ITHEA

International Journal INFORMATION MODELS & ANALYSES

2016 Volume 5 Number 3

International Journal
INFORMATION MODELS & ANALYSES
Volume 5 / 2016, Number 3

IJ IMA is official publisher of the scientific papers of the members of the ITHEA® International Scientific Society

IJ IMA rules for preparing the manuscripts are compulsory.
The **rules for the papers** for ITHEA International Journals are given on www.ithea.org .
The camera-ready copy of the paper should be received by ITHEA® Submission system http://ij.ithea.org .
Responsibility for papers published in IJ IMA belongs to authors.

International Journal "INFORMATION MODELS AND ANALYSES" Volume 5, Number 3, 2016

Edited by the **Institute of Information Theories and Applications FOI ITHEA**, Bulgaria, in collaboration with
Institute of Mathematics and Informatics, BAS, Bulgaria,
V.M.Glushkov Institute of Cybernetics of NAS, Ukraine,
Universidad Politechnika de Madrid, Spain,
Hasselt University, Belgium,
Institute of Informatics Problems of the RAS, Russia,
St. Petersburg Institute of Informatics, RAS, Russia
Institute for Informatics and Automation Problems, NAS of the Republic of Armenia,

Publisher: **ITHEA®** Sofia, 1000, P.O.B. 775, Bulgaria. www.ithea.org, e-mail: info@foibg.com
Technical editor: **Ina Markova**

Printed in Bulgaria

ISSN 1314-6416 (printed) **ISSN 1314-6432 (Online)**

A METHOD FOR EVALUATION OF INFORMATIONAL SERVICES - STEP 2: COMPUTING THE INFORMATIONAL SERVICES' PERFORMANCE PROPORTIONALITY CONSTANTS

Krassimira Ivanova, Ivan Ivanov, Mariyana Dimitrova, Krassimir Markov, Stefan Karastanev

Abstract: *Enhancing the hardware power does not cause linear enhancing of the informational services' performance. To discover the value of growth one has to test both source and enhanced systems running equal or similar services. If we need to discover the growth of services' performance for different computers' configurations we have to have common basis for comparing one software service with those of other systems which are tested on different computer configurations. In paper [Ivanova et al 2016] the first step of a method for solving such problem was presented. In this paper we outline the second step of the method. This step consists of computing the informational services' performance proportionality constants. Further paper will present the last step of the method. All examples in the paper are based on results from real experiments presented in the [Markov et al, 2015].*

Keywords: *Evaluation of informational services; computing the software systems' performance proportionality constants.*

ACM Classification Keywords: *H.3.4 Systems and Software - Performance evaluation (efficiency and effectiveness); H.3.5 Online Information Services.*

Introduction

In series of three papers we present a method for evaluation of informational services. It consists of three steps:

1. Computing the hardware proportionality constants;
2. Computing the software systems' performance and proportionality constants;
3. Analysis of experiments: Rank-based multiple comparison.

In the paper [Ivanova et al, 2016] we outlined the first step of the method - computing the hardware proportionality constants. This step is important due to differences in the hardware and corresponding operational systems. Further paper will present the third step of the method. Now we will discuss the

second step of the method: computing the informational services' performance proportionality constants. All examples in the paper are based on results from real experiments presented in the [Markov et al, 2015].

Let remember the main problem to be solved.

Enhancing the hardware power does not cause linear enhancing of the informational services' performance. To discover the value of growth one has to test both source and enhanced systems running equal or similar software.

In our case we have the same problem. In [Ivanova et al, 2016] we show that computer configurations A, K, B, and C, may be ordered by their Average CPU Marks as well as their General scores. In all cases we need to discover the growth of software performance for different configurations. This is needed because we want to have common basis for comparing our load time with those of other systems which are tested on different computer configurations.

For this purpose we will follow simple algorithm.

Let informational service **X** is tested on two computer configurations: **U** and **W**, where **W** is enhanced configuration; and informational service **Y** is tested on different computer configuration **V** of the same class and similar characteristics as **U**. We have couples (**X**,**U**), (**X**,**W**), and (**Y**,**V**).

Computer configurations **U** and **W** are not available for testing and all work has to be done on computer configuration **V**.

Computer configurations' global scores are respectively:

$$\mathbf{E_U} = 0.3, \mathbf{E_V} = 1, \text{ and } \mathbf{E_W} = 3.$$

X is tested on **U** by dataset **S1** with 200 instances and on **W** with similar dataset **S2** with 250 instances.

Y is tested on configuration **V** by datasets **S1** and **S2**.

Loading times are respectively:

$\mathbf{L_{(X,U,S1)}}$=1000 sec., $\mathbf{L_{(X,W,S2)}}$=5 sec.;

$\mathbf{L_{(Y,V,S1)}}$=400 sec., $\mathbf{L_{(Y,V,S2)}}$=500 sec.

The problem we have to solve is:

"What will be the loading time of informational service **Y** if it will be run on computer configuration **W** with dataset **S2**?" i.e. $\mathbf{L_{(Y,W,S2)} = ?}$.

The algorithm

Firstly we will illustrate the algorithm and after that we will give it in details.

We have the diagram (Figure 1):

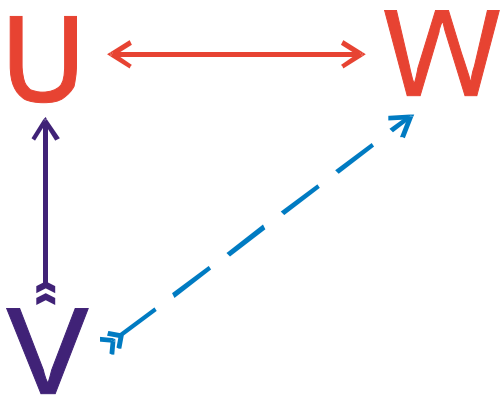


Figure 1. Interrelations between computer configurations

Using published data we may estimate interrelations between computer configurations **U** and **W** as well as between two versions of informational service **X** run on **U** and **W**. We have to use hardware proportionality constants (given in [Ivanova et al, 2016]) to make data comparable and to compute the ratio coefficient of software growth by dividing the loading time on **W** by one on **U**.

To make data from experiments on **V** comparable with these on **U** and **W** we assume that **V** and **U** are from the same class of computer power and there is no software growth for a informational service **Y** in the transition from **V** to **U**. In other words, to estimate interrelations between computer configurations **V** and **U** we need only hardware proportionality constants. After this step we will have data from experiments on **V** transferred for the **U**, i.e. we will have results from informational service **Y** as if the informational service **Y** is tested on configuration **U**.

We assume that the possible software growth of informational service **Y** from computer **U** to **W** is the same as for the informational service **X**, i.e. we can use the same coefficient for software growth for systems **X** and **Y**. This way we will have comparable data for computer configuration **W**.

Below the algorithm is given in details:

1. Reduce loading time $\mathbf{L_{(X,W,S2)}}$ of informational service **X**, run on computer configuration **W** and dataset **S2** with |**S2**|=250 instances, to loading time $\mathbf{L_{(X,W,S2')}}$ of **X** for hypothetical dataset **S2'** with |**S2'**|=|**S1**|=200 instances, using the formula

$$\mathbf{L_{(X,W,S2')}} = |\mathbf{S2'}| * (\mathbf{L_{(X,W,S2)}}/ |\mathbf{S2}|) =$$
$$= |\mathbf{S1}| * (\mathbf{L_{(X,W,S2)}}/ |\mathbf{S2}|) = 200*(5/250) = \mathbf{4}$$

2. Compute ratio coefficient of growth $\mathbf{G_{UW}}$ from (**X**,**U**) to (**X**,**W**) by equation:

$$\mathbf{G_{UW}} = \mathbf{L_{(X,U,S1)}}/\mathbf{L_{(X,W,S2')}} = 1000/4 = \mathbf{250}$$

3. Compute loading time $L_{(Y,U,S2)}$ of informational service **Y** with dataset **S2** if it is hypothetically ran on configuration **U**, using hardware proportionality constant H_{VU}:

$$V \propto U: \quad H_{VU} = E_V/E_U = 1 / 0.3 = \mathbf{3.33}$$

and formula:

$$L_{(Y,U,S2)} = H_{VU} * L_{(Y,V,S2)} = 3.33 * L_{(Y,V,S2)} = 3.33 * 500 = \mathbf{1665}$$

4. Compute loading time $L_{(Y,W,S2)}$ of informational service **Y** with dataset **S2** if it is hypothetically ran on configuration **W,** using ratio coefficient of growth G_{UW}, hypothetical loading time $L_{(Y,U,S2)}$, and formula:

$$L_{(Y,W,S2)} = L_{(Y,U,S2)}/G_{UW} = L_{(Y,U,S2)} / 250 = 1665/250 = \mathbf{6.66}$$

This way we have received comparable value of loading time of informational service **Y** with informational service **X** for computer configuration **W**, i.e.

$$L_{(X,W,S2)}=5 \text{ sec. and } L_{(Y,W,S2)} = 6.66 \text{ sec.}$$

and we may conclude that informational service **X** will have a little better loading time than informational service **Y** if both are run on computer configuration **W** with dataset **S2**.

One may suppose that we may use directly proportionality constant H_{WV}:

$$W \propto V: \quad H_{WV} = E_W/E_V = 3 / 1 = \mathbf{3}$$

and to reduce $L_{(Y,V,S2)}$=500 sec. three times, i.e. 500/3 = 166.66.

This is not correct because the ***software growth*** is not taken in account.

We have to calculate possible software growth from **V** to **W** again going through **U** and using G_{UW} to calculate possible G_{VW}. This may be done by using the proportionality constant H_{VU} because we need to calibrate growth from **U** to **W** by hardware proportionality of **V** and **U**. In other words, to receive value of growth G_{VW} from **V** to **W** we have to compute:

$$G_{VW} = G_{UW}/H_{VU}$$

Finally:

$$L_{(Y,W,S2)} = L_{(Y,V,S2)}/G_{VW}$$

Let see it for concrete values:

$$G_{UW} = L_{(X,U,S1)}/L_{(X,W,S2')} = 1000/4 = \mathbf{250}$$

$$H_{VU} = E_V/E_U = 1 / 0.3 = \mathbf{3.33}$$

$$G_{VW} = (G_{UW}/H_{VU}) = (250/3.33) = \mathbf{75.07}$$

$$L_{(Y,W,S2)} = L_{(Y,V,S2)}/G_{VW} = 500 / 75.07 = \mathbf{6.66}$$

We received the same result as algorithm above. This proves that we have equivalent approaches.

The algorithm may be presented by a formula:

$$L_{(Y,W,S2)} = R_{YVW} * L_{(Y,V,S2)}$$

where

$$R_{YVW} = \frac{E_v * |S1| * L_{(X,W,S2)}}{E_U * |S2| * L_{(X,U,S1)}}$$

i.e.

$$L_{(Y,W,S2)} = \frac{E_v * |S1| * L_{(X,W,S2)}}{E_U * |S2| * L_{(X,U,S1)}} * L_{(Y,V,S2)}$$

where:

- **X**, **Y** - informational services;
- **U**, **V**, **W** – computer configurations;
- (**X**,**U**), (**X**,**W**), (**Y**,**V**) – couples “informational service – computer configuration”;
- E_U, E_V, E_W - computer configurations’ global scores;
- **S1**, **S2** – datasets;
- $L_{(X,U,S1)}$, $L_{(X,W,S2)}$, $L_{(Y,V,S1)}$, $L_{(Y,V,S2)}$, $L_{(Y,W,S2)}$ - loading times of given informational service, computer configuration, and dataset;
- H_{VU} – computer configurations’ proportionality constant;
- G_{UW} – ratio coefficient of growth of informational service during migration from a computer configuration to enhanced one.

Experimental environment

Our experimental environment includes informational services, computer configurations, datasets and experimental data like published benchmark results, different constants, ratio coefficients, etc.

The main elements of our experimental environment (for concrete names see [Markov et al, 2015]) are:

- Informational services to be compared are: **R**, **V**, **J**, and **S**.

  V, **J** and **S** have several variants depending of databases used. These variants have different loading times on the same computer configurations. In our comparisons we will take the best result from the all benchmarks on given configuration.
- Computer configurations used for benchmarking are **A**, **K**, **B**, **C**;
- Couples "informational service – computer configuration" are:

  (**R**, **K**);

  (**V**, **A**), (**V**, **B**), (**V**, **C**);

  (**J**, **A**), (**J**, **B**), (**J**, **C**);

  (**S**, **A**), (**S**, **B**), (**S**, **C**).
- Computer configurations' global scores are $\mathbf{E_A}$, $\mathbf{E_K}$, $\mathbf{E_B}$, and $\mathbf{E_C}$;
- Middle-size datasets are: **B50K**; **H.nt**; **B250K**; **G.nt**; **B1M**; **B5M**.
- Large size datasets are: **I.nt**; **B25M**; **B100M**.
- Proportionality constant between computer configurations **K** and **A** is $\mathbf{H_{KA}}$ [Ivanova et al, 2016];
- Ratio coefficient of growth of informational services during migration from computer configuration **A** to enhanced ones **B** and **C** are $\mathbf{G_{AB}}$ and $\mathbf{G_{AC}}$;

Corresponded loading times **L** will be presented at the places where they will be used.

Software proportionality constants

To provide concrete comparisons of our experimental loading time data, we have to compute $\mathbf{H_{KA}}$, $\mathbf{G_{AB}}$, and $\mathbf{G_{AC}}$.

For purposes of this discussion, it is enough to compute average constants $\mathbf{H_{KA}}$, $\mathbf{G_{AB}}$, and $\mathbf{G_{AC}}$ based on average loading data for all chosen systems. We will use published benchmark results (done by Freie Universität Berlin, Web-based Systems Group (BSBM team)) and available both as printed publication and free accessible data in the Internet.

Software proportionality for configurations K, A, and B

Benchmark results for dataset **S1** (**H.nt**; 200 036 instances) used for benchmarks on **Configuration A** are published in [Becker, 2008] and reproduced in Table 1.

Table 1. Benchmark results for dataset S1 (**H.nt**)

system	loading time in seconds	the best time in seconds
V	1327	1327
J Variant 1	5245	3557
J Variant 2	3557	
J Variant 3	9681	
S	2404	2404
Total average time in seconds:		2429.333

Benchmark results for dataset **S2** (***B250K***; 250 030 instances) used for benchmarks on **Configuration B** are published in [BSBMv2, 2008] and reproduced in Table 2.

Table 2. Benchmark results for dataset S2 (***B250K***)

system	loading time in seconds
V	33
J	24
S	18
Total average time in seconds:	25

Due to equal informational services and range of their loading times on the same computer configuration, we will use total average times as loading times of virtual informational service **X**, i.e. $L_{(X,A,S1)}$ = **2429.333** and $L_{(X,B,S2)}$ = **25**.

