neural network with the removed neuron, while afterwards in case of unsatisfactory results the removed neuron is returned back to the ANN. Diligent reader can note that there are several possibilities in regards to how and which neuron should be returned back into a neural net. We used the same neuron, but we did not explore possibilities of adding a random neuron.

VI. CONCLUSION

In the current paper, we presented the algorithm for pruning artificial neural networks along with experimental data (UCI classification data sets were used), showing that it was able to significantly simplify a network structure, if a data set did not have a complex structure. Algorithm with the proposed stopping criteria applied to UCI datasets in many cases is giving much simpler ANN models with only a few neurons while having slightly worse or in some cases better classification accuracy rates in comparison with original non-pruned neural networks. Such “simpler” models are faster to execute and are better candidates for knowledge extraction. Further research directions are exploration of other techniques for returning a neuron back after a retraining phase.

REFERENCES

- [1] A.-Krizhevsky, I. Sutskever, G. Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”, *Advances in Neural Information Processing Systems* 25, NIPS, 2012.
- [2] G. Hinton, L. Deong, D. Yu, G. Dahl and others, “Deep Neural Networks for Accoustic Modelling in Speech Recognition”. *IEEE Signal Processing Magazine*, Nov., 2012.
<http://dx.doi.org/10.1109/MSP.2012.2205597>
- [3] X. Qiang, G. Cheng, Z. Wang, “An Overview of Some Classical Growing Neural Networks and New Developments”, *IEEE, Education Technology and Computer (ICETC), 2nd International conference*, vol. 3. 2010.
- [4] V. Chaudhary, A. K. Ahlawat, R. S. Bhatia, “Growing Neural Networks using Soft Competitive Learning”. *International Journal of Computer Applications* (0975-8887) vol. 21, no. 3, May 2011.
- [5] R. Reed, “Pruning Algorithms – A Survey”, *IEEE Transactions on Neural Networks*, vol. 4., no. 5., Sep. 1993.
<http://dx.doi.org/10.1109/72.248452>
- [6] M. C. Mozer and P. Smolensky, “Skeletonization: A Technique for Trimming the Fat From a Network via Relevance Assessment,” in

- Advances in Neural Information Processing*, pp. 107–115, Denver, 1989.
- [7] B. E. Segee and M. J. Carter, "Fault Tolerance of Pruned Multilayer Networks," in *Proc. Int. Joint Conf. Neural Networks*, vol. 2, Seattle, pp. 447–452, 1991.
 - [8] E. D. Karnin, "A Simple Procedure For Pruning Back-Propagation Trained Neural Networks," *IEEE Trans. Neural Networks*, vol. 1., no. 2, pp. 239–242, 1990. <http://dx.doi.org/10.1109/72.80236>
 - [9] R. Setiono and H. Liu, "Understanding Neural Networks via Rule Extraction," *IJCAI*, 1995.
 - [10] R. Setiono and W. H. Leow, "Pruned Neural Networks for Regression" in *PRICAI 2000 Topics in Artificial Intelligence, Lecture Notes in Computer Science*, vol. 1886, 2000, pp. 500–509. http://dx.doi.org/10.1007/3-540-44533-1_51
 - [11] Y. Le Cun, J. S. Denker, and S. A. Solla, "Optimal Brain Damage," in *Advances in Neural Information Processing (2)*, D. S. Touretzky Ed. (Denver 1989), 1990, pp. 598–605.
 - [12] B. Hassibi, D. G. Stork, G. J. Wolf, "Optimal Brain Surgery and General Network Pruning."
 - [13] Y. Chauvin, "A Back-Propagation Algorithm With Optimal Use of Hidden Units" *Advances in Neural Information Processing*, (1) D. S. Touretzky ed. (Denver 1998), 1989, pp. 519–526.
 - [14] A. S. Weigend, D. E. Rumelhart, and B. A. Huberman, "Back-Propagation, Weight Elimination and Time Series Prediction," in *Proc. 1990 Connectionist Models Summer School*, D. Touretzky, J. Elman, T. Sejnowsky, and G. Hinton, Eds., 1990, pp. 105–116.
 - [15] A. S. Weigend, D. E. Rumelhart, and B. A. Huberman, "Generalization by Weight-Elimination Applied to Currency Exchange Rate Prediction," in *Proc. Int. Joint Conf. Neural Networks*, vol. 1, (Seattle), 1991, pp. 837–841.
 - [16] A. S. Weigend, D. E. Rumelhart and B. A. Huberman, "Generalization by Weight-Elimination With Application to Forecasting," in *Advances in Neural Information Processing (3)* R. Lippmann, J. Moody, and D. Touretzky, Eds., 1991, pp. 875–882.
 - [17] C. Ji, R. R. Snapp, and D. Psaltis, "Generalizing Smoothness Constraints From Discrete Samples," *Neural Computation*, vol. 2, no. 2, 1990, pp. 188–197. <http://dx.doi.org/10.1162/neco.1990.2.2.188>
 - [18] D. C. Plaut, S. J. Nowlan, and G. E. Hinton, "Experiments on Learning by Back Propagation," *Tech. Rep. CMU-CS-86-126*, Carnegie Mellon Univ., 1986.
 - [19] S. J. Nowlan, and G. E. Hinton, "Simplifying Neural Networks by Soft Weight-Sharing," *Neural Computation* vol. 4, no. 4, 1992, pp. 473–493. <http://dx.doi.org/10.1162/neco.1992.4.4.473>
 - [20] L. Prechelt, "Adaptive Parameter Pruning in Neural Networks," *International Computer Science Institute*, Mar. 1995.
 - [21] W. Finnoff, F. Hergert, and H. G. Zimmermann, "Improving Model Selection by Nonconvergent Methods", *Elsiever Neural Networks*, vol. 6, no. 6, 1993, pp. 771–783. [http://dx.doi.org/10.1016/S0893-6080\(05\)80122-4](http://dx.doi.org/10.1016/S0893-6080(05)80122-4)
 - [22] J. K. Kruschke, "Creating Local and Distributed Bottlenecks in Hidden Layers of Back-Propagation Networks," in *Proc. 1988 Connectionist Models Summer School*, D. Touretzky, G. E. Hinton, and T. Sejnowsky, Eds., 1988, pp. 120–126.
 - [23] J. K. Kruschke, "Improving Generalization in Back-Propagation Networks with Distributed Bottlenecks," in *Proc. Int. Joint Conf. Neural Networks*, Washington DC, vol. 1, 1989, pp. 443–447.
 - [24] K. Bache, M. Lichman, (2013), UCI Machine Learning Repository [Online]. Available: <http://archive.ics.uci.edu/ml>, Irvine, CA: University of California, School of Information and Computer Science. Accessed Sept 15, 2014.
 - [25] P. Golik, P. Doetsch, and H. Ney, "Cross-Entropy vs. Squared Error Training: a Theoretical and Experimental Comparison", in *Interspeech*, pp. 1756–1760, Lyon, France, August 2013.

Andrey Bondarenko received the B.Sc. degree in 2004, and MSc. degree in 2006 from the Institute of Transport and Communication, University of Latvia, (Riga, Latvia). Since 2010 he has continued studies at Riga Technical University towards obtaining of the Doctoral degree in Computer Science. Presently his major field of study is the extraction of concise rules from a trained artificial neural network. His research interests include: computational intelligence, machine learning.
E-mail: andrejs.bondarenko@gmail.com.

Arkady Borisov received the Doctoral degree in Technical Cybernetics from Riga Polytechnic Institute in 1970 and *Dr. habil. sc. comp.* degree in Technical Cybernetics from Taganrog State Radio Engineering University in 1986.

He is a Professor of Computer Science with the Faculty of Computer Science and Information Technology, Riga Technical University. His research interests include fuzzy sets, fuzzy logic, computational intelligence and bioinformatics. He has more than 235 publications in the field.

He is a member of IFSA European Fuzzy System Working Group, Russian Fuzzy System and Soft Computing Association, honorary member of the Scientific Board, member of the Scientific Advisory Board of the Fuzzy Initiative Nordrhein-Westfalen (Dortmund, Germany).
Address: 1 Kalku Str., Riga, LV-1658; phone: +371 6708953
E-mail: arkadijs.borisovs@cs.rtu.lv.