Following our algorithm, we reduce loading time $L_{(X,B,S2)}$ of virtual informational service **X**, run on computer configuration **B** and dataset **S2** with |**S2**|=250 030 instances, to loading time $L_{(X,B,S2')}$ of **X** for hypothetical dataset **S2'** with |**S2'**|=|**S1**|=200 036 instances, using the formula

$$L_{(X,B,S2')} = |S1| * (L_{(X,B,S2)}/ |S2|) = 200036*(25/250030) = \mathbf{20.00}.$$

We compute ratio coefficient of growth G_{AB} from (**X**,**A**) to (**X**,**B**) by equation:

$$G_{AB} = L_{(X,A,S1)}/L_{(X,B,S2')} = 2429.333/20 = \mathbf{121.46665}.$$

Hardware proportionality constant H_{AK} is:

$$A \propto K: \quad H_{AK} = E_K/E_A = 1 / 0.32 = \mathbf{3.125}$$

Really measured **R** loading time on Configuration **K** for dataset **S2** is **575.069** sec. We compute loading time $L_{(R,A,S2)}$ using formula:

$$L_{(R,A,S2)} = H_{AK}*L_{(R,K,S2)} = 3.125*575.069 = 1797.09.$$

At the end, we compute loading time $L_{(R,B,S2)}$ of informational service **R** with dataset **S2** if it is hypothetically run on configuration **B,** using ratio coefficient of growth G_{AB}, hypothetical loading time $L_{(R,A,S2)}$, and formula:

$$L_{(R,B,S2)} = L_{(R,A,S2)}/G_{AB} = 1797.09 / 121.46665 = \mathbf{14.796}$$

To **verify** our computations and to show *the easiest way* to find $L_{(R,B,S2)}$, we will use our formula

$$L_{(RDFArM,B,S2)} = R_{RDFArM,K,B} * L_{(RDFArM,K,S2)}$$

i.e. we have to compute $R_{R,K,B}$ one time and to use it in benchmarks for all datasets. $R_{R,K,B}$ may be computed by formula:

$$R_{RDFArM,A,B} = \frac{E_K * |S1| * L_{(X,B,S2)}}{E_A * |S2| * L_{(X,A,S1)}}$$

or in linear view:

$$R_{R,K,B} = (E_K * |S1| * L_{(X,B,S2)}) / (E_A * |S2| * L_{(X,A,S1)}) =$$
$$= (1 * 200036 * 25) / (0.32 * 250030 * 2429.333) =$$
$$= 5000900 / 194369961.5968 = \mathbf{0.025729}.$$

We compute loading time $L_{(R,B,S2)}$ of informational service **R** with dataset **S2** if it is hypothetically run on configuration **B,** using ratio coefficient $R_{R,K,B}$:

$$L_{(R,B,S2)} = L_{(R,K,S2)} * R_{R,K,B} = 575.069 * 0.025729 = \mathbf{14.796}.$$

We receive the same result.

Software proportionality for configurations K, A, and C

Software proportionality for configurations **K**, **A**, and **C** will be computed based on the performance of systems **V** and **J** because missing information about **S** in the benchmark publications.

Benchmark results for dataset **S1** (***I.nt***; 15,472,624 instances) used for benchmarks on **Configuration A** are published in [Becker, 2008] and reproduced in Table 3.

Table 3. Benchmark results for dataset S1 (**I.nt**)

system	loading time in seconds	the best time in seconds
V	**7017**	**7017**
J Variant 1	**70851**	**70851**
J Variant 2	**73199**	
J Variant 3	**734285**	
Total average time:		**38934**

Benchmark results for dataset **S2** (***B100M***; 100 000 748 instances) used for benchmarks on **Configuration C** are published in [BSBMv6, 2011] and reproduced in Table 4.

Table 4. Benchmark results for dataset S2 (***B100M***)

system	loading time in seconds
V	**6566**
J	**4488**
Total average time:	**5527**

Following our algorithm, we reduce loading time $\mathbf{L_{(X,C,S2)}}$ of virtual informational service **X**, run on computer configuration **C** and dataset **S2** with |**S2**|=100 000 748 instances, to loading time $\mathbf{L_{(X,C,S2')}}$ of **X** for hypothetical dataset **S2'** with |**S2'**|=|**S1**|=15 472 624 instances, using the formula:

$$\mathbf{L_{(X,C,S2')}} = |\mathbf{S1}| * (\mathbf{L_{(X,C,S2)}}/ |\mathbf{S2}|) =$$
$$= 15472624*(5527/100000748) = \mathbf{855.166}.$$

We compute ratio coefficient of growth $\mathbf{G_{AC}}$ from (**X**,**A**) to (**X**,**C**) by equation:

$$\mathbf{G_{AC}} = \mathbf{L_{(X,A,S1)}}/\mathbf{L_{(X,C,S2')}} = 38934/855.166 = \mathbf{45.528}.$$

Hardware proportionality constant $\mathbf{H_{AK}}$ is:

$$\mathbf{A \propto K:} \quad \mathbf{H_{AK}} = \quad \mathbf{E_K/E_A} = \quad 1 / 0.32 \quad = \mathbf{3.125}.$$

Really measured **R** loading time on Configuration **K** for dataset **S2** is 43652.528 sec. We compute loading time $\mathbf{L_{(R,A,S2)}}$ using formula:

$$\mathbf{L_{(R,A,S2)}} = \mathbf{H_{AK}}*\mathbf{L_{(R,K,S2)}} = 3.125*43652.528 = \mathbf{136414.15}.$$

At the end, we compute loading time $\mathbf{L_{(R,C,S2)}}$ of system **R** with dataset **S2** if it is hypothetically run on configuration **C,** using ratio coefficient of growth $\mathbf{G_{AC}}$, hypothetical loading time $\mathbf{L_{(R,A,S2)}}$, and formula:

$$\mathbf{L_{(R,C,S2)}} = \mathbf{L_{(R,A,S2)}}/\mathbf{G_{AC}} = 136414.15/45.528 = \mathbf{2996.27} \text{ sec.}$$

To **verify** our computations and to show *the easiest way* to find $\mathbf{L_{(R,C,S2)}}$, we will use our formula

$$L_{(RDFArM,C,S2)} = R_{RDFArM,K,C} * L_{(RDFArM,K,S2)}$$

i.e. we have to compute $\mathbf{R_{R,K,C}}$ one time and to use it in benchmarks for all datasets. $\mathbf{R_{R,K,C}}$ may be computed by formula:

$$R_{RDFArM,A,C} = \frac{E_K * |S1| * L_{(X,C,S2)}}{E_A * |S2| * L_{(X,A,S1)}}$$

or in linear view:

$$\mathbf{R_{R,K,C}} = (\mathbf{E_K} * |\mathbf{S1}| * \mathbf{L_{(X,C,S2)}}) / (\mathbf{E_A} * |\mathbf{S2}| * \mathbf{L_{(X,A,S1)}}) =$$

$$= (1 * 15472624 * 5527) / (0.32 * 100000748 * 38934) =$$

$$= 85517192848 / 1245897319242.24 = \mathbf{0.068639}.$$

We compute loading time $\mathbf{L_{(R,C,S2)}}$ of informational service **R** with dataset **S2** if it is hypothetically run on configuration **C,** using ratio coefficient $\mathbf{R_{R,K,C}}$:

$$\mathbf{L_{(R,C,S2)}} = \mathbf{L_{(R,K,S2)}} * \mathbf{R_{R,K,C}} = 43652.528 * 0.068639 = \mathbf{2996.27}.$$

We receive same result.

Ratio coefficients

To compare our results from experiments on computer configuration **K** we will use ratio coefficients:

For published results received on computer configuration **A**:

$$\mathbf{L_{(R,A,S2)} = L_{(R,K,S2)} * 3.125;}$$

For published results received on computer configuration **B**:

$$\mathbf{L_{(R,B,S2)} = L_{(R,K,S2)} * 0.025729;}$$

For published results received on computer configuration **C**:

$$\mathbf{L_{(R,C,S2)} = L_{(R,K,S2)} *0.068639.}$$

Conclusion

The goal of this work was to outline the second step of a method for estimating further development of any informational service. This step consists of computing the software proportionality constants and ratio coefficients.

We assumed that the "software growth" will be done in the same grade as one of the known systems. Estimation of experimental systems was provided to make different configurations comparable. Using proportionality formula, experiments become comparable. We have provided series of experiments which were needed to estimate the storing time of a concrete informational services for middle-size and very large datasets. Our experimental environment included program systems, computer configurations, datasets and experimental data like published benchmark results, different constants, ratio coefficients, etc. All examples in the paper were based on results from real experiments presented in the [Markov et al, 2015].

A further paper will present the last steps of the method.

Acknowledgement

The paper is published with partial support by the Project "Methods for modeling and evaluation of informational services" of the University of Telecommunications and Posts, Sofia, Bulgaria.

Bibliography

[Becker, 2008] Christian Becker, "RDF Store Benchmarks with Dbpedia", Freie Universität Berlin, 2008, http://wifo5-03.informatik.uni-mannheim.de/benchmarks-200801/ (accessed: 05.04.2013)

[BSBMv2, 2008] Berlin SPARQL Benchmark Results, V2 2008, http://wifo5-03.informatik.uni-mannheim.de/bizer/berlinsparqlbenchmark/results/V2/index.html (accessed: 31.07.2013)

[BSBMv6, 2011] Berlin SPARQL Benchmark Results, V6, 2011, http://wifo5-03.informatik.uni-mannheim.de/bizer/berlinsparqlbenchmark/results/V6/index.html (accessed: 31.07.2013)

[Ivanova et al, 2016] Krassimira Ivanova, Emiliya Saranova, Krassimir Markov, Stefan Karastanev A Method for Evaluation of Informational Services - Step 1: Computing the Hardware Proportionality Constants.

[Markov et al, 2015] Krassimir Markov, Krassimira Ivanova, Koen Vanhoof, Vitalii Velychko, Juan Castellanos, „Natural Language Addressing", ITHEA® Hasselt, Kyiv, Madrid, Sofia, IBS ISC No.: 33, 2015, ISBN: 978-954-16-0070-2 (printed), ISBN: 978-954-16-0071-9 (online), 315 p.

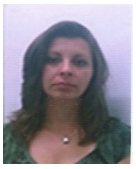
Authors' Information

Krassimira Ivanova *– Assoc. prof. Dr.; University of Telecommunications and Posts, Sofia, Bulgaria; Institute of Mathematics and Informatics, BAS, Bulgaria; e-mail: krasy78@mail.bg;*

Major Fields of Scientific Research: *Software Engineering, Business Informatics, Data Mining, Multidimensional multi-layer data structures in self-structured systems*

Ivan Ivanov *– Assist. prof.; University of Telecommunications and Posts, Sofia, Bulgaria; e-mail: e-mail: i.ivanov@vutp.bg*

Major Fields of Scientific Research: Information and Network Security, Cryptographic Methods and Algorithms



Mariyana Dimitrova *– student; University of Telecommunications and Posts, Sofia, Bulgaria; e-mail: mariyana.dimitrova1@gmail.com*

Major Fields of Scientific Research: Business Modeling

Krassimir Markov *– Institute of Mathematics and Informatics, BAS, Acad. G.Bontchev St., bl.8, Sofia-1113, Bulgaria; ITHEA Institute of Information Theories and Applications, P.O. Box: 775, Sofia-1000, Bulgaria; e-mail: markov@foibg.com*

Major Fields of Scientific Research*: General theoretical information research, Multi-dimensional information systems; Software Engineering, Business Informatics, Data Mining.*

Stefan Karastanev *– Assist. prof.; Institute of Mechanics, BAS, Bulgaria;*

e-mail: stefan@imbm.bas.bg

Major Fields of Scientific Research: *Software Engineering, Data Processing and Mining, Data structures in information systems.*

INTELLIGENT FRAMEWORK FOR RECOMMENDATION OF MOBILE SERVICES TO CONSUMERS

Ivan Ganchev

Abstract: *This paper describes an intelligent framework for the recommendation of mobile services to users, based on a combination of two different approaches. The first one utilizes Wireless Billboard Channels (WBCs) to push (by broadcasting) service advertisements to mobile users/devices in order to enable them to discover and associate with the 'best' service instances under the Always Best Connected and best Served (ABC&S) communications paradigm. The second approach uses a dedicated service recommendation system, which finds and suggests to each individual user the 'best' service instances, depending on the current user-, network-, and service context. The main parts of the proposed framework are explained and the key technological solutions required to support it are outlined.*

Keywords: *Ubiquitous Consumer Wireless World (UCWW), Always Best Connected and best Served (ABC&S), Wireless Billboard Channels (WBCs), Advertisement, Discovery and Association (ADA), mobile services, service recommendation system, intelligent framework.*

ACM Classification Keywords: *C.0 GENERAL – System architectures, C.2.1 Network Architecture and Design – Wireless communication, C.2.4 Distributed Systems – Client/server, D.2.2 Design Tools and Techniques – Evolutionary prototyping.*

1. Introduction

In the near future, a next generation network (NGN) wireless environment, called the Ubiquitous Consumer Wireless World (UCWW) [O'Droma, 2007, O'Droma, 2010], will emerge, where users will act more like consumers of mobile services rather than subscribers as nowadays. In the UCWW, services will be available anywhere-anytime-anyhow, will be customized to the user's needs and adapted to the current user- and network context, in the best possible way, independent of the user's movement across heterogeneous access networks, e.g. 3G/4G, Wi-Fi, WiMAX, etc., according to his/her own criteria (e.g. on the basis of price/performance offerings), while maintaining active service sessions, i.e. without service interrupting, restarting applications, or losing data. This new communications paradigm, called the Always Best Connected and best Served (ABC&S) [O'Droma, 2006], will be facilitated by a novel

Consumer-Based techno-business model (CBM), which will enable a loose dynamic (even casual) consumer-type association between mobile users and access network providers (ANPs).

The UCWW could be considered as a global communications environment, which brings a different approach to the wireless communications business. It will provide greater flexibility and freedom to mobile users, full user mobility among participating access networks, and a greater degree of service choice. Besides these new benefits for users, the UCWW has the potential to stimulate the creation of a number of new interesting business opportunities and to create a more liberal, more open and fairer wireless marketplace for existing and new ANPs and (mobile) service providers (xSPs), allowing their primary business success indicator to shift from subscriber numbers to the volume of consumer transactions. In the future, this will increase the range of competitive price/performance and price/quality offerings, specialist and niche service offerings, and so forth, all of which will drive forward innovation in the wireless communications and mobile services market.

To enable this user-oriented, user-friendly, and user-driven ABC&S wireless communication environment, the mobile user should be made aware of the presence of communications services and mobile services around him/her. One possibility to accomplish this is to use Wireless Billboard Channels (WBCs) to solely push (by means of broadcasting) service advertisements to mobile users/devices in order to enable them to discover and associate with 'best' service instances under the ABC&S paradigm. This technological solution along with the supplementary procedure for services' Advertisement, Discovery and Association (ADA) is described in greater detail in the next section.

Another option is to use a dedicated service recommendation system, which to do the same without a need for additional infrastructural network elements, simply by utilizing the (mobile) Internet connection of users. Details about this option are provided in Section 3. The building of an intelligent framework for recommending mobile services to users, by combining these two approaches, is described in Section 4. Conclusions and future research work are presented in Section 5.

2. Wireless Billboard Channels (WBCs) and Advertisement, Discovery and Association (ADA)

Wireless Billboard Channels (WBCs) are simplex push-based channels, used for pro-active broadcasting of service advertisements to mobile devices (in a particular coverage area) in order to enable mobile users to discover and associate with 'best' service instances (in a background mode of operation) under the ABC&S communications paradigm (Figure 1). If needed, they can operate also as down-link (DL) out-band Cognitive Pilot Channels (CPCs).

At first glance, the WBCs may look as a form of CPCs. However, the WBC concept was elaborated in 2004 by two researchers (O'Droma and Ganchev) at the University of Limerick, Ireland [O'Droma, 2004a], i.e. a few years before the CPC concept [Bourse, 2007]. The CPC idea was promoted by the

European Telecommunications Standards Institute (ETSI) [ETSI, 2009] in order to provide collaboration between networks and mobile devices as to enable the information transfer to mobile devices of available knowledge of the wireless network environment, including available radio access networks (RANs), radio access technologies (RATs), corresponding coverage areas and frequency bands, network policies, etc. CPCs are used mostly in relation to the wireless communications services' provision, whereas WBCs provide also (advertisement) information about the (most appropriate to the user) mobile services, which makes them much richer solution as regards the functionality.

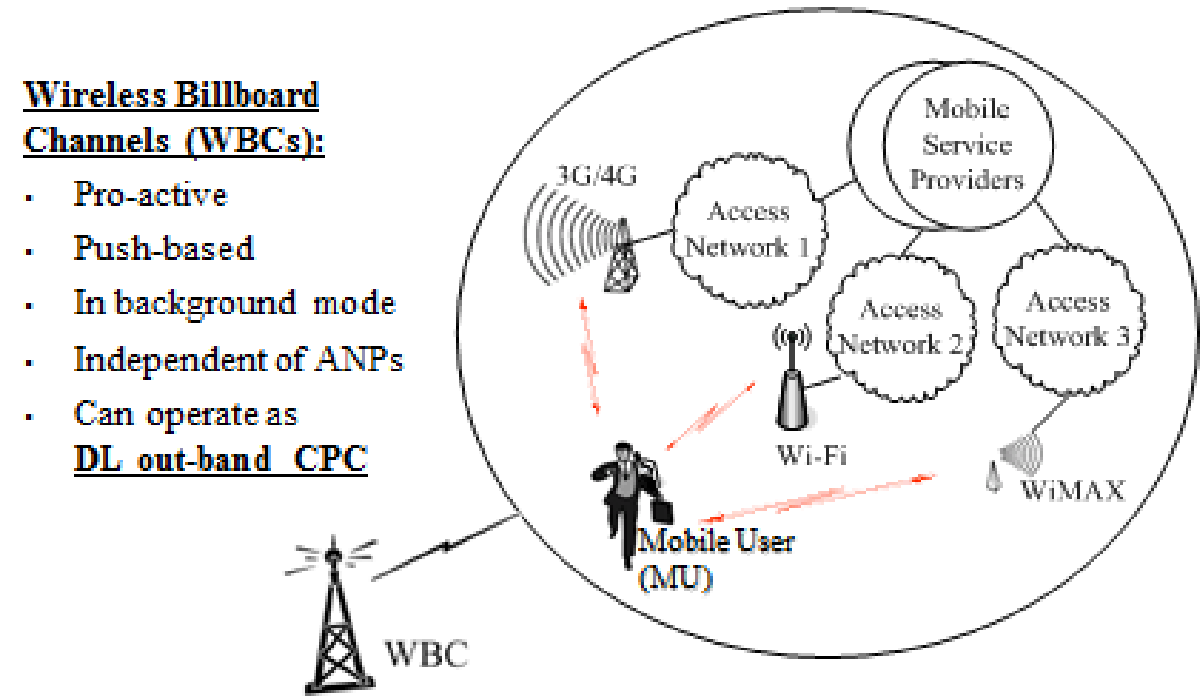


Fig. 1. The use of WBCs for the ABC&S realization (adapted from [Ji, 2010a])

If the user (acting as a consumer) is located in the footprints of several wireless access networks, first thing s/he needs is to discover these networks well enough to make ABC&S decisions in respect of the various services s/he may wish to access. Second, the user needs to know supplementary information about the services provided by these networks along with corresponding price policies. Third, s/he needs to find out what network is best for use for each particular (mobile) service (instance). Information about all these is provided via the WBC deployed in the area of presence. The user's mobile device can than compare this information with its own capabilities, the current user profile's preferences, the current location, and other contextual information such as the time of day, day of the week, etc. to select (in a

background mode) the 'best' services (instances) for use to achieve a particular goal (e.g. to make a phone call). As well as advertising the service, the WBC will also provide information to help with the process of discovering and associating with that service (e.g. access networks' physical layer information), [Flynn, 2006]. All this is referred to as the Advertisement, Discovery and Association (ADA) procedure, which could be (almost) fully automated and completed transparently to the user.

WBCs are wireless equivalent of roadside advertisement billboards and as such could be used as 'push' advertisements means for providers (ANPs and xSPs) to let users know about their presence and current service offerings. After receiving this information in the form of WBC broadcasts, the user's mobile device would be able to dynamically compile information about available services in the current location and to match these service offerings against ABC&S criteria under different user/device roles/profiles, all in all facilitating the ABC&S network-service match decisions and proposing ABC&S solutions to the user through (optional) device reconfigurability and application service adaptability functionalities. The user then, according to one of his/her roles, e.g. family parent, will select the available 'best' access network for a particular service and the 'best' service instance, using criteria such as price/performance ratio and current location. For instance, different mobile service and access network will be selected to call a family member (i.e. VoIP over Wi-Fi) than to call a business partner (i.e. an ordinary phone call via a 3G/4G cellular network), based on time/day/week/location configuration.

In the future, mobile devices will be able to select any access network they consider being the 'best' among all available access networks in the current area, even those whose communications technologies, protocols, etc. are not supported (by default) by the device itself. This is because mobile devices will become more and more reconfigurable, and eventually be able to access any and all existing and new access networks, and because users will more and more want autonomy in availing of access networks' communications services as with any other consumerist services [O'Droma, 2004b]. Through the Software-Defined Networking (SDN) technology, networks will also become more reconfigurable in order to match users' (individual and collective – groups', communities') variable mobile wireless service needs and desires thus achieving the ABC&S reconfigurability goal.

The following are the main WBC characteristics and related attributes [Flynn, 2006]:

- *Point-to-multipoint (broadcast)*

  A WBC, deployed in a particular area, will deliver service advertisements to multiple (all active) mobile devices, currently located in the same area.

Simplex

The simplex (i.e. unidirectional) attribute has the additional benefit of easing WBC physical deployment and operation. If the channel were duplex, then bandwidth-spectrum allocation will

become a much more significant issue, closer to the complexities involved in existing cellular spectrum allocations.

- *Limited bandwidth*

  Given the proposed usage – point-to-multipoint, unidirectional service of advertisements –, bandwidth requirements will be relatively narrow. This has the added advantage of enhancing WBC likely success, e.g. of global agreements on spectrum allocations for WBC.

- *Maximum coverage area*

  The WBCs should ideally be available anywhere-anytime. No matter where it is, a mobile device should have the ability to discover what services are available to it from local to regional and international service providers. Device mobility should not affect the ability to receive information on the channel.

- *Different versions for different areas*

  The number and types of WBCs could eventually correspond to the local, regional, national, and international interests of advertisers and users. For instance, there could be one national WBC channel, advertising all the services that are relevant on a national level, which could include advertisements of local, regional or interregional significance, and then separate regional WBC channels, advertising the services available in that particular region (Figure 2).

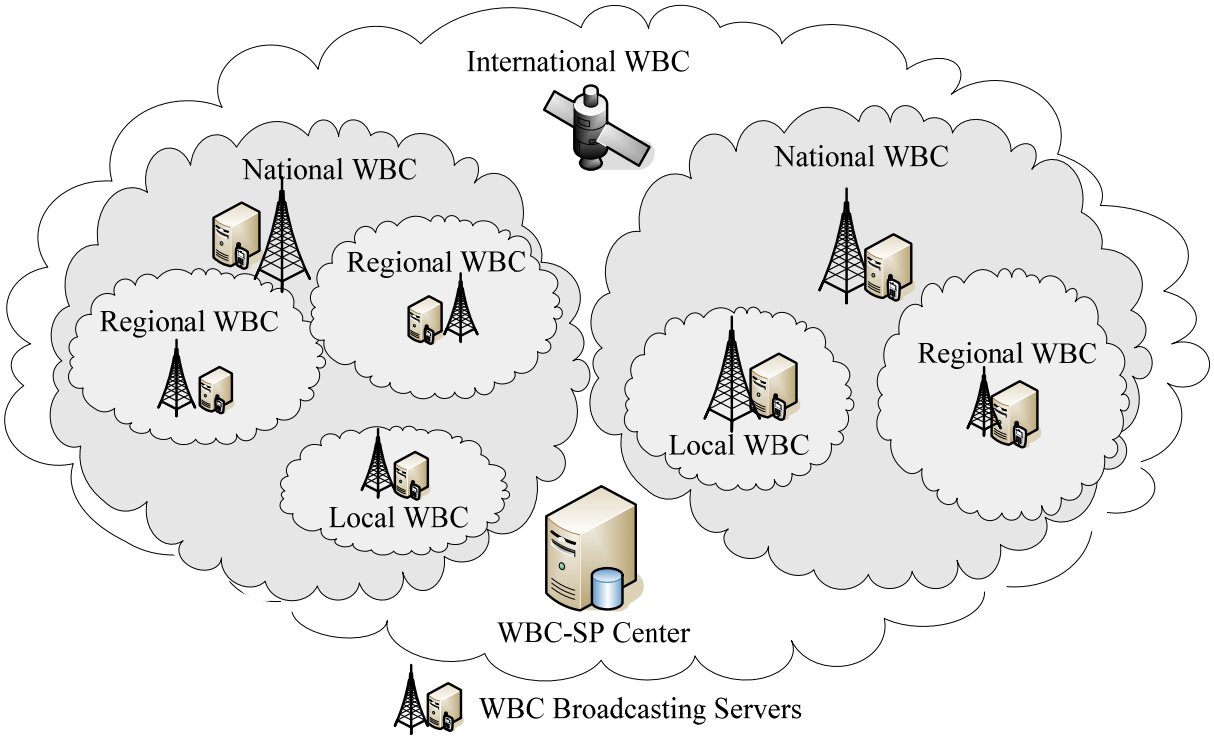


Fig. 2. Different WBC versions (adapted from [Ji, 2007])

- *Operated by non-ANP service providers*

  WBCs will need to be regulated and be fully independent and physically separate from ANPs and their access networks. This is needed to ensure fair competition and equity of access to WBC advertisement space, i.e. equally open to all ANPs. For this it is better that they be operated by non-ANP service providers, e.g. by existing radio and TV broadcasters.

- *Carrier technologies*

  There are several seemingly suitable broadcast technologies to consider, which fall into two main categories – terrestrial and satellite. Terrestrial examples include Digital Audio Broadcast (DAB), Terrestrial Digital Multimedia Broadcast (T-DMB), Digital Radio Mondiale (DRM), Digital Video Broadcast Handheld (DVB-H), Multimedia Broadcast / Multicast Service (MBMS). Satellite examples include Satellite DMB (S-DMB) or the Digital Audio Radio Satellite technologies being used by, for example, the WorldSpace system, XM Radio and Sirius.

The main WBC purpose is to allow services[1] to be discovered by mobile users/devices. The standard approach used by service discovery protocols, such as Jini, SLP, and Salutation, relies on a central registry of service descriptions (SDs). Service providers register their SDs with that registry. Clients query the central registry about available services, based on service attributes. The central registry responds to the clients with SDs that match their queries. The clients then can start using the services they discovered.

A modified version of this model was elaborated for use in the WBC by taking into account its specifics [Flynn, 2006]. As the WBC is a simplex "push" channel which does not facilitate queries to a registry, the solution was to broadcast all SDs in turn on the channel and the mobile device just to wait for the required SD to be broadcast. The WBC service discovery model is shown in Figure 3 and described below:

A. All service providers (ANPs and xSPs) register the SDs of their services with the WBC service provider's (WBC-SP's) central registry.

[1] The term 'service' here means both access networks' communications services and mobile services. The former are the actual access networks through which mobile devices connect. The latter is an encompassing term for all non-access-network services, e.g. from e-learning, e-government portals to on-line Internet shopping, e-mail, web-browsing, peer-to-peer services, etc. To use a mobile service, a mobile device will utilize an access networks' communications service.

B. The WBC-SP broadcasts all collected (and paid) advertisements / SDs, repeatedly, on a WBC. The advertisement structure should be flexible enough to include all, or at least all relevant, SDs. Advertisements should be streamed cyclically with frequency dictated by the WBC-SP's business model.

C. Each mobile device (MD) tunes to the channel and collects all SDs that the mobile user (MU) is interested in.

D. The MU/MT may seek further information (e.g. following a URL in the advertisement), makes a choice of the 'best' ANP for a service it requires, associates with that ANP, and begins to use the 'best' instance of that service.

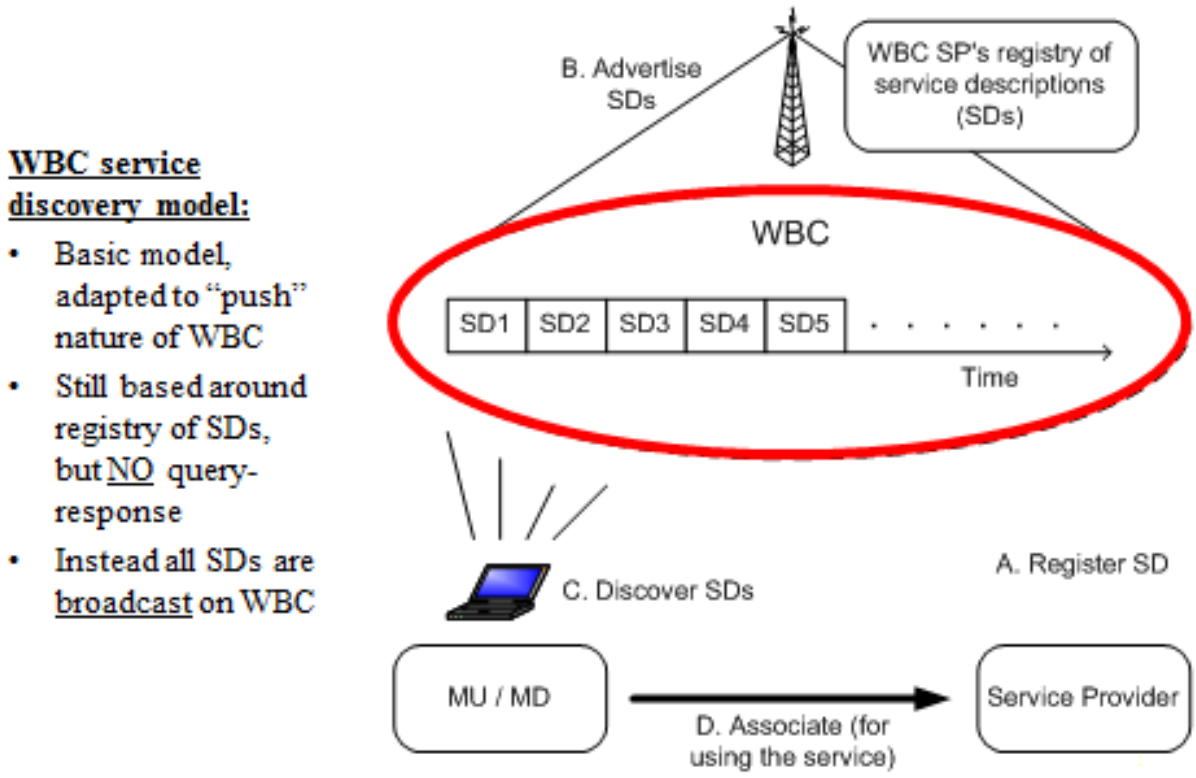


Fig. 3. The WBC service discovery model (adapted from [O'Droma, 2012])

The service advertisements, which are broadcast on the WBC, are composed of SDs. A three-part design of the SD was adopted in [Flynn, 2006]:

Service Type

The purpose of this field is to group together similar services, making the advertised services (and their SD) easier to find in the WBC streaming. Every service to be advertised has an assigned *Service Type*. Each *Service Type* has a template, which SDs follow. These templates are managed by the WBC-SP and published for all providers to follow.