Andrejs Bondarenko, Arkadijs Borisovs. Mākslīgo neironu tīklu vispārināšana un vienkāršošana, izmantojot samazināšanu

Šobrīd mēs piedzīvojam otro mākslīgo neironu tīklu (MNT) renesansi. Iemesls – veiksmes apmācības nozarē. Pašlaik kļūvis iespējams apmācīt daudzslāņu tīklus, izmantojot lielus datu apjomus. Tomēr, tāpat kā iepriekš, MNT izmantošanu ierobežo tas, ka dotais modelis ir melna kaste, kura nepaskaidro, kā notiek klasifikācija, kādi faktori un kāda veidā ietekmē klasifikācijas rezultātu. Vēl viena problēma, izmantojot MNT, ir tīkla hiperparametru atlase. Viens no tiem – tīkla arhitektūra ievērojami ietekmē MNT vispārināšanas iespējas. Lai zināšanas izvilkto no tīkla (likumu iegūšana) un vienkāršotu tīklu, kā arī lai palielinātu vispārināšanas iespējas, ir divas pieejas: apmācīt ļoti lielu tīklu, un pēc tam to vienkāršot, vai pakāpeniski apmācīt mazu tīklu, palielinot neironu skaitu. Šis raksts sniedz pārskatu par MNT samazināšanas metodēm. Ir dots pārskats par šādām samazināšanas metodes klasēm: jutīguma analīzes metode, neironu svaru soda metode, neironu svaru pakāpeniska samazināšana, interaktīva samazināšana, automātiska samazināšana u. c. Šajā rakstā tiek ierosināts un apskatīts algoritms, kurš samazina tīklu. Testiem ir izmantoti dati no UCI glabātuves. Metode balstās uz neironu izņemšanu un, ja vajag, atdošanu tīklā. Izmantotais neironu tīkls – daudzslāņu perceptrons – apmācīts svaru samazināšanai. Algoritmam ir automātiskās apstāšanās kritērijs. Iegūtie rezultāti liecina, ka atkarībā no datu sarežģītības, samazinātais un vienkāršotais MNT var būt ievērojami mazāks un vispārīgāks, nekā oriģinālais tīkls. Bieži vienkāršošana samazina klasifikācijas kļūdu. Iegūtais MNT visos gadījumos ir mazāks par oriģināliem tīkliem un to var izmantot noteikumu iegūšanai.

Андрей Бондаренко, Аркадий Борисов. Обобщение и упрощение искусственных нейронных сетей через урезание

В данный момент мы наблюдаем второй ренессанс искусственных нейронных сетей (ИНС). Причина этого – успехи в области глубокого обучения. На данный момент стало возможным обучение многослойных сетей на больших объемах данных. Однако, как и прежде, ограничением к использованию ИНС является то, что данный вид модели является черным ящиком, не дающим объяснения, как проводится классификация и какие факторы и как влияют на результат классификации. Другой проблемой при использовании ИНС является подбор гиперпараметров сети. Один из них – архитектура сети значительно влияет на обобщающие возможности ИНС. Как для формализации модели (извлечение правил), так и для упрощения сети и частично увеличения обобщающих возможностей проводят обучение излишне большой сети с последующим урезанием, либо обучение маленькой сети с последующим добавлением нейронов. В данной статье приведен обзор методов урезания. Рассмотрены такие подходы, как: анализ чувствительности, методы на основе штрафа, методы распада весов нейронных связей, интерактивного урезания, автоурезания и другие. В данной статье предложен и рассмотрен алгоритм урезания сети на примере данных из UCI репозитория на основе удаления и, по необходимости, возврата нейронов в сеть. Нейронная сеть – многослойный перцептрон, обученный при помощи обратного распространения ошибки. Алгоритм обладает критерием остановки. Полученные результаты говорят о том, что в зависимости от сложности данных модель ИНС может быть значительно упрощена и обобщена. Зачастую урезание уменьшает классификационную ошибку. Полученные ИНС во всех случаях становятся меньше оригинальной сети и могут быть использованы для извлечения правил.

ITIL Self-assessment Approach for Small and Medium Digital Agencies

Zigurds Binders¹, Andrejs Romanovs^{2, 1, 2} *Riga Technical University*

Abstract – In a growing world of Information Technology, entrepreneurs, start-ups and SMEs (small and medium-sized enterprises) require a good mechanism of delivering a good service. Information Technology Infrastructure Library (ITIL) is identified to be the best and most popular “best practices” framework. However, implementing ITIL not only is very difficult but there are also no best practices for implementing it. As a result, ITIL implementations are usually long, expensive and risky. This paper proposes an approach to identify and assess small business specific ITIL implementation and provide a roadmap for improvement based on the assessment results, business needs and guidelines. The proposed guidelines were used practically to assess and implement ITIL processes in a real-world digital agency.

Keywords – Digital agencies, IT service management (ITSM), ITIL, maturity model, self-assessment, service lifecycle, SMT.

I. INTRODUCTION

In today's IT world, businesses rely on IT to enable them to achieve their companies' vision, goals and business strategy.

Organizations use the IT for various reasons but mainly to:

- Revolutionize the way they operate, communicate and do business;
- Develop and innovate, gain market advantage and differentiate themselves to their end customers;
- Drive increased productivity and efficiency, improve business processes, make cost savings, and increase sales and growth;
- Communicate with larger, more global marketplace [1].

All of the above-mentioned reasons fully correspond to the main use of IT by the digital agency, where its quality of IT is reflected in its reputation and brand, and has direct impact upon sales and revenue. In order to achieve benefits from the IT investment, the resulting IT service must be well planned, designed, managed and delivered. This is the key of a good IT service management.

IT service management has many definitions, but the one that is more specific is the following.

IT service management is the implementation and management of quality IT services that meet the needs of the business. IT service management is performed by IT service providers through an appropriate mix of people, processes and information management [2].

Providing a good IT service management is essential to gain business benefits through IT at an agreed and controlled cost.

II. ITIL SERVICE MANAGEMENT MODEL

ITIL is a number of best practices that can be applied to managing and delivering IT service management (ITSM). It focuses on aligning IT services with the needs of business. ITIL provides a practical framework for identifying, planning, delivering and supporting IT services to the business. ITIL can be described as the mixture of many concepts and frameworks to produce a roadmap [3]. In implementing ITIL, it is crucial to understand the differences among the concepts and the frameworks [4]. The processes, procedures, tasks and checklists described by ITIL are used by an enormous number of organizations to establish the integration with their strategy, value delivery and maintaining their competency level. ITIL can be a solid backing for organizations that want to improve their IT services, reduce costs, improve customer satisfaction and enhance productivity. ITIL also helps the IT service providers to identify the effective approach for specifying metrics and measuring the outcomes.

ITIL approaches service management from the lifecycle aspect of a service. The service lifecycle is an organizational model that provides insights into:

- The way service management is structured;
- The way various lifecycle components are linked to each other;
- The impact that changes in one component will have on other component and on the entire lifecycle system [5].

The ITIL service lifecycle consists of five phases (shown in Fig. 1):

1. Service Strategy;
2. Service Design;
3. Service Transition;
4. Service Operation;
5. Continual Service Improvement.

Each stage of the lifecycle also consists of a defined set of processes that have their own purpose (concept), activities and inputs/outputs. In total, there are 26 processes that are spread across the lifecycle and can be implemented separately from each other.

Service strategy delivers guidance with designing, developing and implementing service management as a strategic asset [5]. It describes principles underpinning the practice of service management, which are useful for developing service management policies, guidelines and processes across the ITIL service lifecycle. Two aspects of strategy are covered in the ITIL service strategy:

- Defining strategy whereby a service provider will deliver service services to meet a customer's business outcomes;
- Defining a strategy for how to manage those services [10].



Fig. 1. The service lifecycle [6].

Service strategy is critical in the context of all processes along the ITIL service lifecycle. The purpose of the service strategy phase is to define the perspective, position, plans and patterns that a service provider needs to be able to execute to meet an organization's business outcomes. Service strategy is located at the center of the service lifecycle, where the value creation begins with understanding organizational objectives and customer needs. The development and application of service strategy requires constant revision, just as in all other components of the cycle.

ITIL service strategy is aimed mainly at executives and managers, who are responsible for defining the strategy of a service provider.

The processes specifically addressed in the service strategy are the following:

- Strategy management for IT services;
- Service portfolio management;
- Financial management for IT services;
- Demand management;
- Business relationship management [10].

ITIL service design provides guidance for the service design stage of the ITIL service lifecycle. It deals with the design and development of services and their related processes [5]. The purpose of the service design stage of the lifecycle is to design IT services, together with governing IT practices, processes and policies, to implement the service providers strategy and to facilitate the introduction of the service into supported environments ensuring quality service delivery, customer satisfaction and cost-effective service provision. Design encompasses the whole IT organization, for it is the organization as a whole that delivers and supports the services. Service design is the stage in the lifecycle that turns a service strategy into a plan for delivering the business objectives.

The service design phase in the lifecycle begins with the demand for new or changed requirements from the customer. Good preparation and an effective and efficient infusion of people, processes, products (services, technology and tools) and partners (suppliers, manufacturers and vendors) – ITIL's four P's – are necessary if the design, plans and projects are to succeed [11].

The processes specifically addressed in the service design are the following:

- Design coordination;
- Service catalogue management;
- Service level management;
- Availability management;
- Capacity management;
- IT service continuity management;
- Information security management;
- Supplier management [11].