Scope List

This field identifies which scopes a particular service belongs to. Scopes are a way to group services together. For example, a service provider may offer a number of news alert services to mobile users who pay a monthly subscription. All these services can then be assigned the same scope. The users, when discovering services using the WBC, will see their scope and will know they have access to them. Non-subscribing users will see that these services have a scope, which they have not been assigned, and can ignore any such SDs.

Attribute List

This field carries structured information on a service being advertised. For ANP's wireless access services, sufficient information to enable a mobile user/device to associate with the access network should be present in these SD attributes. To this end, SD templates, with relevant attribute lists, for different service types are formulated in [Flynn, 2006]. The format of the *Attribute List* in a SD depends upon the *Service Type* of that SD. Each *Service Type* has a service template specifying the format of the *Attribute List*. Templates are managed by the WBC-SP.

Making decisions between different service instances is semi-automatic. The mobile device reads the SDs from the WBC and provides relevant information to the user. For a given service type, that the user is interested in, the device will show all available service instances (sorted in order of preference based on attributes such as cost, quality, and supported features specified in the service profile) and among these the user will choose the 'best' service instance for the desired service. After that, the mobile device will need to know how to associate with that service (instance). The first thing needed is the client-side software to be installed on the mobile device (if not already installed/pre-loaded). For this, an

attribute is envisaged, which to tell the device how to download this software. Another attribute specifies the software itself and its version so that the device can see if it needs to be updated. It could be better to have one software-download service (which would also be advertised as any other service on the WBC), which allows for downloading of all additional software needed for the use of services advertised on the WBC (this could be used for software defined radio, SDR, downloads as well). Having installed the software, there are some attributes specific to a particular service (e.g. IP addresses and port numbers) that need to be known as well.

For the SD encoding, the use of the Abstract Syntax Notation (ASN.1) was proposed in [Flynn, 2006]. ASN.1 offers various different encoding rules, among which the Packed Encoding Rules (PER) are the most efficient that can yield encoding close to optimal. This is important for the WBC as bandwidth is a big issue.

An efficient system for SD advertisement, collecting, clustering, scheduling, indexing, discovery, and association was elaborated in [Ji, 2008a] along with a novel Advertisements Delivery Protocol (ADP) [Ji, 2008b]. In addition to collecting SDs from service providers, the other goal of *collecting* is to provide information about the user demand for a particular service and the advertisement cost paid by the corresponding service provider, which has to be taken into account when generating the SD broadcasting frequencies. The goal of *clustering* is to group SDs into WBC segments in an optimal way so as to reduce the user's access time and tuning time. The goal of *scheduling* is to apply a reasonable scheduling scheme for minimizing the access time of the entire system. Scheduling could be based on a push-based non-flat broadcasting, which will broadcast the more popular SDs more often. To achieve this, the WBC segments could be divided into two groups – *hot segments*, which are broadcast more frequently depending on the current user demand (i.e. several times per broadcast cycle), and *cold segments*, which are inserted equally into the scheduling cache by filling the remaining gaps (i.e. ones per broadcast cycle). A modified Broadcast Disks algorithm was developed to accomplish this goal more efficiently than the classic Broadcast Disks algorithm [Ji, 2008a].

Tuning time could be reduced by employing an *indexing* scheme because without it, a mobile device would have to tune into the WBC and listen to the broadcast continuously until the required SD is transmitted. By adding indexing data to the broadcast, mobile devices can tune in, find out when the required SD will be transmitted, then tune out and wait until that time to tune back in again. By adding redundant data to the broadcast, however, the average access time will be increased. Suitable indexing schemes, providing a good trade-off between the tuning time and access time, were investigated in

[Flynn, 2006] and an indexing scheme adjusted to the WBC specifics was proposed in [Ji, 2008a], based on the *(1, m)* indexing algorithm.

The goal of *discovery* and *association* is to discover and associate with the 'best' service instance by utilizing the information stored in various user's profiles, such as an identification/authentication profile, an advertisement profile, a discovery profile, an association profile, a rules profile, a history profile, etc.

To smooth the SD processing in the WBC, a new reliable and scalable ADP protocol was elaborated, based on the standard Asynchronous Layered Coding (ALC) protocol, to convert WBC segments into IP packets. The designed ADP protocol includes Building Blocks (BBs) and Protocol Instantiation (PI). Four modified BBs (i.e. Layered Coding Transport BB, Forward Error Correction / FEC BB, Congestion Control BB, and authentication BB) and two types of PIs (i.e. ALC using FEC) and NACK Oriented Reliable Multicast (NORM) relying on FEC with ARQ) were developed for the ADP in [Ji, 2008b].

A pilot 'WBC over DVB-H' prototype system, based on a 3-layer architecture and utilizing novel algorithms, schemes, and protocols, was developed, evaluated, and tested [Ji, 2010b]. Besides the design and implementation of a layered, distributed, intelligent, and heterogeneous WBC system prototype, the research work to date included the proofing and refining of different aspects of the design. Also completed is the operational testing, performance evaluation, and scalability evaluation of the core WBCs elements.

Information about the WBC has appeared in a recent International Telecommunication Union Radiocommunication Sector (ITU-R) report [ITU-R, 2015] as a realization of a coverage-area out-band CPC, piggybacked on a broadcast digital platform.

3. Service Recommendation System

Another possibility to recommend (mobile) services to consumers-users is to use a dedicated service recommendation system, like the one described in [Ganchev, 2013] as a means for users matching their need to discover the 'best' service instances, and facilitating, and supporting, the association with them by following a user-driven ABC&S paradigm.

The area of service recommendations has attracted great attention in the last few years. For instance, [Lee, 2010] proposes a personalized digital TV program recommendation system, working within a

cloud computing environment, which is able to analyze and use the viewing pattern of consumers in order to personalize the TV program recommendations. A personal photo recommendation system is proposed in [Tian, 2013] by fusing contextual and textual features on mobile devices. A music recommendation system, based on analysis of users' sentiments extracted from sentences posted on social networks is presented in [Rosa, 2015]. In [Songhui, 2012], a context-aware architecture of a car navigation recommendation system is described, which computes and dynamically adjusts the optimal travel path(s), based on real-time traffic information. [Wu, 2014] describes an intelligent urban car parking recommendation system for facilitating drivers with fully efficient, real-time and precise parking lot guiding suggestions. [Zhang, 2015] applies a semi-automated, extensible, and ontology-based approach for the discovery and selection of Infrastructure-as-a-Service (IaaS) cloud offers, by utilizing a multi-criteria decision-making technique, based on real-time end-to-end quality of service (QoS) parameters, for meeting service-level agreements (SLAs). [Nagarathna, 2012] proposes a service recommendation system, based on trust, reputation, and QoS requirements, for use in a Service Oriented Grid (SOG) by utilizing a mechanism of similarity computation and ranking of service providers based on users' feedback. [Pääkkö, 2012] applies knowledge-based recommendation techniques for dynamic, runtime, proximity-based service compositions for mobile devices.

However, the service recommendation system presented here is considered as a global solution applicable to all types of mobile services and also to many other Internet services. Taking into account the 'big data' aspect of information about (and gathered from) consumers, networks, and services, a cloud-based version of such a recommendation system is envisaged as being more capable for facilitating the delivery of increasingly contextualized mobile services to support the consumer-choice optimization process.

A UCWW mobile application [Ganchev, 2015a], installed on the user's mobile device and associated with such a system, could be used for finding and recommending to users, or even automatically selecting if the user's profile settings are so set, the 'best' mobile services, depending on the current context, including in that decision process the user's personal profile requirements. The complex functional requirements of such application make for a demanding app design, testing, and validation. A possible design solution, realized through a structured composition of three tiers – a mobile application tier, a web tier, and a cloud tier, is presented in [Ganchev, 2014a].

On the back-end, a UCWW cloud is envisaged to facilitate the storage of user data harvested via mobile devices, and based on the analysis of this data, to offer predictions as to the applicability and ABC&S

suitability of services to particular users, and to enable ever-enhanced contextualization and personalization functionalities. Over time the data collected relating to particular users can give an accurate view of particular cohorts, based on common interests, repetitive access of particular services, etc. By monitoring this information, the system then can accurately predict the types of services most applicable to individuals, and in turn, recommend these to them. Furthermore, efficient algorithms must be applied to facilitate service utilization predictions locally on the mobile devices or as part of the UCWW cloud as an alternative to mining the stored data [Ganchev, 2015b]. For instance, [Zhang, 2016] proposes a hybrid method that integrates user trust relations with item-based collaborative filtering. This is achieved by incorporating user social similarities into the computation of item similarities. Performance evaluation results demonstrate that the proposed approach achieves better accuracy than the traditional item-based collaborative filtering.

The UCWW cloud could be established to operate as a middleware. At the lowest layer, the user's mobile device collects context data from the environment, and at the highest layer the UCWW client application makes use of this data. Between them operates the middleware of the system, which could be entirely implemented as cloud services. [Ganchev, 2013] describes the flow of context data between a mobile device and the UCWW cloud as well as the mechanism of sending requests and receiving responses from the decision support subsystem, i.e. providing ratings (ranking) of the service providers available for a particular type of service requested by the user.

This service recommendation system could be deployed as an 'anywhere-anytime-anyhow' oriented component, supplemented by a Data Management Platform (DMP) [Ganchev, 2016b] that acts as a machine learning platform for turning raw data into actionable analytic dataset, i.e., user behavior profiles, including user preferences, content consumption preferences, shopping preferences, interest preferences, app usage, etc., abiding by the user-privacy principles. More specifically, the DMP provides data collecting, processing, analyzing, and consumer targeting operations. It could be used for managing consumer identification and generating audience segments, in order to target consumers with most appropriate / 'best' (mobile) services. For this, it utilizes real-time user's profiling algorithms and off-time data processing algorithms [Ganchev, 2016a], and could be implemented with the Publish/Subscribe design pattern [Ganchev, 2016b].

The service recommendation system communicates with the DMP, keeps updating the user behavior profiles, user interests and requirement tendencies, and sends a personalized list of 'best' recommended service instances to each user in a real-time manner by utilizing relevant

recommendation algorithms and updated recommendation rules. A recommendation engine acts a central element in this system. It allows uploading the recommendation algorithms to the system and defining/updating the recommendation rules. This engine can be built with a Lambda Architecture for providing real-time recommendations (at the speed layer) and off-time analytical operations (at the batch layer) [Ganchev, 2016a].

The service recommendations, provided by such a system, will depend greatly on the current context. Besides the context that relates to the mobile *services* available on offer (i.e. the category, type, scope and attributes of the service, the request time, the application initiating the request, the current Quality of Service / Quality of Experience (QoS/QoE) index of the service, price, etc.), the context data may relate to the *user* (e.g., the user location, the user preferences and profiles, current battery charge and other operational characteristics of the user's mobile device, type of activity, intentions, social interests, the upper bound on the price and the lower bound on QoS/QoE accepted by the user for each particular service, privacy and security requirements, etc.), and/or relate to the constraints of the wireless access *network* currently utilized by the user (e.g., the communication channel state information (CSI), network congestion level, the current data usage pattern, the current QoS/QoE index, the cost of using the network, pricing scheme, etc.). Then determining the 'best' service instance at any moment for a particular user is based on a set of context parameter values, categorized in three groups – user-related ***(u)***, service-related ***(s)***, and (access) network-related ***(n)***, forming a 3D ***(u,s,n)*** context space, as illustrated in Figure 4. The selection of the 'best' service instance ***s*** for user ***u*** in the network context ***n*** is based on finding the following maximum value [Ganchev, 2014b]:

$$Max_{s_1 \ldots x} \sum_{i=1}^{n} Best(\mathrm{u}_i, \mathrm{s}_i, \mathrm{n}_i)$$

The concept of context allows making smart decisions based on mining of data stored in cloud repositories. [Ganchev, 2013] proposes context to include both the data sensed in the environment (as in a typical context-aware system), and the history of the user and the collective history of users who have acted in a similar environment. This constitutes a novel approach in providing context-aware services with elements of community-based personalized information retrieval (PIR), applied to mobile network environments.

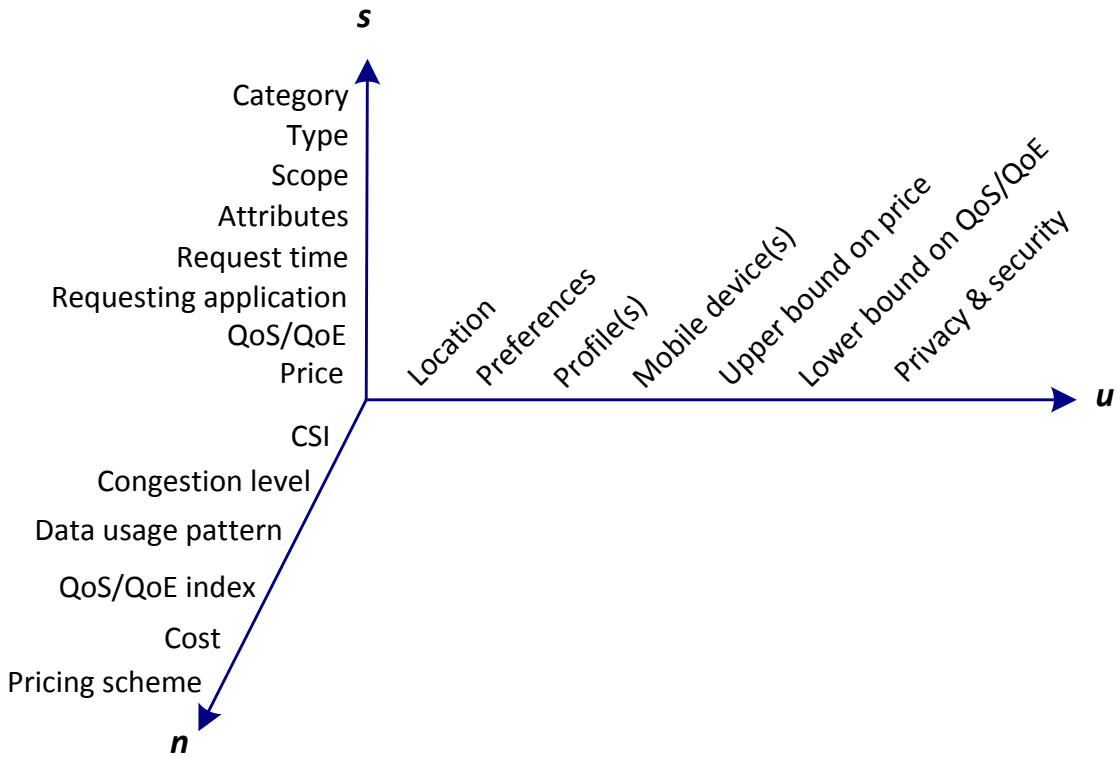


Fig. 4. The 3D context space.

4. Intelligent Framework for Recommendation of Mobile Services to Consumers

By combining the two aforementioned service advertisement approaches, an intelligent framework could be built with a modular structure consisting of four main parts, used for stream computing, SD server hosting, cloud-based data mining, and distributed log collection, respectively (Figure 5), [Ganchev, 2014b].