Service transition consists of the management and coordination of the processes, systems and functions required for the building, testing and deployment of new and changed services [5]. It provides guidelines for the development and improvement of capabilities for transitioning new and changed services into the supported environment, including release planning, building, testing, evaluation and deployment.

Service transition establishes the service as specified in the design phase, based on the customer and stakeholder requirements. The purpose of service transition stage of the service lifecycle is to ensure that new, modified or retired services meet the expectations of the business as documented in the service strategy and service design stages of the lifecycle [12].

Service transition also introduces the service knowledge management system, which can support organizational learning and help to improve the overall efficiency and effectiveness of all stages of the service lifecycle. This will help people to benefit from the knowledge and experience of others, support informed decision-making and improve the management of services.

The processes specifically addressed in the service transition are the following:

- Transition planning and support;
- Change management;
- Service asset and configuration management;
- Release and deployment management;
- Service validation and testing;
- Change evaluation;
- Knowledge management [12].

Service operation involves coordination and carrying out activities and processes required to provide and manage services for business users and customers within a specified agreed service level [5]. Service operation is also responsible for management of the technology required to provide and support the services.

Service operation is a critical stage of the service lifecycle. Well-planned and well-implemented processes will be of no avail if the day-to-day operation of those processes is not properly conducted, controlled and managed. Nor will service improvements be possible if day-to-day activities to monitor

performance, assess metrics and gather operational data are not systematically conducted during service operation.

The processes specifically addressed in the service operation are the following:

- Event management;
- Incident management;
- Request fulfillment;
- Problem management;
- Access management [13].

IT departments must continually improve their services in order to remain appealing to the business. This is placed within a lifecycle phase of Continual Service Improvement (CSI) [5].

The purpose of CSI stage of the lifecycle is to align IT services with changing business needs by identifying and implementing improvements to IT services that support business processes [14]. These improvement activities support the lifecycle approach through a service strategy, service design, service transition and service operation. Feedback from any stage of the service lifecycle can be used to identify improvement opportunities for any other stage of the lifecycle. CSI is always seeking ways to improve service effectiveness, process effectiveness and cost effectiveness.

In order to identify improvement opportunities, the measurement of current performance is an important factor. Despite the performance measurement, CSI also measures and monitors the following matters:

- Process compliance;
- Quality;
- Business value of a process [5].

The processes/functions specifically addressed in the CSI are the following:

- Seven-step improvement;
- Service reporting (function) [14].

ITIL is the world's most widely accepted and adopted IT service management and delivery approach, where real organizations have benefited in a number of ways.

From the business point of view, adoption of ITIL practices ensures many benefits, such as:

- IT services which align better with business priorities and objectives, meaning that the business achieves more in terms of its strategic objectives;
- Known and manageable costs, ensuring the business better plans its finance;
- Increased business productivity, efficiency and effectiveness, because IT services are more reliable and work better for the business users;
- Financial savings from improved resource management and reduced rework;
- More effective change management, enabling the business to keep pace with change and drive business change to its advantage;
- Improved user and customer satisfaction with IT;
- Improved end-customer perception and brand image [1].

Regardless of how big the organization is, whichever industry it belongs to or how many layers of management it consists of, there is always a space for improvement from ITIL.

III. FITSM – STANDARD TO LIGHTWEIGHT IT SERVICE MANAGEMENT

In order to assess the level of ITIL maturity within the organization, it is necessary to go through all 26 processes described in ITIL. It would be very time consuming and useless to go through all of them as they might not be relevant at all or might not even exist within a small or medium organization.

FitSM is a standard family aimed to facilitate lightweight service management implementation. FitSM is designed to be compatible with the International Standard ISO/IEC 20000-1 (requirements for service management) and the ITIL [7].

FitSM introduces several documents, providing guidance and input on ITIL, like overview and vocabulary, requirements, objectives and activities, role model, maturity and capability assessment schema etc. For the purpose of this implementation, only three are going to be used – process specific requirements document, process definition template, and maturity and capability assessment schema. The requirements document sets a minimum set of requirements for effective ITSM. This is intended to be achievable and useful in a small digital agency that may have existing but not entirely defined or implemented ITIL processes. The process specific requirements document targets an overall ITIL maturity and is composed of 14 ITIL processes from across all five service lifecycle phases. The processes are the following:

- Service Strategy
 - Service Portfolio Management
 - Business Relationship Management
- Service Design
 - Service Level Management
 - Service Availability & Continuity Management (FitSM describes Availability and IT Service Continuity Management as one process)
 - Capacity Management
 - Information Security Management
 - Supplier Management
- Service Transition
 - Service Asset and Configuration Management
 - Change Management
 - Release and Deployment Management
- Service Operation
 - Incident Management
 - Problem Management
- Continual Service Improvement
 - Service Reporting (function)
 - Seven-step Improvement.

The process definition template will facilitate the documentation and organization of ITIL processes. It will capture such relevant information as goals and objectives, role model, activities, key performance indicators (KPI), potential service management tools etc. These 14 documents will be

kept as a source of truth and reference during the assessment process. In capability assessment schema it is considered that there are 6 capability levels (0 to 5) as described in ISO/IEC 15504 standard but only levels 0-4 will be described. This is done as only levels 0-3 are considered in the implementation scenarios FitSM addresses. Level 4 is included in the assessment schema largely to show the difference from level 3. Within the scheme we have two kinds of requirement: Tasks/activities are things that are done while output/achievement are things, which are produced.

IV. EXPERIMENTS

ITIL self-assessment within a small or medium-sized digital agency requires a new, efficient, easy-to-use and lightweight approach. The purpose of this study is to describe and apply the approach to a live digital agency.

The approach to ITIL process self-assessment usually covers three key areas on which ITSM is implemented, namely: people, processes and technology [9].

In this study, the main focus will be on processes and technology.

The ITIL processes will be assessed using the approach introduced by Tudor IT Process Assessment (TIPA) framework [8]. The framework describes how to document the assessment that includes individual process SWOT analysis, analyze the results and produce a road map for ITIL process adoption and next steps. The maturity level will be defined using the FitSM approach of a standard ISO/IEC 15504 scale and process criteria against a reference model that is defined in terms of purpose, expected results, base practices, inputs and outputs. To identify the extent to which the process is performed, achieves its purpose and produces its outcomes, a workshop will be conducted to go through each process assessment criteria. All of the above-mentioned activities will help to facilitate the assessment and have a documented result of each assessed process and see where its strengths, weaknesses, opportunities and threats are.

While carrying out the ITIL process maturity assessment, it is also important to look at the process assessment from a technological point of view.

Carrying out each management process in scope of ITSM may require the concurrence of appropriate service management tools (SMT), which can range from documentation describing the process till sophisticated software that assists the service provider. [9] The lack of these tools may limit the value of maturity of ITSM or even make it not possible. The assessment process will follow the framework introduced by FedSM organization. The idea is to assess the identified 14 service management processes by appointing them to one of three levels:

- Level 0: No tools;
- Level 1: General tools;
- Level 2: Specific tools [9].

All three levels have different criteria for every process that needs to be assessed. Levels that are defined in this way are helpful for the organizations that provide service to make a decision to improve the tool support in case it is needed as

well as facilitating the maturity of service management within the organization.

The analysis of both process maturity and service management tool assessment will allow identifying if there is a mismatch between ITSM process and the technology used to facilitate it and produce a more accurate process assessment.

The overall ITIL process assessment plan is the following:

- Get the organization acquainted with the service management processes that are going to be assessed and the methods used;
- Perform the assessment of a target level of each process maturity using the workshop approach;
- Perform assessment using the TIPA approach and FitSM process maturity model and predefined criteria;
- Perform SMT assessment using FedSM tool maturity baseline assessment framework;
- Analyze both assessment results to identify processes that need to be improved/implemented;
- Define the improvement plan based on the analysis results.

The organization will be introduced to the ITSM and the processes being assessed via a handout that will contain the description of each process and the overall approach to the assessment. The process description will contain the following:

- Introduction and overview;
- Goals and objectives;
- Inputs/Outputs;
- Activities;
- Key performance indicators.

The next step is to have a workshop and perform the assessment of a target maturity level of each process that will help to identify the processes that require improvement or need to be implemented later on. Following the description of each level of process maturity, the target level will be identified and recorded as described in Table I below, showing the target level for two service strategy processes.

FitSM describes maturity level criteria for all 14 processes that will be used during the assessment. To identify to which extent a particular level has been achieved, the process will be rated using a scale shown below:

- “Fully” achieved: Achieved between 86 % and 100 %
- “Largely” achieved: Achieved between 51 % and 85 %
- “Partially” achieved: Achieved between 16 % and 50 %
- “Not” achieved: Achieved between 0 % and 15 %

The percentage will be calculated based on the process criteria. The representation of the assessment results will be captured in a table. See Table II as an example. The table presents assessed processes horizontally, levels and attributes vertically.