Among these, the SD server is the main part. It collects aggregated user behavior monitored by the UCWW client app installed on the user's mobile device. The monitored behavior includes user's searching for services, user's associations with services, actual service usage, etc. In the first step, the SD retrieval module searches for the relevant keyword in the SD index database, facilitated by the user's profile, web page attributes database, and real-time bidding system for demand (RTBD), and as a result an initial service recommendation list is generated. The list is then pushed to SD ranking module and, after computing with the click-through rate (CTR) module, a final list is generated. The WBC management module collects the list, sends it back to the user (via the UCWW server), and pushes it to the log data real-time collection part. If SDs are for new/popular/emerging services, these SDs will be

cached in the WBC management module for broadcasting on the WBC. The memory key-value NoSQL database Redis (http://redis.io) could be used to provide persistence operation and the Ngnix (http://nginx.org) – for load-balancing services in the web tier [Ganchev, 2014b].

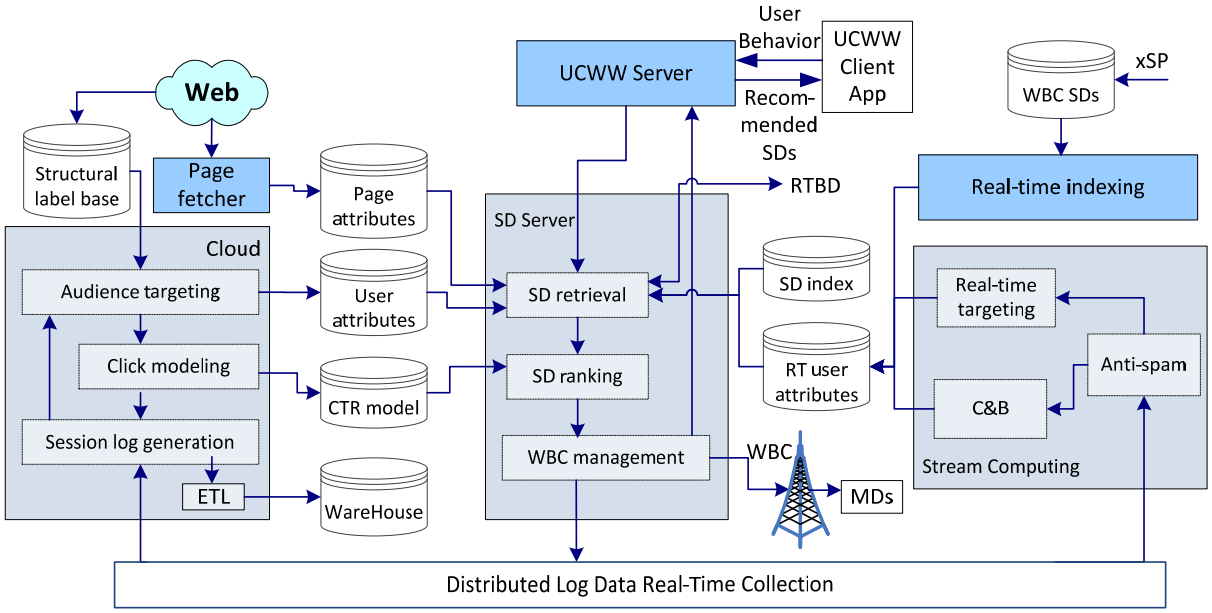


Fig. 5. The intelligent service recommendation framework [Ganchev, 2014b].

The distributed log data real-time collecting part utilizes a producer-consumer paradigm to exchange data, i.e., the SD server is a producer, and the cloud and the stream computing parts are consumers. The Apache Flume (http://flume.apache.org/) could be used to implement this part [Ganchev, 2014b].

In the cloud, the corresponding session data is generated with a unique user ID and serialized to the data warehouse by the extraction transformation loading (ETL) module. Then the audience targeting function collects the behavior and updates it to the user attributes key-value database and CTR model database. Besides collecting the log data from the ad server, the cloud also fetches the corresponding service's web page, updates it to a page attributes database, labels the keywords, and saves the new information to a structural label database. A Hadoop environment could be utilized to provide input/output, remote procedure call (RPC), and Map/Reduce functions [Ganchev, 2014b].

The stream computing part is a real-time computing system, which uses an anti-spam function to filter user's behavior, integrates a charging and billing (C&B) function for services, and updates the behavior to the real-time user attributes Redis database.

4. Conclusion

This paper has reported on the development of a cloud-based next generation networking (NGN) conceptual framework for the supply of service recommendations to consumers (mobile users), built on the revolutionary concept for the realization of the next phase of a NGN-based consumer-oriented wireless networking, founded on the key attributes of the Ubiquitous Consumer Wireless World (UCWW).

The described cloud-based framework provides a personalized data collecting, processing, analyzing, and consumer targeting functions. It could be used to manage consumer identification and generate audience segments in the UCWW, in order to target consumers with the 'best' suitable mobile services they might be interested in, thus facilitating the realization of a truly consumer-centric Always Best Connected and best Served (ABC&S) 'anywhere-anytime-anyhow' provision.

The framework has a modular structure and as such contains a supplementary module for feeding with updated information a Wireless Billboard Channel (WBC), proposed as a global solution for services' Advertisement, Discovery and Association (ADA) in the UCWW.

The integration of such semantic-based recommendation framework into the UCWW has the potential of creating an infrastructure in which consumers-users will have access to (mobile) services with a radically improved contextualization. As a consequence, the framework is expected to radically empower individual consumers in their decision making and thus positively impact the society as a whole by facilitating and enabling direct consumer–service provider relationships. Besides benefitting the consumers, this will open up the opportunity for stronger competition between providers, therefore creating a more liberal, more open, and fairer marketplace for existing and new providers, in which they can deliver a new level of services that are both much more specialized and reaching a much larger number of consumers (mobile users).

Future research work will seek further elaboration of the design, followed by implementation, testing, and evaluation of a fully operational system prototype, employing an efficient and effective relevance measurement approach for the UCWW heterogeneous service network, along with an effective graph-based feature extraction method for building the consumer profiles for facilitating real-time service recommendations, supplied to consumers in order to discover and choose the 'best' (mobile) service instances, under the ABC&S communications paradigm. With this design, the resultant framework will be flexible, scalable, and could be easily integrated into the UCWW cloud.

Acknowledgements

This work is funded and supported by the Telecommunications Research Centre (TRC), University of Limerick, Ireland and by the NI15-FMI-004 project of the Scientific Fund of the Plovdiv University "Paisii Hilendarski", Bulgaria.

The author wishes to specially thank the TRC Director, Dr. Máirtín O'Droma, for the fruitful discussions and inspiring thoughts on the subject through many years.

Bibliography

[Bourse, 2007] D. Bourse, R. Agusti, P. Ballon, P. Cordier, S. Delaere, B. Deschamps, D. Grandblaise, A. Lee, P. Martigne, K. Moessner, M. Muck, O. Sallent. The E2R II Flexible Spectrum Management (FSM) Framework and Cognitive Pilot Channel (CPC) Concept - Technical and Business Analysis and Recommendations, E2R II White Paper 2007. Pp. 1–52.

[ETSI, 2009] Reconfigurable Radio Systems (RRS) – Cognitive Pilot Channel (CPC). ETSI TR 102 683 V1.1.1, 2009, http://www.etsi.org

[Flynn, 2006] P. Flynn, I. Ganchev, M. O'Droma. "Wireless Billboard Channels: Vehicle and Infrastructural Support for Advertisement, Discovery, and Association of UCWW Services," In: Annual Review of Communications, Vol. 59 (Chicago, Ill.: International Engineering Consortium), Pp. 493–504. 2006. ISBN: 978-1-931695-53-4.

[Ganchev, 2013] I. Ganchev, M. O'Droma, N. Nikolov, Z. Ji. "A UCWW Cloud System for Increased Service Contextualization in Future Wireless Networks" (invited paper). Proc. of the 2nd International Conference on Telecommunications and Remote Sensing (ICTRS'13), Pp. 69-78. 11-12 July 2013, Noordwijkerhout, The Netherlands. ISBN: 978-989-8565-57-0.

[Ganchev, 2014a] I. Ganchev, Z. Ji, M. O'Droma. "A Cloud-based Service Recommendation System for Use in UCWW". Proc. of the IEEE 11th International Symposium on Wireless Communication Systems (IEEE ISWCS'2014). Pp. 791-795, 26-29 August 2014, Barcelona, Spain. ISBN: 978-1-4799-5863-4/14. DOI: 10.1109/ISWCS.2014.6933461.

[Ganchev, 2014b] I. Ganchev, M. O'Droma, Z. Ji. "UCWW Pilot System Prototype". Proc. of the 17th Royal Irish Academy Research Colloquium on Communications and Radio Science into the 21st Century, Pp. x.1-x.4. 30 April - 1 May 2014, Dublin, Ireland. Proceedings Book (©RIA; ISBN: 978-1-908996-33-6).

[Ganchev, 2015a] I. Ganchev, Z. Ji, M. O'Droma. UCWW Cloud-based ABC&S Mobile App. Proc. of the URSI Atlantic Radio Science Conference 2015 (URSI AT-RASC 2015). 18-22 May 2015, Gran Canaria, Canary Islands.

[Ganchev, 2015b] I. Ganchev, Z. Ji, M. O'Droma. "Making the UCWW a Reality". Proc. of the 2015 IEEE International Symposium on Technology in Society (IEEE ISTAS 2015). Pp. 1-4, 11-12 November 2015, Dublin, Ireland. ISBN: 978-1-4799-8283-7. DOI: 10.1109/ISTAS.2015.7439435.

[Ganchev, 2016a] I. Ganchev, Z. Ji, M. O'Droma. "A Conceptual Framework for Building a Mobile Services' Recommendation Engine". Proc. of the IEEE International Conference 'Intelligent Systems' (IEEE IS 2016). Pp. 285-289. 4-6 September 2016, Sofia, Bulgaria. ISBN: 978-1-5090-1353-1/16.

[Ganchev, 2016b] I. Ganchev, Z. Ji, M. O'Droma. "The Creation of a Data Management Platform for Use in the UCWW". Proc. of 2016 SAI Computing Conference. Pp. 585-588. 13-15 July 2016, London, UK. ISBN:978-1-4673-8460-5/16. DOI: 10.1109/SAI.2016.7556040.

[ITU-R, 2015] Radiocommunication Sector of the International Telecommunication Union (ITU-R). Cognitive radio systems in the land mobile service. REPORT ITU-R M.2330-0. M Series: Mobile, radio determination, amateur and related satellite services. 69 pp. Geneva, Switzerland. 2015.

[Ji, 2007] Z. Ji, I. Ganchev, M. O'Droma. "On WBC Service Layer for UCWW". Proc. of the 9th IFIP International Conference on Mobile and Wireless Communications Networks (IFIP MWCN'07), Pp. 106-110, 19-21 September 2007. Cork, Ireland. ISBN 978-1-4244-1719-3.

[Ji, 2008a] Z. Ji, I. Ganchev, and M. O'Droma, "Efficient Collecting, Clustering, Scheduling, and Indexing Schemes for Advertisement of Services over Wireless Billboard Channels". Proc. of International Conference on Telecommunications (ICT 2008), St. Petersburg, Russia, pp. 225-230, 16-19 June, 2008.

[Ji, 2008b] Z. Ji, I. Ganchev, M' O'Droma. "Reliable and Efficient Advertisements Delivery Protocol for Use on Wireless Billboard Channels". Proc. of the 12th IEEE International Symposium on Consumer Electronics (IEEE ISCE 2008), Pp. 261-264, 14-16 April 2008. Algarve, Portugal. ISBN 978-1-4244-2422-1.

[Ji, 2010a] Z. Ji, I. Ganchev, P. Flynn, M. O'Droma. "Formal Description of Services for Advertisement on Wireless Billboard Channels". Proc. of the IEEE 14th International Symposium on Consumer Electronics (IEEE ISCE2010), Pp. 366-371, 7-10 June 2010. Braunschweig, Germany.

[Ji, 2010b] Z. Ji, I. Ganchev, M. O'Droma. "A `WBC over DVB-H` Prototype System". Royal Irish Academy Research Colloquium on Wireless as an Enabling Technology - Innovation for a Critical Infrastructure, 22 April 2010. Dublin, Ireland. Proceedings Book (©RIA; ISBN 978-1-904890-68-3), Pp. 34-38.

[Lee, 2010] S. Lee, D. Lee and S. Lee, "Personalized DTV program recommendation system under a cloud computing environment," IEEE Transactions on Consumer Electronics, vol. 56, no. 2, pp. 1034-1042, May 2010.

[Nagarathna, 2012] N. Nagarathna, M. Indiramma and J. S. Nayak, "Optimal Service Selection Using Trust Based Recommendation System for Service-Oriented Grid," Proc. of the 2012 International Symposium on Cloud and Services Computing (ISCOS), Mangalore, 2012, pp. 101-106.

[O'Droma, 2004a] M.S. O'Droma and I. Ganchev. "Enabling an Always Best-Connected Defined 4G Wireless World," In: Annual Review of Communications, Vol. 57 (Chicago, Ill.: International Engineering Consortium), Pp. 1157-1170. 2004. ISBN 0-931695-28-8.

[O'Droma, 2004b] M. O'Droma and I. Ganchev. 2004. "Techno-Business Models for 4G" (invited paper), Proc. of the International Forum on 4th Generation Mobile Communications, Pp. 3.5.1-30, 20-21 May, King's College London, London.

[O'Droma, 2006] M. O'Droma, I. Ganchev, H. Chaouchi, H. Aghvami, V. Friderikos. "`Always Best Connected and Served` Vision for a Future Wireless World". Journal of Information Technologies and Control, Year IV, No 3-4, 2006, Pp. 25-37+42. ISSN: 1312-2622.

[O'Droma, 2007] M. O'Droma and I. Ganchev. "Towards a Ubiquitous Consumer Wireless World". IEEE Wireless Communications, Feb. 2007, Pp. 2-13. ISSN: 1536-1284.

[O'Droma, 2010] M. O'Droma and I. Ganchev. "The Creation of a Ubiquitous Consumer Wireless World through Strategic ITU-T Standardization" (invited paper). IEEE Communications Magazine, Vol. 48, Issue 10, Pp. 158-165. October 2010. ISSN: 0163-6804. DOI: 10.1109/MCOM.2010.5594691.

[O'Droma, 2012] M. O'Droma, I. Ganchev, N. Nikolov, P. Flynn. 2012. "A Minimally Intrusive Wireless Solution for Context- and Service Awareness Enablement in Mobile Communications" (invited paper). Proc. of the First International Conference on Telecommunications and Remote Sensing (ICTRS'12), Pp. 118-128. 29-30 August, Sofia, Bulgaria. ISBN: 978-989-8565-28-0.

[Pääkkö, 2012] J. Pääkkö, M. Raatikainen, V. Myllärniemi and T. Männistö, "Applying Recommendation Systems for Composing Dynamic Services for Mobile Devices," Proc. of 19th Asia-Pacific Software Engineering Conference, Hong Kong, 2012, pp. 40-51.

[Rosa, 2015] R. L. Rosa, D. Z. Rodriguez and G. Bressan, "Music recommendation system based on user's sentiments extracted from social networks," IEEE Transactions on Consumer Electronics, vol. 61, no. 3, pp. 359-367, Aug. 2015.

[Songhui, 2012] L. Songhui, S. Junqing, Y. Zhe and Z. Xuan, "On key technology of service recommendation system for car navigation," Proc. of 31st Chinese Control Conference (CCC), Hefei, 2012, pp. 7298-7303.