The results of the assessment will be shown in a spider diagram (Fig. 2). The example diagram shows the result of only 5 processes out of 14.

TABLE I
TARGET LEVEL ASSESSMENT

		Level 1 Ad-hoc/Initial	Level 2 Repeatable/Partial	Level 3 Defined/Complete	Level 4 Managed/Aligned
Service Portfolio Management	Target	F	F		
Business Relationship Management	Target	F	F	F	

TABLE II
PROCESS PROFILE

		Level 1 Ad-hoc/Initial	Level 2 Repeatable/Partial	Level 3 Defined/Complete	Level 4 Managed/Aligned
Service Portfolio Management	Assessed	Largely	Largely	Partially	Not
Business Relationship Management	Assessed	Fully	Largely	Partially	Not

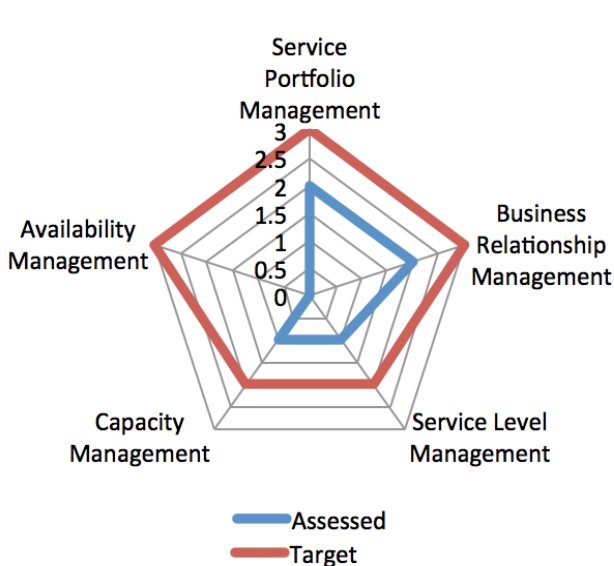


Fig. 2. Results of maturity assessment.

Each radial arm of the web represents a service management process. The diagram shows strengths in some areas and weaknesses in other as well as the comparison between the assessed and target level.

Similar approach will be taken to perform SMT assessment, but without assessing target levels. Assessing SMT and producing a spider graph will help to identify processes with a good and poor set of SMTs to perform the process phases (Fig. 3).

Analysis of both assessments will cover the explanation of maturity and SMT level achieved, process inputs and outputs, process SWOT analysis, recommendations and expected outcomes.

The final step of the assessment approach is to define the improvement or implementation plan. Based on the results of the assessment, the organization (together with all those involved) has to prioritize the order in which processes will be improved or implemented and put in place a high-level plan (which usually is a typical project plan).

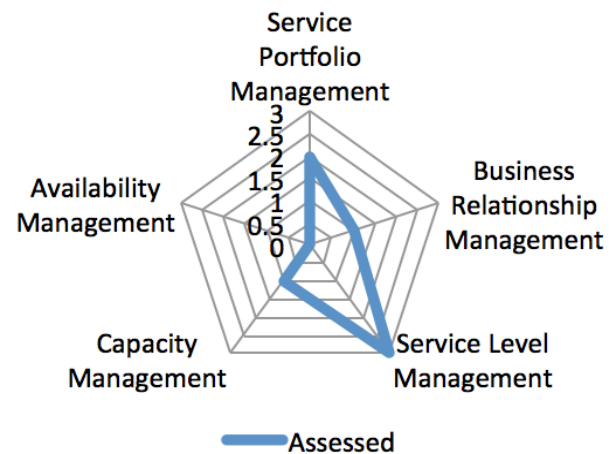


Fig. 3. Results of SMT assessment.

The implementation/improvement of a particular process needs to realize the following steps:

- To review process objectives/outcomes;
- To write process policy;
- To define process triggers;
- To define process KPIs;
- To defining process roles and responsibilities;
- To designing process flow;
- To write process procedures;
- To define SMTs (Service Management Tools).

V. CONCLUSION

The described approach will be implemented in a real life small digital agency. While conducting the assessment, all the necessary documentation and notes will be taken for post-assessment analysis, to identify if the approach needs improvement.

The results of this approach will allow the organization to identify the service management processes that need to be improved or implemented and the method of how to actually accomplish that. The approach itself can be used by any small IT organization (SME) that would like to improve or implement ITIL service management and benefit from it. The

future work will be focused purely on the implementation/improvement of the identified processes based on the recognized obstacles to the target levels that organization has indicated. The next challenge would be to focus on the third area of ITSM – people. Dealing with people involves performing an organizational change, plan to involve people and communicate effectively in order to finalize the ITIL implementation.

REFERENCES

- [1] M. Kneller, "Executive Briefing: The Benefits of ITIL," The Stationery Office, U.K., 2010.
- [2] P. Farenden, *ITIL for Dummies*, England: John Wiley & Sons, Ltd, 2012.
- [3] H. Marquis, "ITIL: what it is and what it isn't," *Business Communication Review*, vol. 36, no. 12, pp. 49–52, Dec. 2006.
- [4] A. A. Latif, M. M. Din and R. Ismail, "Challenges in adopting and integrating and CMMI in ICT Division of a public utility company," *2nd International Conference on Computer Engineering and Application*, Mar. 19–21, pp. 81–86, 2010.
<http://dx.doi.org/10.1109/ICCEA.2010.279>
- [5] J. van. Bon, *ITIL V3 A Pocket Guide*, Van Haren Publishing, 2011. p. 164.
- [6] R. Ryan, T. Raducha-Grace, *The Business of IT: How to Improve Service and Lower Costs*. U.S: International Business Machines Corporation, 2010.
- [7] The FitSM Standard Family: Standard for lightweight IT service management [Online]. Available: <http://www.fedsm.eu/fitm>. Accessed: 26.09.2014.
- [8] About TIPA: Assess And Improve Your IT Processes [Online]. Available: <http://www.tipaonline.org/en/tipa>, Accessed: 26.09.2014.
- [9] T. Szeplieniec, Service management tools implementation and maturity baseline assessment framework: FedSM, 2013.
- [10] The Stationery Office, ITIL Service Strategy: 2011 Edition. Information Technology Infrastructure Library 2011, Norwich, U.K.: The Stationery Office, 2011.
- [11] The Stationery Office, ITIL Service Design: 2011 Edition. Information Technology Infrastructure Library 2011, Norwich, U.K.: The Stationery Office, 2011.
- [12] The Stationery Office, ITIL Service Transition: 2011 Edition. Information Technology Infrastructure Library 2011, Norwich, U.K.: The Stationery Office, 2011.
- [13] The Stationery Office, ITIL Service Operation: 2011 Edition. Information Technology Infrastructure Library 2011, Norwich, U.K.: The Stationery Office, 2011.
- [14] The Stationery Office, ITIL Continual Service Improvement: 2011 Edition. Information Technology Infrastructure Library 2011, Norwich, U.K.: The Stationery Office, 2011.

Zigurds Binders, B. sc. ing., is a Master Student at the Information Technology Institute, Riga Technical University (Latvia). He received Bachelor degree in Information Technology from Riga Technical University in 2011. His research interests are the project and service management and their governance, agile methodologies and web/software development. Since 2013, he has been working as a Project Manager at "Zoocy Consulting Limited" (United Kingdom). Before he worked at "Opus Capita" (Latvia) as a Change Manager/Developer.
E-mail: zigurds.binders@rtu.lv

Andrejs Romanovs, Dr. sc. ing., is an Associate Professor and Leading Researcher at the Information Technology Institute, Riga Technical University. He has 25 years of professional experience teaching post-graduate courses at RTU and developing more than 50 industrial and management information systems. His professional interests include modeling and design of management information systems, IT governance, IT security and risk management, integrated information technologies, as well as education in these areas. A. Romanovs is a senior member of the IEEE, LSS; the author of more than 50 papers in scientific journals and conference proceedings in the field of Information Technology.
E-mail: andrejs.romanovs@rtu.lv

Zigurds Binders, Andrejs Romanovs. ITIL pašnovērtējuma pieeja mazām un vidējām digitālām aģentūrām

Informācijas tehnoloģiju pasaulē, jauniem uzņēmējiem un MVU (maziem un vidējiem uzņēmumiem) ir nepieciešams labs mehānisms, lai spētu nodrošināt labu pakalpojuma piegādi. Uzņēmumiem, kas ignorēs IT pakalpojumu vadības principus, būs arvien grūtāk attaisnot ieguldījumus IT un konkurēt ar citiem tirgus dalībniekiem. ITIL (Information Technology Infrastructure Library jeb Informācijas tehnoloģiju infrastruktūras bibliotēka), tiek uzskatīta par vislabāko un vispopulārāko "labo prakšu" kopu, ko var lietot IT pakalpojumu pārvaldības nodrošināšanai. Tas ir atvērts procedūru un ieteikumu apkopojums, kurā apkopotas zināšanas un pieredze, kas uzkrāta gadu gaitā dažādu iestāžu un organizāciju iekšienē. ITIL ieviešana ir grūts un laikietilpīgs darbs, pie tam vēl neeksistē "labo prakšu" ITIL ieviešanā. Tādējādi, ITIL ieviešanas process ir dārgs un riskants pasākums. Šajā rakstā tiek aprakstīta pieeja, ko var pielietot mazajos un vidējos uzņēmumos, lai novērtētu ITIL brieduma līmeni un sniegtu ITIL procesu uzlabojumu un/vai ieviešanas plānu. Pieveca balstās uz divām no trim galvenajām IT pakalpojumu pārvaldības jomām – procesiem un rīkiem. Ir novērtēti tikai 14 ITIL procesi, kurus apraksta FitSM un tika izmantota pieeja, ko apraksta TIPA un FedSM, kas par pamatu izmanto ISO/IEC 15504 standartu, bet ar nelielām izmaiņām kritēriju un brieduma līmeņu aprakstā. Pakalpojumu pārvaldības rīku novērtēšanā tiks pielietota līdzīga pieeja, kas par pamatu izmanto FedSM definētos kritērijus un līmeņus. Aprakstītā pieeja tiks pielietota reālā digitālā aģentūrā, izvērtējot un īstenojot ITIL procesus. Turpmākais darbs būs identificēto ITIL procesu implementācija vai uzlabošana, kā arī aģentūras darbinieku kā organizatorisko pārmaiņu elementu izpēte.

Зигурдс Биндерс, Андрей Романов. Подход самооценки ITIL для малых и средних цифровых агентств

В мире информационных технологий новым предпринимателям и малым и средним предприятиям необходим адекватный механизм для обеспечения поставки услуг хорошего уровня. Предприятиям, которые будут игнорировать принципы управления ИТ услугами, будет все труднее оправдать инвестиции в ИТ и конкурировать с другими участниками рынка. ITIL (Инфраструктурная библиотека информационных технологий) является наилучшим и наиболее популярным подходом использования «лучших практик», который может быть применен для обеспечения управления ИТ услугами. ITIL является открытым обобщением процедур и рекомендаций, в котором собраны знания и опыт, полученные за многие годы в различных организациях и учреждениях. Внедрение ITIL является тяжелой и трудоемкой работой, к тому же еще не существует «лучших практик» для внедрения ITIL. Следствием этого является высокая стоимость и риск всего процесса внедрения ITIL. В данной статье описан подход, который может быть применен малыми и средними предприятиями для оценки уровня зрелости ITIL и для составления плана улучшения и/или внедрения процессов ITIL. Подход основан на двух из трех главных областях управления ИТ услугами – на процессах и средствах. Оценены только 14 процессов ITIL, которые описываются FitSM, и использован подход, описанный в TIPA и FedSM, который как основу использует стандарт ISO/IEC 15504 с небольшими изменениями в описании критериев и уровня зрелости. В оценке средств управления услугам применяется похожий подход, за основу принимающий критерии и уровни, заданные в FedSM. Описанный подход будет применен для оценки и реализации процессов ITIL в реальном цифровом агентстве. Дальнейшая работа будет направлена на внедрение или улучшение идентифицированных процессов ITIL, а также на исследование работников агентства как элементов организационных перемен.

Scenario Development Approach to Management Simulation Games

Jana Bikovska, *Riga Technical University*

Abstract – The paper introduces a scenario development approach to management simulation games. Nowadays the high efficiency of simulation games as a training method is obvious due to their ability to help students understand the complexity of decision making within a dynamic environment. Modern computer technologies allow creating complex games with an adaptable environment that requires development of appropriate game scenarios. It is not a trivial problem that requires elaborating of a formal approach, which will be described in the paper. As a result a general scheme for scenario development and modelling for a simulation game is proposed in the paper. Some sample scenarios for logistics and supply chain games are provided in the paper as well.

Keywords – Game scenarios, logistics and supply chain game, management simulation game, scenario management.

I. INTRODUCTION

Simulation games in education are widely used as they provide realistic, experiential learning environments, thus, provide a more comprehensive way to teach concepts and theories than traditional lecturing [1]. This paper focuses on management simulation games, a subset of simulation games that provide a dynamic, uncertain and competitive business environment to train learners for management decision making. The players of such games make management-type decisions in an offered situation, and their decision choices generally affect the environmental conditions under which the subsequent decisions must be made. In a very brief time, this gives the game participants a chance to study both long and short-term results of their decisions. This provides a possibility to react to these effects and make decisions in the light of altered circumstances as in the real business. As a result, the game participants are able to gain knowledge and practice skills in the game by repeating experiments with the simulated reality through the simulation game. In addition, such games help developing such skills as teambuilding, social networking, collaborating, etc.

Recent advances in technology allow identifying several research and development directions in management simulation gaming. As indicated in [2], in the past two decades the number of Internet-based management simulation games has grown as it allows running them through a central server with administrator-selected parameters and entering participants' decisions directly to the server. Thus, management simulation games use computers-servers to run the models and to process participants' decisions. This allows creating more realistic and complex game scenarios with respect to the number of decision variables included, the number of products and markets defined, and the amount of

feedback available to the participants. Software agents embodied in management simulation games facilitate modelling and exploring possible scenarios in the game while controlling the game process, thus, affecting the environment and direction of the simulation [3].

Modern management simulation games should possess the following features: realism, accessibility, compatibility, flexibility and scale, simplicity of usage, decision support systems, and communication [2]. This paper focuses on enhancing realism, flexibility and scale of management simulation game by proposing a general scheme for the game scenario development.

The objectives of this paper are: (1) to introduce a scenario concept for management simulation games; (2) to propose a general procedure for management simulation game scenario development; (3) to provide examples of scenario development for ILMG and ECLIPS management simulation games.

II. GAME SCENARIO CONCEPT

Scenario approach is widely used in different fields, for instance, forecasting, optimisation etc. There are different views on the definition of scenario, depending on the context in which the term is used. This term appears in a wide variety of fields, ranging from state administration, information systems, requirement engineering to human computer interaction. Interpretation of scenarios seems to depend on their usage and how they are generated. In general, a scenario can be defined as a set of possible sequences of future events [4], [5]. There are different approaches to scenario development: formal, semiformal, and informal, based on experts' opinion [6]. The paper concentrates on a semiformal approach as it includes automatic procedure of scenario generation adjusted by an expert.

The game simulation could be performed according to a predefined scenario chosen (or specially elaborated) by the game manager (instructor). Scenario plays a very important role within the game structure and it has to be formalized for further implementation.

With a respect to management simulation games, a scenario can be conceived as a means of transforming the game initial state to a final state, following main development trends, which are influenced by both internal events and external activities. It could be presented by the 5-tuple [2] as it is shown in (1):

$$\langle SI, T, E, A, SF \rangle \quad (1)$$

where:

SI – the initial state of the game;

T – predefined game business environment development trends;
 E – predefined internal events;
 A – performed external activities;
 SF – game final state once the scenario has terminated.

The generic scenario structure is shown below (see Fig. 1).

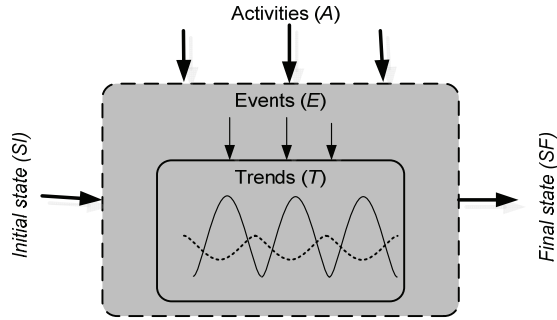


Fig. 1. Generic scenario structure of management simulation game.

Trends T_i introduced to a game scenario dynamics over time t are defined by different functions, e.g. (2):

$$T_i = f_i(t), T_i \in T \quad (2)$$

Patterns of these trends depend on the type of function and its parameters which can be changed by the game manager.

All internal and external influences depend on the current game state and they are aimed at achieving a particular result of the game functioning or predefined learning objectives. Any internal and external influences lead to short-term or long-term changes in the pattern of trends, i.e., (3).

$$f_i^E : E \rightarrow T_i \text{ and } f_i^A : A \rightarrow T_i, \quad (3)$$

where:

$E = \{e_1, e_2, \dots, e_n\}$ is a set of internal events;

$A = \{a_1, a_2, \dots, a_n\}$ is a set of external activities.

Game initial state is defined by the proposed business environment characteristics and structure as well as by value of the trends' functions (4) at the moment when the game session starts:

$$SI = T_i(t_0), \quad (4)$$

where t_0 defines the initial time instant.