[Tian, 2013] Y. Tian, W. Wang, X. Gong, X. Que and J. Ma, "An enhanced personal photo recommendation system by fusing contextual and textual features on mobile device," IEEE Transactions on Consumer Electronics, vol. 59, no. 1, pp. 220-228, Feb. 2013.

[Wu, 2014] E. H. K. Wu, J. Sahoo, C. Y. Liu, M. H. Jin and S. H. Lin, "Agile Urban Parking Recommendation Service for Intelligent Vehicular Guiding System," IEEE Intelligent Transportation Systems Magazine, vol. 6, no. 1, pp. 35-49, Spring 2014.

[Zhang, 2015] M. Zhang, R. Ranjan, M. Menzel, S. Nepal, P. Strazdins, W. Jie and L. Wang, "An Infrastructure Service Recommendation System for Cloud Applications with Real-time QoS Requirement Constraints," IEEE Systems Journal, vol. PP, no. 99, pp. 1-11. 2015.

[Zhang, 2016] H. Zhang, I. Ganchev, N. S. Nikolov, M. O'Droma. "A Trust-Enriched Approach for Item-Based Collaborative Filtering Recommendations". Proc. of the IEEE 12th International Conference on Intelligent Computer Communication and Processing (2016 IEEE ICCP). Pp. 65-68. 8-10 September 2016, Cluj-Napoca, Romania. ISBN: 978-1-5090-3899-2/16.

Author's Information

Ivan Ganchev *– DipEng (summa cum laude), PhD, SMIEEE, ITU-T (Invited Expert), IJTMCC Regional Editor (Europe).*

TRC Deputy Director, University of Limerick, Limerick, Ireland & Associate Professor, Plovdiv University "Paisii Hilendarski", Bulgaria; e-mail: Ivan.Ganchev@ul.ie

Major Fields of Scientific Research*: novel telecommunications paradigms, future networks and services, smart ubiquitous networking, context-aware networking, mobile cloud computing, Internet of Things (IoT), Internet of Services (IoS), Ambient Assisted Living (AAL), Enhanced Living Environments (ELE), trust management, Internet tomography, mHealth and mLearning ICT solutions.*

THE MODEL OF IT-STARTUP THAT GROWS IN UNIVERSITY ECOSYSTEM AND APPROACH TO ASSESS ITS MATURITY

Maxim Saveliev, Vitalii Lytvynov

Abstract: *This paper is dedicated to description of method of the maturity estimation of startup and spin-off IT-companies created with the help of University Business Centers as an approach for cooperation between universities and industrial companies. The control loop, based on the formation and subsequent evaluation of organizational maturity of IT-start-up based on CMMI model is shown. The conception of tool for assessment of IT companies maturity is proposed and theoretical ground for implementing it on the basis of the apparatus of fuzzy logic is provided.*

Keywords: *academic IT entrepreneurship, IT start-up, maturity of IT companies, fuzzy logic, CMMI.*

ACM Classification Keywords: *K.6.1 Management of Computing and Information Systems - Project and People Management*

Introduction

In the end of XX century lead western countries face to the problem of universities graduates not readiness to the requirements of industry and work market needs. One of the emerging problems in Europe is shortage of IT personnel. According to estimates of the European Commission in 2020, a shortage of skilled IT professionals in the EU could reach 825 thousand [European Commission, 2013]. And that happen when the number of graduates in the EU in the IT field is kept at the level of 100 thousand professionals per year.

One of the solution of that problem is involving University to cooperation with business. But this subject is not well studied especially in European countries that formerly were the part of the Soviet bloc and which have no tradition of entrepreneurship and the free market.

The formalized University-Business Cooperation (UBC) models were offered and the first practical experience of such cooperation was analysed by Prof. V.Kharchenko and Prof. V.Sklyar [Kharchenko, 2012]. The model of cooperation between University and companies through academic consortiums was presented in work of Prof. Y.Kondratenko [Kondratenko, 2015]. Interaction between Industry and

Universities was topic of the works of such scientists as C.Phillips, S.Lange, A.Tormasov, H.Edmondson et al.

One of the promising approach for UBC is the academic entrepreneurship when University acts as an ecosystem for growing new start-ups [Lytvynov, 2015]. As a rule Universities have a specific department called Business-Center (BC). Such BCs cultivate companies which main capital is new unique technology or “know-how”. In fact BCs specialized in providing services that is not directly related to company «know-how», but to the functioning of the company as a business structure.

The process of growing IT companies is multifaceted and difficult, especially if this company consists of young students united by one idea but completely without any real experience. On each stage of growing IT-startup from unformal group to independent business it requires from University making difficult decisions to provide or cancel support the company. Such decisions require objectively proved models, methods and tools to estimate different characteristics of IT-start-up and its project. And it should be noted that the problem of such tools is not well studied for now and need to be solved.

This paper will describe a model Academic IT-start-up – a newly created small company formed in University ecosystem which general activity belongs to creation of IT product or service and the tool to assess the maturity characteristics of IT-start-up as an IT business company.

The model of IT-start-up in University ecosystem

There are 4 main component of IT-startup described in work «Business models dynamics for start–ups and innovating e-businesses» by Bouwman at al. [Bouwman, 2007].

Product – that describe value proposal for the market.

Technology – the functionality required to implement the product.

Organization – structure of actors, people and stakeholders required to create the product.

Finance – mechanism of financial support product development and its entrance to the market.

Software Quality Institute in Texas defines three categories that must be managed with a view to the successful implementation of IT projects (creating software), they are: product, project and staff. Moreover, each category requires its own set of skills.

It should be noted that not all factors and characteristics inherent to mature and big companies are belong to IT-startups that grows in university ecosystem. It is possible to distinguish 5 main components that have direct impact to the goal state of IT-startup on its way to independent company. Here it is.

Team – group of individuals who work together for reaching common Project goal.

Technology – a set of methods, technics, equipment and knowledge required to implement the Product.

Processes – organization and business processes and functionality of the Team required managing the development of the Product at the defined level of quality.

Product – product or service that IT-startup wants to create and present to the market.

Project – Team activities, aimed at creating a unique Product (service).

Fig.1 is representing this model in graphics.

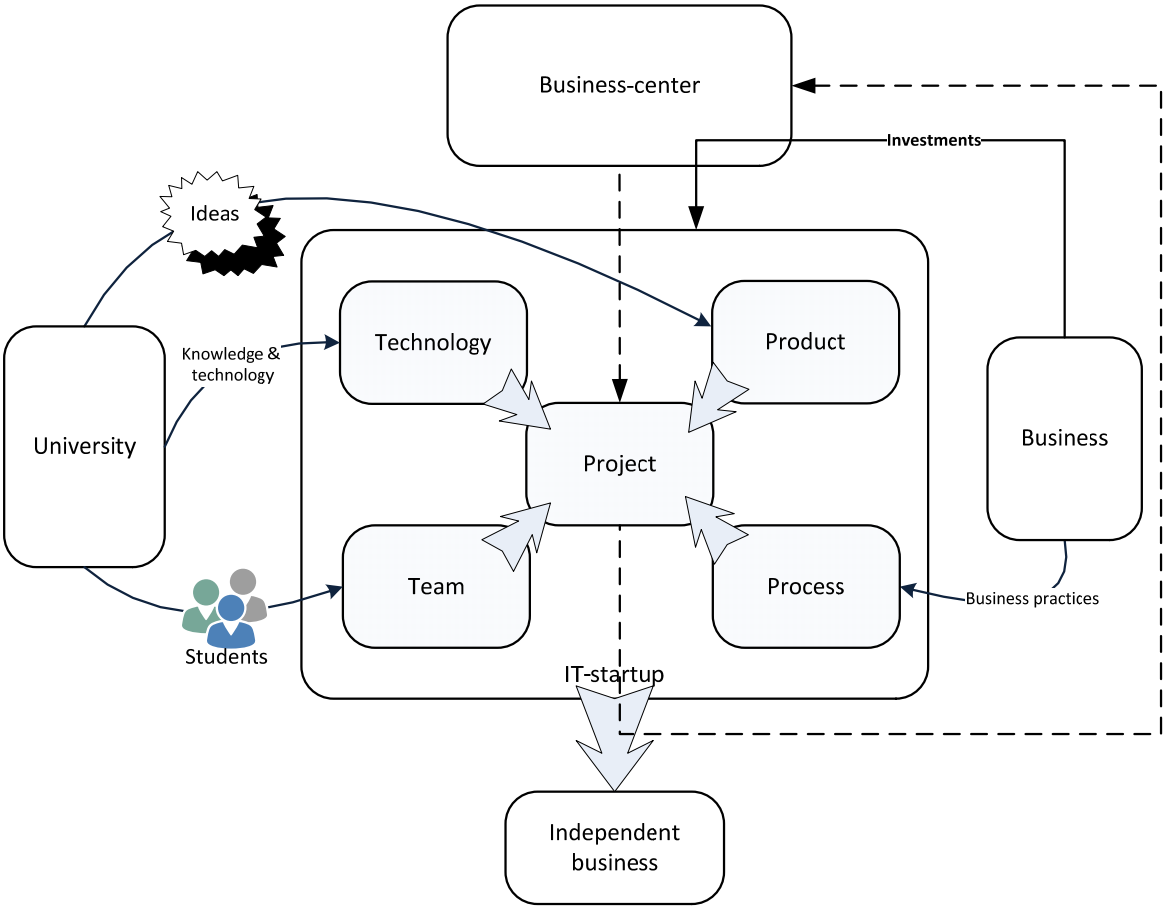


Figure 1. Conceptual model of academic IT-start-up

In university ecosystem, the University feeds startup with ideas for new business linked with modern technology and new know-how that usually born here. The University also feeds startup with students and young scientists who want to transfer their ideas to new business. Startups itself draw knowledge from University about modern technology and methods of its usage.

From its side, the Business provides financing and investments and also time proved business and engineering practices that helps to organize technological business processes in startup as for a business organization. Normally it happens directly or indirectly through University Business-Centre.

And here the control loop for BC to manage it-startup to the stage of independent business could be defined. This control loop uses indicators of project state like delays from baseline schedule or project risks and trends, and indicators of IT-start-up maturity like development in organization of different business processes.

In other words, the control action on the IT-startup will be implemented through the issuance of recommendations for inclusion in the project schedule works related to the improvement of its organizational maturity, production processes, as well as the necessary competences, both at the individual level of individual roles in the team and at the company level generally. These impacts will be supported by the terms of access to the resources of the business center (funding, laboratories, data centers, etc.). And BC will require a number of assessment tools for defining right control action on the IT-startup and its project. One is the maturity assessment tool of Academic IT-startup.

Conceptual model for the assessment of IT-startup maturity level

One well-known approach in assessment of business structures maturity is Capability Maturity Model Integration originally proposed as Software CMM in the 1989 by Watts Humphrey in his book “Managing the Software Process” [Humphrey, 1989]. Now this model is supported successfully by the Carnegie Mellon Software Engineering Institute (SEI).

CMMI groups Key Practices into Special Goals. Then Special Goals are grouped into Process Areas. Process Areas are linked to maturity levels. The development of the key practices of CMMI is evaluated by experts using linguistic variables that could take one value from the set {“not implemented”, “partially implemented”, “largely implemented”, “fully implemented”}.

It means that the level of maturity could be assessed using hierarchical fuzzy logic system [Kondratenko, 2011]. In this case maturity of IT companies will be determined by the following equation:

$$M = f\left(f_1\left(P_1, P_2, \ldots, P_7\right), f_2\left(P_8, P_9, \ldots, P_{18}\right)\right) \quad (1)$$

where

- M is the maturity of IT-startup;
- $f_1(\cdot)$ and $f_2(\cdot)$ are functions to assess maturity for 2 and 3 levels;
- $P_1, P_2, \ldots, P_{18}$ are the fuzzy value of development of process areas in CMMI-DEV model.

In turn, the development of each area of process will be determined by the relation:

$$P_i = h_i\left(g_{i1}\left(p_{i11},\ldots,p_{i1l}\right),\ldots,g_{in}\left(p_{in1},\ldots,p_{inm}\right)\right) \quad (2)$$

where

- $h_i()$ is a fuzzy function to output the development of process area with index ***i*** ;
- $g_{ij}()$ is a fuzzy function to determine how IT-startup reach Special Goal with index ***j*** for the Process area with index ***i***;
- p_{ijk} is a fuzzy linguistic variable for assessment of special practice with index ***k*** development in IT-startup that belongs to Special Goal with index ***j***, that in its turn linked with Process Area with index ***i***.

Let define NI, PI, LI, FI as fuzzy numbers for the linguistic variable p_{ijk} . This numbers are defined as fuzzy sets *N, P, L, F* belongs to the set of real numbers R. Carriers sets of *N, P, L, F* are defined as intervals on *R* – [n_L, n_R], [p_L, p_R], [l_L, l_R], [f_L, f_R] respectively.

There are functions $\mu_i(x)$ such that $\forall x \in R \mid \mu_i(x) \in [0,1]$ where $i \in \{N,P,L,F\}$ defined on *R* for each number NI, PI, LI, FI.

For fuzzy numbers with so called L-R type functions $\mu_a(x)$ that is defined by relation (3) it exists distance metric.

$$\mu_a(x) = \begin{cases} 0 & x < a1 \\ left\left(\dfrac{x-a1}{a2-a1}\right) & a1 \le x < a2 \\ 1 & a2 \le x \le a3 \\ right\left(\dfrac{a4-x}{a4-a3}\right) & a3 < x \le a4 \\ 0 & x > a4 \end{cases} \quad (3)$$

It was shown at Grzegorzewski works [Grzegorzewski, 1998] that best metric is expansion of the Euclidean distance, see relation (4).

$$d^2(a,b) = \int_0^1 \left(A_L(\alpha) - B_L(\alpha)\right)^2 d\alpha + \int_0^1 \left(A_U(\alpha) - B_U(\alpha)\right)^2 d\alpha \quad (4)$$

where the fuzzy number is defined by the concept of α-section, such that $\forall \alpha \in (0,1)\ \exists a1 \le x \le a4$ and $\mu_a(x) \ge \alpha$ and $A_L(\alpha) = \mu_{a\uparrow}^{inv}(x)$ - function reverse to $\mu_a(x)$ over the interval of increase and $A_U(\alpha) = \mu_{a\downarrow}^{inv}(x)$ - function reverse to $\mu_a(x)$ in the interval of its decrease.

Consider the function (5) to determine the reachability an IT company specialized goals in CMMI process area:

$$g(x_1, x_2, \ldots x_n) = x_1 \oplus x_2 \oplus, \ldots \oplus x_n \quad (5)$$

where the operation $\oplus$ is the operation of addition of fuzzy numbers according to the formula (6).

$$a \oplus b = \int_{a1+b1}^{a2+b2} \frac{\min\left(\mu_a(x), \mu_b(y)\right)}{x+y} \quad (6)$$

Achievement of special goal G_i of process development by the IT-startup could be described by the linguistic values: fully reached (FR); partially reached (PR); not reached (NR).

Let consider that special goal G_i is fully reached (G_i = FR) if the sum of experts answers regarding development of key practices as fuzzy number $g(x_1, x_2, \ldots x_n) = x_1 \oplus \ldots \oplus x_n$ close to fuzzy number $FI * n = FI_1 \oplus \ldots \oplus FI_n$.

Let consider that special goal G_i partially reached (G_i = PR) if fuzzy number $g(x_1, x_2, \ldots x_n) = x_1 \oplus \ldots \oplus x_n$ close to fuzzy number $Mean = (LI * n \oplus PI * n)/2$.

Similarly, we shall consider, in general, that the special goal G_i is not reached (G_i = NR) if fuzzy number $g(x_1, x_2, \ldots x_n) = x_1 \oplus \ldots \oplus x_n$ close to fuzzy number $NI * n = NI_1 \oplus \ldots \oplus NI_n$.

Let apply the same reasoning to assess the area of CMMI process development by IT-starup. The development process area could be defined as a sum of fuzzy values G_i obtained in the previous step. Let define development of Process area by another linguistic variable that could have one of the

following values: NF, PF, FF that corresponds to «area not mastered», «area mastered partially» и «area fully mastered». Which will be calculated in a general form by the following rules:

- Process area $F_j = NF$ if $\sum_{i=1}^{n} G_i$ close to fuzzy number $\sum_{1}^{n} NR$;
- Process area $F_j = PF$ if $\sum_{i=1}^{n} G_i$ close to fuzzy number $\sum_{1}^{n} PR$;
- Process area $F_j = FF$ if $\sum_{i=1}^{n} G_i$ close to fuzzy number $\sum_{1}^{n} FR$.

To assess the maturity level of IT-startup, functions $f()$, $f_1()$, $f_2()$ from equation (1) should be defined. For functions $f_1()$ and $f_2()$ similar reasoning could be applied, defining it as a sum of process areas development. But similar reasoning could not be done for the function $f()$.

The CMMI defines that company reach 3d “Repeatable” level of maturity if it already reach 2d “Defined” level and mostly developed process areas that specific to the 3d level. In this case, if values of functions $f_1()$ and $f_2()$ defined by the term-set {NM, PM, FM} = {non mature, partially mature, fully mature} then following fuzzy rules could be defined:

$$\text{IF } (f_1(P_1,\ldots,P_7) = \text{NM}) \text{ THEN } M = I.$$

$$\text{IF } (f_1(P_1,\ldots,P_7) = PM) \text{ THEN } M = D.$$

$$\text{IF } (f_1(P_1,\ldots,P_7) = FM) \text{ THEN } M = D.$$

$$\text{IF } (f_1(P_1,\ldots,P_7) = PM) \text{ AND } (f_2(P_8,\ldots,P_{18}) = NM) \text{ THEN } M = D.$$

$$\text{IF } (f_1(P_1,\ldots,P_7) = PM) \text{ AND } (f_2(P_8,\ldots,P_{18}) = PM) \text{ THEN } M = R.$$

$$\text{IF } (f_1(P_1,\ldots,P_7) = FM) \text{ AND } (f_2(P_8,\ldots,P_{18}) = PM) \text{ THEN } M = R.$$

$$\text{IF } (f_1(P_1,\ldots,P_7) = FM) \text{ AND } (f_2(P_8,\ldots,P_{18}) = FM) \text{ THEN } M = R.$$

IT-startup maturity calculator

The mentioned above theoretical consideration became a basis in creation of the tool to assess IT-startup maturity. This tool, named “IT-startup maturity calculator” was designed to identify “bottlenecks”

in IT-startup organizational processes. It also helps to provide recommendations in fixing problems and improving efficiency of business process. The formal results of IT-startup maturity assessments could be used by University Business Center to support decision making regarding IT-startup.

There are several structured modules of IT-startup maturity calculator:

- "Questionnaire", that forms database of expert assessments of IT-startup performance indicators;
- IT-startup "maturity calculation" module, that assess the level of maturity of the company performing its fuzzy calculations;
- "Bottlenecks" identification module, which is based on a comparison of the actual assessments of process areas development to the proper level of maturity;

Module of retrospective analysis of performance indicators and it changes, which is based on a comparison of previously collected data.

The identification of "bottlenecks" is not a problem, because in fact, it is ether the lack of or weak development of the company's key practices in CMMI. Another thing is that the reasons for the existence of such "bottlenecks" can be different. The first is the lack of necessary skills and experience on the part of employees. Second is the lack of resources because special practices will require allocation of separate roles for their support.

A retrospective analysis of changes of maturity indicators of company plays a role in the assessing of the IT-startup's dynamics. It can serve as an indirect indicator of changes in the values of the team, the objectives of the project and other activities, the difficulties faced in front of the team. In addition, a stable negative dynamics of the maturity of IT-startup could be a criterion for withdrawal of support provided by the University Business Centre.

Created tool was tested to evaluate the maturity of the IT start-ups created by students of Slavutych branch of National Technical University of Ukraine "KPI" named Igor Sikorsky, as well as for the existing small IT companies that operates in Slavutych region.

It should be noted that both standard and proposed fuzzy method gives same results in assessment of the level of IT-startup maturity. However, the lack of possibility the result de-fuzzing of the standard method assessment does not allow to monitor the dynamics of changes in the company's maturity. It makes the proposed fuzzy method in the evaluation of maturity more attractive relative to the standard.

Conclusion

Finally, we can draw the following conclusions:

University-Business Cooperation and especially the academic entrepreneurship are promised approach to solve the problem of universities graduates not readiness to the requirements of industry and work market needs.

The process of growing IT companies is multifaceted and difficult, especially if this company consists of young students united only by common idea. University Business-Centers have to make difficult decisions to provide or cancel support the company. The proposed model of IT-startup in University ecosystem provides a control of IT-startup on its way to independent business.

One of the tools of such control is IT-startup maturity calculator, which could be successfully implemented using apparatus of fuzzy logic.

Bibliography

[European Commission, 2013] Digital Agenda Scoreboard 2013. European Commission, Brussels, 2013

[Kharchenko, 2012] Kharchenko V., Sklyar V. The concept and model of interaction between university research and the IT industry: S2B–B2S. In: KARDBLANSH, N.8-9, P.45-52, 2012. (in Russian).

[Kondratenko, 2015] Kondratenko Y. at al. Models of universities and IT-companies cooperation , decision-making system on fuzzy logic: monograph. Ed. Y.Kondratenko, Kharkov, 2015. (in Ukrainian).

[Lytvynov, 2015] Lytvynov V. at al. Tool-Based Support of University-Industry Cooperation in IT-Engineering. Lytvynov V., Kharchenko S., Lytvyn S., Saveliev V., Trunova E. and Skiter I. National University of Technology, Chernihiv, 2015.

[Bouwman, 2007] Bouwman H. Business models dynamics for start–ups and innovating e-businesses. M. de Reuver, H. Bouwman, I. MacInnes. In: Int. J. Electronic Business Vol. 7, No. 3. pp 269–286, 2007.

[Humphrey, 1989] Humphrey W. Managing the Software Process. Watts Humphrey, Addison-Wesley Professional, 1989

[Kondratenko, 2011] Kondratenko Y. Features of synthesis and modeling of hierarchically-organized DSS based on fuzzy logic. Kondratenko Y., Sidenko E. In: Vestnyk Khersonskogo natsionalnogo technicheskogo universiteta, N.2(41), pp. 150–158, 2011 (in Ukrainian).

[Grzegorzewski, 1998] Grzegorzewski P. Metrics and orders in space of fuzzy numbers. P. Grzegorzewski In: Fuzzy Sets and Systems, Vol. 97, Issue 1, pp. 83–94, 1998.

Authors' Information

Maxim Saveliev – Institute of Mathematical Machines and Systems Problems, Ukraine;

e-mail: saveliev.maxim@gmail.com

Major Fields of Scientific Research: Software Engineering, Automated System Life Circle Models, Requirements Evolution, System Analysis, System-of-Systems

Vitalii Lytvynov – Dr. Sc., Prof. Chernihiv, National University of Technology, 95, Shevchenko street, Chernihiv-27, Ukraine, 14027; vlitvin@ukrsoft.ua

Major Fields of Scientific Research: modeling of complicated systems, computer aided management systems, decision support systems

PROTECTION OF COMPUTER INFORMATION SYSTEMS OF AGRICULTURAL ENTERPRISES

Valentyn Nekhai, Igor Skiter, Elena Trunova

***Abstract**: The article deals with some modern methods and technologies used in solving problems of information support of the effective management of the agricultural enterprise. It contains the principles of construction systems of information protection in computer information systems of agricultural enterprises.*

***Key words:** information technology, information support, information systems, protection of information, policy of information safety, information protection system.*

***ACM Classification Keywords**: K.6.5 Management of computing and information systems - Security and Protection.*

Introduction

Activities of agricultural enterprises are characterized by complexity and system city of tasks that are to be solved: increase of arable land fertility; prevention of land degradation; improvement of yields and quality of agricultural products; minimization of costs for agro-technical measures; intra-logistics optimization and downtime reduction; minimization of economic risks of the enterprise's activity.

Everything mentioned above requires the transition to new methods of agriculture management information support, the usage of automatized control systems and modern information technologies. In turn, the rapid development of information technologies takes the form of global information revolution, which encourages the formation and development of innovative global substances - information environment and information society [Buriachok, 2013].

Over recent years, the understanding of information support impact on the process of making management decisions, takes the information to the next level – as a resource that has a certain value. Information becomes the most important strategic resource of any enterprise; its development and consumption become an important basis for the effective operation and development of various spheres of social and economic activity. The efficient activity of agricultural enterprises requires the information

that includes the complex of many factors data: grown crops peculiarities, climatic conditions, soil condition and quality, usage of fertilizers, pesticides etc.

The rapid technological development of information society, modern communicative capabilities and rapidly growing information space significantly increase the number of information sources, and thus extend the actual and/or potential sources of internal and external cyber influence, that makes the management of enterprises pay more attention to the protection of computer information systems.

The problems of agricultural enterprises' information management were described in the researches made by: I.V. Bal'chenko [Bal'chenko, 2013], V.V. Litvinov [V.V. Litvinov, 2013], V.P. Klimenko, Sayko V.F [Sayko V.F, 2006] and others. The significant part of work made by Buriachok V.L. [V.L. Buriachok, 2013], N.A. Gaydamakin [N.A. Gaydamakin, 2008] was dedicated to the problem of information security systems development and operation. But the rapid development of information technologies and the specificity of agricultural enterprises' activities require the search of new approaches to the information security organization.

The goal of the study is the research and analysis of existing information systems and their compliance with the current information needs in agricultural enterprises management, identification of vulnerabilities in terms of information security and information security system building.

Data protection

The development and implementation of automatized control systems show that none of the security information tools (methods, activities and assets) is completely reliable. Methodological and methodical bases of information security are quite general recommendations based on the international experience and the theory of systems.

Data protection is a set of methods and means that ensure the integrity, confidentiality and availability of information in terms of the impact of threats of natural or artificial nature, the implementation of which may result in damage to the owners and users of information.

Today's task of information security system is to adapt the abstract statements to the specific subject area (agricultural enterprises), where unique peculiarities and subtleties will be always present.

The research and analysis of foreign and local experience demonstrate the necessity for building an integrated system of enterprise information security, that includes operational, operational-technical and organizational measures for information protection. This system should provide flexibility and adaptation to rapidly changing factors of internal and external environment. It is impossible to provide this level of

information security without making an analysis of existing threats and potential possibilities for information leakage.

The basis for information security system creation is the development of information security policy for the enterprise. As a result, the protection plan should be created, which will implement the principles that are set out in the Security Policy.

Today, the problem of agricultural enterprises information protection is associated with the creation of large agricultural holdings and the transition to high-intensive farming. The basis for the transition to the innovative farming is the availability of information concerning the exact limits of arable land and their agro-chemical and agro- physical characteristics. This in turn requires the usage of modern information technologies and revision of approaches for agricultural enterprises information system creation.

Enterprise Performance Management

Enterprise Performance Management can be the basis for automatized control system building not only as a management concept, but also as the exact class of information systems that support this concept.

The enterprise information infrastructure can be presented in several hierarchical levels, each of which is characterized by the degree of information aggregation and its role in the management process. “Analytical stack” developed by Gartner can be an example of schematic representation of the information infrastructure. There are several levels in this hierarchy [Ysaev, 2008]:

- the level of transactional systems;
- the level of business intelligence, including data warehouses, data marts and OLAP-systems;
- the level of analytical applications (Picture 1.).

Transactional systems include enterprise resource management systems (ERP-system) and provide the information needs of management at the operational level. Despite the objective differences, all these systems have a common feature: they are designed to handle certain operations (On-Line Transaction Processing (OLTP) - processing transactions in real time). The goals, objectives and sources of information at the operational level are initially defined and have a high degree of structure and formalization.

Transactional systems are the sources of primary information, which after the appropriate processing are used for further analytical processing and presentation for making management decisions. From transactional systems, data can be passed to analytical applications either sequentially through all the levels of analytical stack or by passing one or more levels (“bypass” - “direct transfer”).

Data warehouse (DW) is defined by Bill Inmon [Inmon, 1992] as "subject-oriented, integrated, stable, supporting the chronology of data sets, organized for the purpose of management support, designed to

act as "one and the only one source of truth" that provides managers and analysts with reliable information necessary for rapid analysis and making decisions".

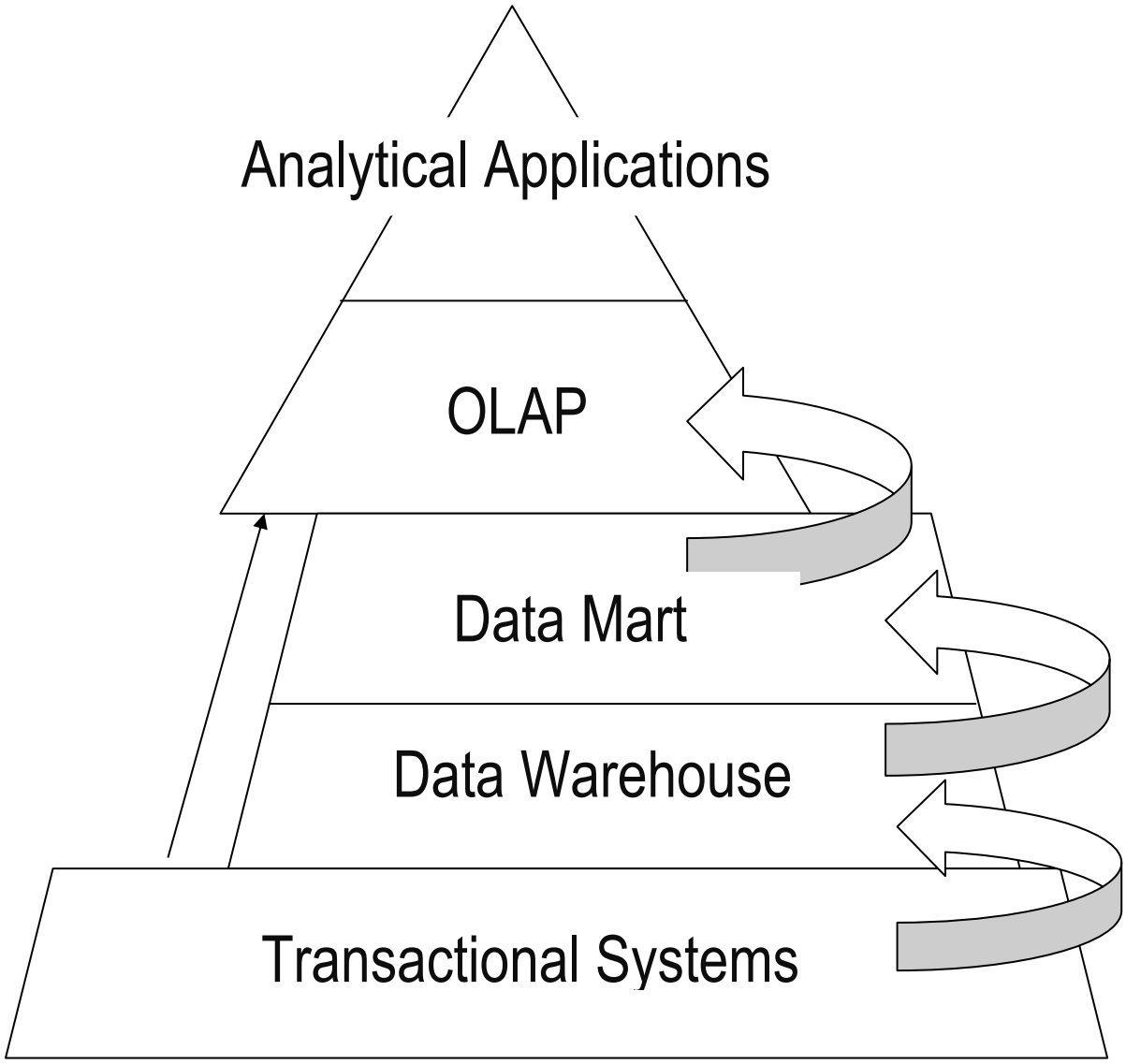


Figure 1. Analytical Stack

However, the large amount of data contained in warehouses, usually make them unavailable for processing in real time. This problem is solved on the following hierarchy levels – data marts and OLAP-systems.