Thus, a scenario provides a flexibility of management simulation games as allows a game manager to define the initial state of game business environment, to determine the development of the game situation through time according to predefined trends, as well as preset the list of important events taking place during simulation. Events can affect trends. Participants make their decisions (or in other words, take actions) that can affect both trends and events. Trends may be configured according to the learning objectives by changing appropriate parameters available in the game.

Once all of the game scenario elements are defined, the scenario development procedure should be applied.

III. SCENARIO DEVELOPMENT APPROACH

The scenario development for management simulation games becomes a complex process as the scenarios have to reflect real life situations [2], in other words, the number of a scenario parameters that define game initial state, game business environment development trends, internal and external events can be very large. In this situation, manual scenario development would be not so efficient. This makes it necessary to propose an advanced scenario development procedure (see Fig. 2).

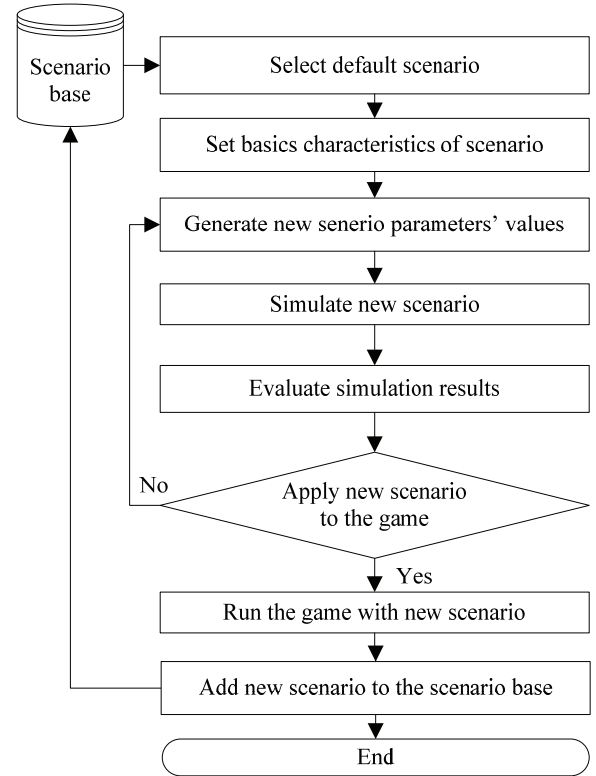


Fig. 2. Scenario development procedure for management simulation games.

Here, a scenario base represents a set of previously developed scenarios that can be used as a base for further scenario development. Game manager chooses a default scenario and sets basic scenario characteristics, thus defining initial state of the game. Then specially elaborated tools are applied to generate new values of scenario parameters that define game business environment development trends, internal and external events.

After the scenario is generated, it is necessary to simulate it with the game. It can be done manually or automatically. As the manual simulation supposes real players to test the game scenario, it would take more time and efforts. Automatic scenario simulation has several advantages: (1) the scenario can be simulated in a very short time period; (2) several replications can be made for more efficient scenario evaluation. To simulate the developed game scenario, special software agents are proposed [7], [8]. They act as the game participants by performing activities related to the company development under the proposed scenario conditions [4].

Depending on the type of the game, several agents can be running simultaneously.

When the game results are obtained, a game manager has to evaluate a new scenario. If it matches learning objectives, and it is proved that certain game results can be achieved, then the scenario could be applied to the regular game session and added to a scenario base. Scenario from the scenario base can also be applied to the game without any changes.

The proposed procedure can be applied to scenario generation of different management simulation game types.

IV. ILLUSTRATIVE EXAMPLES

To provide the examples of scenario development, two management simulation games were selected, namely, International Logistics Management Game and ECLIPS game.

A. International Logistics Management Game

As an example of a management simulation game for introducing scenario development and modelling approach, the International Logistics Management Game (ILMG) was selected [9], [10]. The game is aimed at providing a virtual business environment for training decision-making skills in the area of international logistics, production management, finances etc. The game can be classified as a total enterprise (or general management), interacting, computer-aided and internet-based, real-time processed business simulation. ILMG software functioning according to the structure presented in Fig. 3.

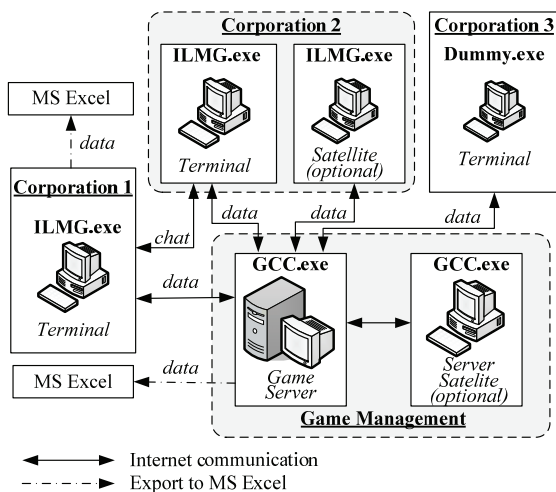


Fig. 3. ILMG software structure.

Different scenarios in the game are introduced to simulate real world situations and force trainees to acquire skills and experience in managing different functions of the company in various situations. During the game, participants operate as different corporations within an international market, producing goods or services (i.e., making strategic and operational decisions) and competing with each other in order to improve their corporation performance and to achieve the

following goals: to earn profit, to capture substantial market share and to achieve high customer satisfaction.

The ILMG scenario is generated by the game manager according to specific learning objectives. Initial conditions of the game and corporations' state are set up by changing different parameters that can be divided into several groups (see Fig. 4): regional parameters (e.g., unit projecting cost, basic wage level), regional-related production parameters (e.g., productivity parameters, overhead parameters), technical coefficients of products (e.g., material per unit, setup cost), market related parameters (e.g., price effect parameters), financial parameters (e.g., prime rate, tax rate) etc.

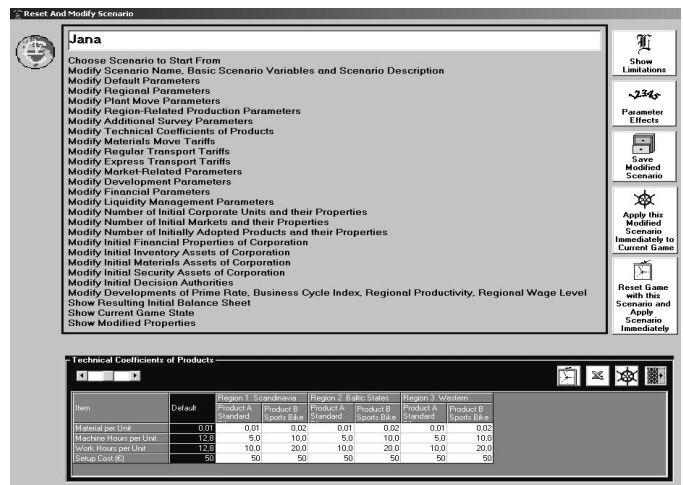


Fig. 4. ILMG scenario development screen.

Most of the parameters have to be defined before the game starts and they cannot be changed afterwards. However, some parameters (i.e., prime rate, business cycle index of the region, regional productivity index, and regional wage level) can be easily edited during the game that allows changing a business environment to simulate different situations.

In ILMG, the game manager has to set over than a hundred different parameters for the simplest scenario (one region, one product, two bank accounts). To support this task, special stand-alone application GameEditorPro was developed (see Fig. 5). It generates a full set of the game scenario parameters according to guidelines provided by the game manager.

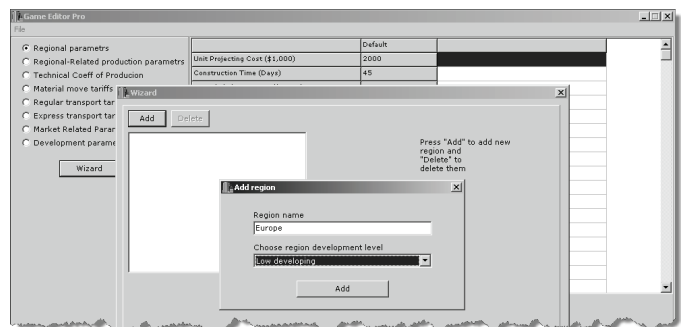


Fig. 5. GameEditorPro for ILMG interface example.

The game manager decides only on a number of regions (R) and products (P) and their characteristics; then the software automatically generates all other scenario parameters. Regions are defined with qualitative characteristics as regions with “advanced”, “less advanced” or “developing” economy. Products are classified as “hi-tech”, “medium-tech” and “low-tech” products. For example, the game manager sets up a new scenario with three regions R1=“advanced”, R2=“developing” and R3=“advanced”, and three different products such as P1=“low-tech”, P2=“hi-tech” and P3=“hi-tech”.

Then an automatic generation of scenario parameters is performed by the following algorithm:

- the total number of parameters is calculated;
- the full range of parameters values is defined;
- each defined range is divided into three equal subsets;
- each subset of parameters characterizes either one region type or one product type;
- within each subset, values of parameters are generated randomly to exclude the possibility obtaining identical regions or products.

The parameters of the scenario received from the GameEditPro module can be exported to the ILMG software. It is used to simulate and test a specific scenario. In this case, a player agent is involved and it makes strategic, marketing, production, purchasing, transport, investment and financial decisions. Finally, the game manager manually evaluates the key indicators received by the player agent and decides if they meet learning objectives such as “a positive Net profit, a customer’s fill rate is more than 50 %, and a liquidity ratio is greater than 1”. If the evaluation is negative, the procedure of scenario generation is repeated and new values of game parameters are generated. Otherwise, the generated scenario is approved and introduced in the game regular session.

For the testing purpose, the game with 3 player agents was run for 8 periods (or 2 years). These agents substituted real players of the game. The key indicators received by the agents’ players are given in Table I and presented in Fig. 6. As far as the obtained results meet the above-mentioned requirements, the tested scenario can be introduced in the game regular session.

TABLE I
KEY INDICATORS OBTAINED FROM SCENARIO MODELLING

	Agent 1	Agent 2	Agent 3
Net Profit (EUR)	80,749,580	81,865,660	90,920,780
Service Level	64%	60%	63%

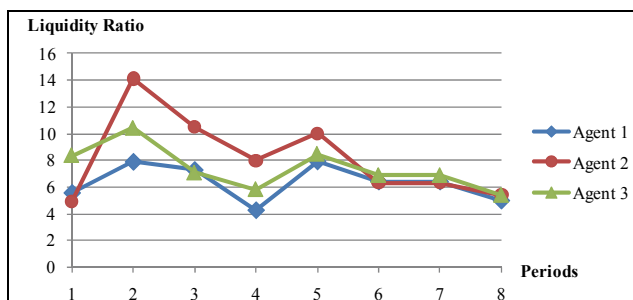


Fig. 6. Liquidity ratio of tested scenario.

Then the scenario is implemented and the game manager continuously reviews the key performance indicators of corporations represented by the human players. If they are out of range as defined by the teaching objectives, the game simulation time is fixed and a new scenario model is generated, e. g., such parameters as prime rate, business cycle index of the region, regional productivity index, or regional wage level could be adjusted. A two-year period is simulated during the regular session of a game.

B. ECLIPS Game

ECLIPS game is a board game that considers an inventory management problem in a multi-echelon supply chain, in particular, detailed workings of different inventory replenishment policies. It provides an insight into organization and functioning of multi-echelon supply chain based on continuous or periodic inventory review planning as well as demonstrates benefits and shortcomings of their implementation under different conditions [11], [12]. ECLIPS game provides flexible supply chain network modelling capability.

To support activities of the game manager and players, a special computer-aided tool was developed. It is aimed at providing environment that allows simulating different scenarios of the ECLIPS game [12]. The example of the simulated supply chain is represented in Fig. 7.

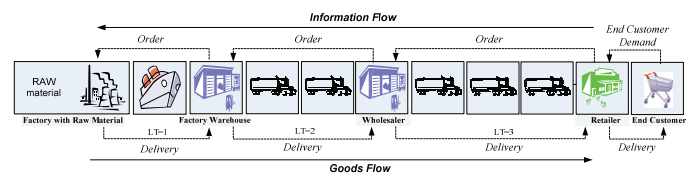


Fig. 7. Three echelon supply chain of the ECLIPS game.

The current version of the software allows testing scenarios by changing different parameters of the supply chain (see Fig. 8), e. g., end customer demand, inventory holding cost, order cost, new production cost, lead times between stock points and service level coefficient. Here, the number of stock points or supply chain echelons is limited by three as well as the end customer demand can be modelled either by one or two dice with sides defined by the game manager.

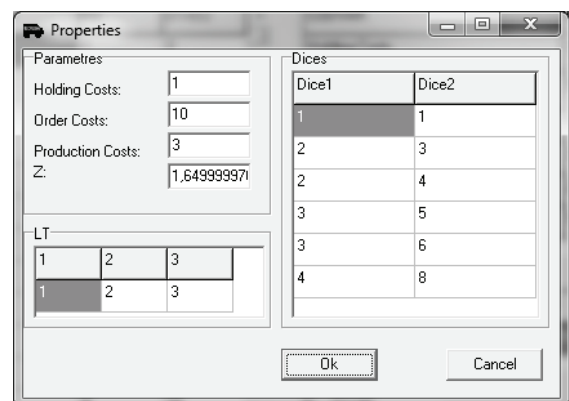


Fig. 8. Scenario parameters of the ECLIPS game.

After setting all parameters, the scenario is simulated to obtain game results. To evaluate reproducibility of the achieved results, the confidence interval method is proposed. Scenario is evaluated by three key indicators: service level (cumulative), average cost and average inventory.

The simulation results from 10 replications of 52 game periods for average inventory under a continuous replenishment policy are aggregated and represented in Table II. The cumulative mean and 95 % confidence intervals of the simulation results are given in Table II, as well. Similar calculations were performed for service level and average cost.

TABLE II

CONFIDENCE INTERVAL CALCULATION FOR AVERAGE INVENTORY

Replication	Average inventory	Cumulative mean	Standard deviation	Confidence Level	Lower interval	Upper interval	% deviation
1	100.92	100.92	n/a	n/a	n/a	n/a	n/a
2	99.08	100.00	0.881	7.914	92.09	107.91	7.9%
3	104.33	101.44	1.784	4.430	97.01	105.87	4.4%
4	99.94	101.07	1.563	2.487	98.58	103.55	2.5%
5	101.63	101.18	1.395	1.732	99.45	102.91	1.7%
6	101.83	101.29	1.283	1.346	99.94	102.63	1.3%
7	100.26	101.14	1.203	1.112	100.03	102.25	1.1%
8	102.40	101.30	1.168	0.977	100.32	102.28	1.0%
9	101.12	101.28	1.098	0.844	100.44	102.12	0.8%
10	100.74	101.23	1.043	0.746	100.48	101.97	0.7%

The cumulative mean and confidence intervals for average cost, average inventory and service level are graphically represented in Fig. 9 to Fig. 11. Here, the cumulative mean in all the graphs is reasonably flat and the confidence intervals narrow fairly rapidly.

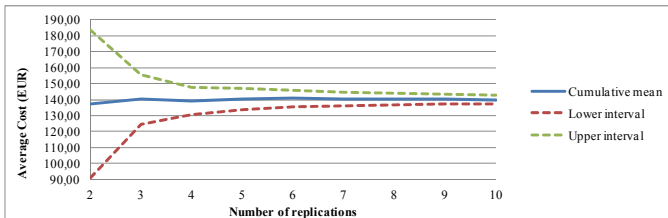


Fig. 9. Cumulative mean and 95 % confidence interval for average cost.

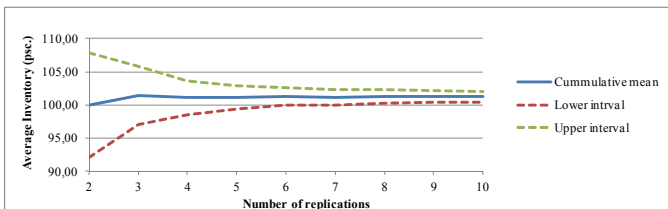


Fig. 10. Cumulative mean and 95 % confidence interval for average inventory.

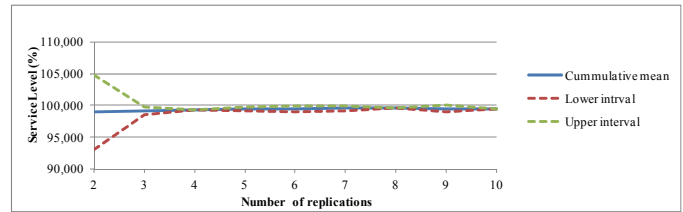


Fig. 11. Cumulative mean and 95 % confidence interval for service level.

The calculations and their graphical interpretation allow concluding that for average inventory and service level the deviation is less than 5 % at three replications. The interval narrows a bit slower for the average cost and the deviation is less than 5 % after the 5th replication. Thus, it is possible to confirm the convergence of the scenario modelling results. Similar calculations are performed for a periodic replenishment policy. Therefore, the scenario can be proposed for the implementation of the regular play of the ECLIPS game.

V. CONCLUSION

Recent advances in information technology have had a great impact on the process of the development and application of management simulation games. The research results provided in this paper show that the complexity of such games can be overcome by introducing a formal game scenario concept.

The proposed scenario development approach allows generating and simulating scenarios of games in order to evaluate their relevance to learning objectives defined by the game manager. Despite the fact that some of the described procedure steps are not yet automated, the provided examples of scenario development for ILMG and ECLIPS games show the procedure advantages for game managers.

Currently, it is not possible to completely avoid the interference of an expert (game manager or instructor) in the process of scenario development; however, further research direction may be related to the elaboration of method for generating values of various scenario parameters, as well as to the use of more sophisticated agents for scenario simulation.

REFERENCES

- [1] G. Merkuryeva, Y. Merkuryev, J. Bikovska, J. Pečerska, J. Petuhova, "Active Learning Logistics Management Through Business Gaming", in *Proceedings of 4th International Conference on Interdisciplinary in Education*, pp. 30–36, 2009.
- [2] A. J. Faria, D. Hutchinson, W. J. Wellington, "Developments in Business Gaming: a Review of the Past 40 Years," *Simulation and Gaming*, vol. 40, no. 4, pp. 464–487, 2009. <http://dx.doi.org/10.1177/1046878108327585>
- [3] G. Merkuryeva, J. Bikovska, T. Ören, "An Agent-directed Multisimulation Framework for Simulation Games Management", *International Journal of Simulation and Process Modeling*, vol. 7, no. 3, pp. 184–192, 2012. <http://dx.doi.org/10.1504/IJSPM.2012.049150>
- [4] M. Jarke, X. T. Bui, J. M. Carroll, "Scenario Management: An Interdisciplinary Approach", *Requirements Engineering*, vol. 3, issue 3–4, pp. 155–173, 1998. <http://dx.doi.org/10.1007/s007660050002>
- [5] D. A. Kononov, V. V. Kulba, S. S. Kovalevsky, S. A. Kosjachenko, *Development of scenario spaces and the analysis of dynamics of behaviour of social and economic system*. Preprint, Moscow, 1999. (In Russian).

- [6] V. L. Shulc, V. V. Kulba, et. al., *Models and methods of analysis and synthesis of scenarios of socio-economic systems*. Book 1, Nauka, Moscow, 2012. (In Russian).
- [7] L. Yilmaz, T. Ören, "Intelligent agents, simulation, and gaming", *Simulation & Gaming*, vol. 37, no. 3, pp. 339–349, 2006.
<http://dx.doi.org/10.1177/1046878106289089>
- [8] M. Remondino, "A Web Based Business Game Built on System Dynamics Using Cognitive Agents as Virtual Tutors", in *Proceedings of the Tenth International Conference on Computer Modeling and Simulation*, pp. 568–572, 2008.
<http://dx.doi.org/10.1109/UKSIM.2008.84>
- [9] R. W. Grubbström, G. Merkuryeva, J. Bikovska, J. Weber. "ILMG: Learning Arrangements and Simulation Scenarios," in *Proceedings of 19th European Conference on Modeling and Simulation*, pp. 715–720, 2005.
- [10] J. Bikovska, G. Merkuryeva, "The International Logistics Management Game: An Innovative Business Environment for Training," *Production-Economic Research in Linköping*, pp. 55–72, 2011.
- [11] Y. Merkuryev, G. Merkuryeva, J. Bikovska, J. Hatem, B. Desmet, "Business simulation game for teaching multi-echelon supply chain management," *International Journal of Simulation and Process Modeling*, vol. 5, no 4, pp. 289–299, 2009.
- [12] Y. Merkuryev, J. Bikovska, "Business Simulation Game Development for Education and Training in Supply Chain Management," in *Proceedings of Asia Modeling Symposium (AMS2012), the Sixth Asia International Conference on Mathematical Modeling and Computer Simulation*, pp. 179–184, 2012.

Jana Bikovska is a Researcher and Lecturer with the Department of Modeling and Simulation, Riga Technical University, Latvia. Her professional interests are in the field of supply chain management, management simulation games, scenario approach in complex systems with application to simulation games. Her teaching experiences include the following management games: ILMG, New ORSIAM Int., ROFA, ECLIPS Game, and Beer Game.
E-mail: Jana.Bikovska@rtu.lv

Jana Bikovska. Vadību imitējošo spēļu scenāriju veidošanas pieeja

Rakstā ir apskatīta vadību imitējošo spēļu veidošanas pieeja. Vadīšanas imitējošās spēles gūst aizvien lielāku popularitāti, tā ir kā alternatīva pieeja mācībām, kas nodrošina virtuālo biznesa vidi vadības lēmumu pieņemšanas iemaņu trenēšanai. Mūsdienās, pateicoties informācijas un komunikācijas tehnoloģiju attīstībai, paplašinās arī imitējošo spēļu funkcionālās iespējas, kas paaugstina to efektivitāti un maksimāli pietuvina reālās dzīves situācijām. Līdz ar to rodas nepieciešamība izstrādāt tādas metodes, pieejas un rīkus, kas nodrošina spēles kompleksās biznesa vides pārvaldību. Šim nolūkam tiek ieviesta vadību imitējošo spēļu scenāriju koncepcija, kura kalpo par pamatu rakstā piedāvātajai spēļu scenāriju veidošanas procedūrai. Šī procedūra paredz vairāku soļu izpildi, lai izveidotu tādu vadību imitējošās spēles scenāriju, kas atbilst apmācības mērķiem. Dažus no šīs procedūras soļiem izpilda spēles vadītājs (instruktors), bet daži ir automatizēti. Piemēram, scenāriju parametru vērtības tiek ģenerētas automātiski, kā arī izveidoto scenāriju modelēšana notiek automātiski, piemēram, pielietojot aģentus, kas aizstāj reālus spēles dalībniekus. ILMG un ECLIPS spēles tika izvēlētas, lai parādītu, kā izstrādātā pieeja tiek praktiski realizēta. ILMG ir universāla, internetā bāzēta, tiešsaistes vadību imitējoša datorspēle, kas nodrošina virtuālo biznesa vidi lēmumu pieņemšanas iemaņu pilnveidošanai tādās sfērās kā starptautiskā loģistika, ražošanas vadība, tirgzinība un finanses. ILMG spēles scenāriju parametru skaits ir diezgan liels, un scenārija manuāla veidošana un modelēšana ir pārāk laikietilpīga. Līdz ar to, pielietojot izstrādāto scenāriju veidošanas procedūru, izdevās krietni samazināt darba apjomu un patērēto laiku. Savukārt ECLIPS ir funkcionāla galda spēle, kuras mērķis ir sniegt priekšstatu par piegādes ķēdes darbības pamatprincipiem un dažādu krājumu vadības stratēģiju ieviešanas priekšrocībām un trūkumiem. ECLIPS spēlei tika izstrādāts elektronisks analogs, kas ļauj ģenerēt, modelēt un novērtēt dažādas spēles scenārijus, kas arī būtiski atvieglo spēles vadītāja (instruktora) darbu.

Яна Биковска. Подход для разработки сценариев в деловых имитационных играх

В статье рассмотрен подход для разработки сценариев в деловых имитационных играх. Использование деловых имитационных игр становится все более популярным, так как обеспечивает альтернативный подход к обучению, предоставляя виртуальную бизнес-среду для освоения навыков принятия управленческих решений. Благодаря современному уровню развития информационных и коммуникационных технологий, расширяются и функциональные возможности деловых имитационных игр, что значительно повышает их эффективность и максимально приближает к ситуациям реальной жизни. В этой связи появляется необходимость в разработке таких методов, подходов и средств, которые обеспечивают управление комплексной бизнес-средой этих игр. Для этого вводится концепция сценария имитационной игры, которая служит основой для предлагаемой в статье процедуры разработки сценариев деловых имитационных игр. Эта процедура предполагает выполнение нескольких этапов с целью разработки такого сценария деловой имитационной игры, который соответствует конкретной цели обучения. Некоторые этапы этой процедуры выполняет инструктор, другие - автоматизированы. Например, значения параметров сценария генерируются автоматически, а также моделирование сценария происходит автоматически, например, с использованием агентов, которые замещают реальных участников игры. Для демонстрации практической реализации разработанного подхода используются две игры: ILMG и ECLIPS. ILMG представляет собой универсальную онлайн деловую имитационную игру, которая обеспечивает виртуальную бизнес-среду для освоения навыков принятия решений в таких сферах, как международная логистика, управление производством, маркетинг и финансы. Количество различных параметров сценария в ILMG игре достаточно большое, следовательно, разработка и моделирование сценария в ручном режиме занимает много времени. Применяя разработанную процедуру разработки сценария, удалось значительно сократить объем работы и потраченное время. В свою очередь, игра ECLIPS представляет собой функциональную настольную деловую имитационную игру, цель которой дать представление об основных принципах работы цепи поставок, а также об основных преимуществах и недостатках различных стратегий управления запасами. Для ECLIPS игры был разработан электронный аналог, который позволяет генерировать, моделировать и оценивать различные сценарии игры, что значительно облегчает работу инструктора.