Data marts are structured information files, but their difference is that they are subject-oriented, the information is stored in data marts in the most favorable form for solving specific analytical problems.

The next level of the analytical stack is occupied by On-Line Analytical Processing (OLAP-system). This is the system of analytical data processing in real time that can provide the solutions of many analytical problems and work with relevant data despite of the company's activities characteristics.

OLAP-systems are characterized by large dimensions of stored data (as opposed to relational tables), preliminary calculation and aggregation of values, which makes it possible to build quick independent requests to operational database using a number of different analytical measures.

At the highest level of the analytical stack there are analytic applications, aimed at the analysis and decision support at the strategic level. The information system on the strategic level (Executive Support Systems, ESS) provides the support of making decisions concerning the implementation of promising strategic aims of enterprise development on the basis of solving unstructured problems, special problems that require professional judgments, estimates and intuition.

Peculiarities of construction of the automated control system of the agricultural enterprise

To ensure the information needs of agriculture enterprises management various information systems are used nowadays:

- monitoring system of agricultural resources conditions and crop yields forecasting;
- supporting system of agricultural products quality control;
- operational control system and productive processes optimization;
- information and reference systems of marketing orientation;
- analytical and modeling systems of tracking the emergencies and their impact on production, agricultural products quality, and many other specialized information systems of different orientation and level of detail [Sayko, 2006].

Based on the agricultural enterprises management needs, the following main aspects of creation of agricultural information systems that allow justifying their structure and functions, can me defined [Savchenko, 2006]:

- the necessity for creation of new management and agriculture systems, that take into account natural conditions and organizational and technological capabilities of the company, maximal use of its soil and climatic potential;
- the inextricable connection between technology and biological objects (soil, plants, etc.), which are characterized by occurring continuous processes and cyclical products;

- the need for continuous monitoring of a large number of parameters, including geographically distributed;
- variety of processes and operations in the processing plants, which are usually set out in huge technological maps;
- significant differentiation of agricultural manufacturers in terms of amount and production structure, sustainability, etc.;
- agronomic data is characterized by significant volume of different data that is difficult formalized.

In works [Bal'chenko, 2013], [Litvinov, 2013] the analysis of modern methods and technologies that are used in the process of solving the issues of effective management of agricultural enterprises is made. The approach for building the automated management system of agricultural enterprises is given. The major functional subsystems of information managing system of agriculture enterprise are specified:

1. Normative reference and infrastructure subsystem (system administrators) - to conduct the regulatory information that is required for use in solving management problems.
2. Subsystem of collecting primary information concerning the management object (system administrator, manager) - to collect primary information on the status and processes of the enterprise and the transfer of urgent messages and instructions from the control center to the performers.
3. Subsystem of crop/livestock work planning of and their resources' provision.
4. Subsystem of operational dispatch management of crop/livestock works and operations of and corresponding resources' provision (managers, agronomists/livestock specialists) - automatizing the distribution and initial data processing process concerning the state of management facilities, supporting in the process of making decisions.
5. Subsystem of facilities management assessment - for use in solving dispatcher problems.
6. Logistics subsystem.
7. Subsystem of keeping mapping information (cartographers, land surveyors) - presenting information concerning the state of management facilities in form of digital maps.
8. Subsystem of notification and exchange of urgent messages and instructions between the control center and the performers - designed for messaging between the employees of distributed system.
9. Subsystem of modeling the enterprise activity - for imitating modeling of possible consequences of the prevailing situation in the enterprise activity.

The fundamental concept in the sphere of computer systems information security is the security policy. It means an integrated set of rules and regulations that govern the information processing, the implementation of which provides the status of information security in the given space of threats. The formal expression of Security Policy (mathematical, schematic, algorithms, etc.) is called the security model.

Security models play an important role in the process of development and research of protected computer systems as they provide the system engineering approach that involves the solving of such critical tasks:

- selection and justification of the basic architecture principles of secure computer systems, that defines the mechanisms of protection means and methods implementation;
- verification of systems' properties (security) is developed by formal confirmation of compliance with security policies (requirements, conditions, criteria);
- making a formal specification of security policy as an essential part of organizing and documenting software protection, developed computer systems [Gaydamakin, 2008].

Security policy of computer informational systems

There are the following types of information computer systems security policy [Devyanin, 2005].

Discretionary security policy is the security policy, based on the Discretionary Access Control, which is defined by two properties:

- all subjects and objects are identified;
- the rights of access to system objects and subjects are based on some external rules in relation to the system.

The main element of discretionary access control systems is the matrix of access - the matrix of size $|S| \times |O|$, the lines of which correspond to subjects and the columns correspond to objects. In such a case every element of the access matrix $M[S, O]$ with R determines the access rights of the subject S to the object O, where R is the set of permissions.

The advantages of discretionary security policy include the relatively simple implementation of access control systems; the disadvantages include the static of defined rules of access therein.

Mandate (authority) security policy is a security policy based on Mandatory Access Control, which is defined by four conditions:

- unambiguous identification of all subjects and objects of the system;
- given hierarchical levels of information confidentiality;
- every system object has the level of confidentiality that determines the value of information;
- every system subject has the access level.

Mandate security policy application helps to prevent the overflow of information from the objects with higher hierarchy level to the objects with low access level; on the other hand, the introduction of systems based on the security policy of this type is complicated and requires significant hardware and software resources of information system.

The approach of information flow security policy should be mentioned. It is based on the sharing of all possible information flows between the objects of the system into two disjoint sets: the set of enabling information flows and the set of adverse information flows, the purpose of implementation of which is to ensure the unavailability of emergences in the computer system information flows.

Role differentiation of access is the development of discretionary differentiation access policy, and the rights of access to system objects are based on their application-specific basis, defining their roles thereby. Role differentiation of access allows realizing flexible access control rules that take into account the dynamics of the computer system operation process.

In addition to the abovementioned policy we can name the policy of isolated software environment implemented by determining the order of safe interaction of system subjects that ensures the impossibility of influence on information and security systems and their settings modification or configuration.

Thus, the development of information system security policy should include three levels: basic, segment and marginal. The security policy of base and segment levels must ensure the protection of information flow within the information system, the marginal level of security provides the protection of information exchange with the environment (figure 2).

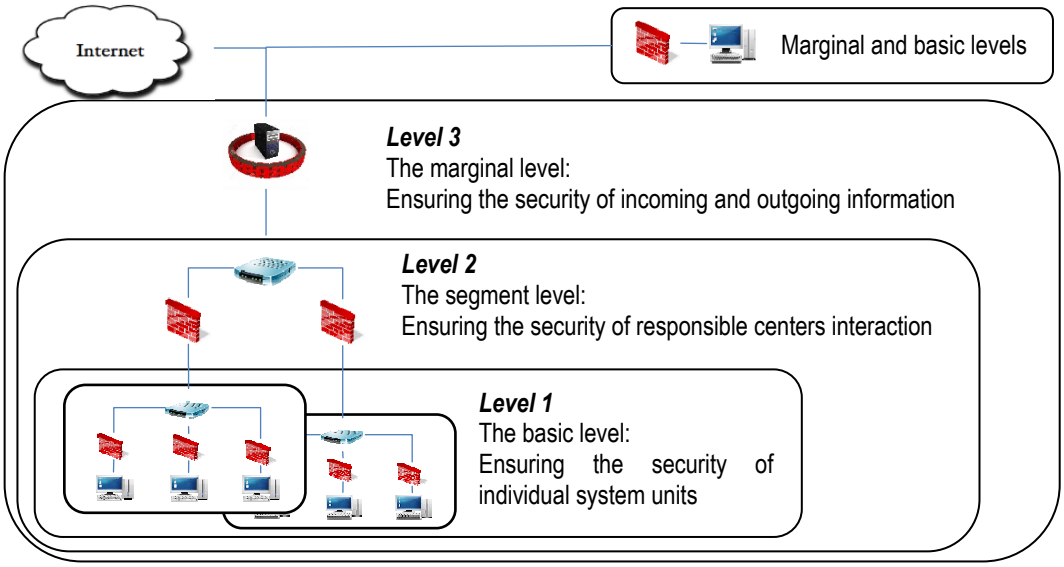


Figure 2. The hierarchical model of information security policy

Conclusions

The basis for constructing a system of information systems protection is the development of the security policy that is based on: organizational and management structure of the company; informational management needs of the enterprise; used organizational, technical and software; processing technology.

The security policy development should be based on a strict hierarchy; this means that the protection degree of different system units cannot be the same. Thus, the data that is being processed in these sites will be under the thread of unauthorized exposure risks. Having divided the information in several categories according to its importance (critical and non-critical), the model of any company's protection can be optimized.

Bibliography

[Buriachok, 2013]. V.L. Buriachok (2013). Bases of state system cyber security formation: Monograph. – Kyiv, 431 p. (in Ukrainian)

[Ysaev, 2008] D.V. Ysaev (2008) Analytical information systems. – Moscow, 60 p. (in Russian)

[Inmon, 1992] William H. Inmon (1992) Building the Data Warehouse, New York: John Wiley & Sons, 1992.

[Sayko, 2006] V.F. Sayko (2006) Scientific support of agriculture and agricultural technologies. Bulletin of Agricultural Science. № 12. - P. 15-19. (in Ukrainian)

[Savchenko, 2006] O.F. Savchenko (2006) Methodological aspects of information systems in agriculture. Advances in science and agribusiness technology. № 11. - P. 5-9 (in Russian)

[Bal'chenko, 2013] I.V. Bal'chenko (2013) Peculiarities of Information Technology Management for Agricultural enterprise. Ed. I.V. Bal'chenko, V.V. Litvinov and V.P. Klimenko. Bulletin of Chernihiv State Technological University. № 67. - P. 211- 218. (in Ukrainian)

[Litvinov, 2013] V.V. Litvinov (2013) Peculiarities of construction of the automated control system of the agricultural enterprise. Ed. I.V. Bal'chenko, V.V. Litvinov and V.P. Klimenko. Mathematical Machines and Systems. № 4. - P. 82-94. (in Ukrainian)

[Gaydamakin, 2008] N.A. Gaydamakin (2008) Theoretical Foundations of Computer Security. - Ekaterinburg, 212 p. (in Russian)

[Devyanin, 2005] P.N. Devyanin (2005) Models of computer systems. - Moscow: Publishing Center "Academy", 144 p. (in Russsian)

Authors' Information

Valentin Nekhai *– Ph.D. Student, Chernihiv National University of Technology, 95, Shevchenko street, Chernihiv-27, Ukraine, 14027; valentin_nehai@meta.ua*

Major Fields of Scientific Research: protection of computer information systems of agricultural enterprises

Igor Skiter *– PhD, Associate Professor of Software Engineering Chernigov National Technological University, 14000, Shevchenko st. 95, Chernigov, Ukraine; e-mail: skiteris@mail.ru*

Major Fields of Scientific Research: The main direction of research: mathematical modeling of systems, decision support systems

Elena Trunova – *PhD, Associate Professor of Software Engineering Chernigov National Technological University, 14000, Shevchenko st. 95, Chernigov, Ukraine; e-mail: e.trunova@gmail.com*

Major Fields of Scientific Research: mathematical modeling of systems, decision support systems, theory and methods of teaching in higher education

PARTIAL DEDUCTION IN PREDICATE CALCULUS AS A TOOL FOR ARTIFICIAL INTELLIGENCE PROBLEM COMPLEXITY DECREASING

Tatiana M. Kosovskaya

Abstract: *Many artificial intelligence problems are NP-complete ones. To decrease the needed time of such a problem solving a method of extraction of sub-formulas characterizing the common features of objects under consideration is suggested. This method is based on the offered by the author notion of partial deduction. Repeated application of this procedure allows to form a level description of an object and of classes of objects. A model example of such a level description and the degree of steps number increasing is presented in the paper.*

Keywords: *Artificial Intelligence, pattern recognition, predicate calculus, level description of a class..*

1. Introduction

While simulation an Artificial Intelligence (AI) problem the most of researchers consider an investigated object as a unit which is characterized by some global features [12]. In particular, a researcher using methods of mathematical logic operates with propositional formulas or Boolean functions [2]. Such an approach is not convenient for a simulation of a complex object which is described by properties of its elements and relations between them.

At the same time there are many papers which offer to use predicate calculus and resolution method for the above mentioned problems [13; 14]. The predicate calculus language is enough adequate to simulate complex or changeable objects. But, unfortunately, the authors do not take into account the time complexity of a problem using such a simulation.

The point is that a problem using such a simulation is an NP-hard [6]. If **P**$\neq$**NP** then such a problem may be solved only in the time exponentially depended of the input [5; 3].

The upper bounds of number of steps for algorithms solving some AI problems described by the predicate calculus language were proved by the author in [6; 10; 7]. The analysis of thees upper bounds allowed to develop hierarchical many-level descriptions of the goal conditions which essentially decrease the solving time for the mentioned problems [8]. But at that time there was no tool for automatic construction of a level description. Intuitive construction of such a description showed that the time decreases.

The notion of partial deduction [9] earlier introduced by the author for the recognition of an object with incomplete information occurred to be such a suitable tool.

Some AI problems using predicate language description which may be simplified with the use of partial deduction are presented in this paper.

2. Logic-objective approach to a recognition problem

Let an investigated object ω is represented as a set of its elements $\omega = \{\omega_1, ..., \omega_t\}$ and predicates $p_1, ..., p_n$ define properties of these elements and relations between them.

Logical description $S(\omega)$ of an object ω is a collection of all true formulas in the form $p_i(\overline{\tau})$ or $\neg p_i(\overline{\tau})$ (where $\overline{\tau}$ is an ordered subset of ω) describing properties of ω elements and relations between them.

Let the set of all investigated objects Ω is a union of classes $\Omega = \cup_{k=1}^{K} \Omega_k$.

Logical description of the class Ω_k is such a formula $A_k(\overline{x})$ that if the formula $A_k(\overline{\omega})$ is true then $\omega \in \Omega_k$. The class description may be represented as a disjunction of elementary conjunctions of atomic formulas.

Here and below the notation $\overline{x}$ is used for an ordered list of the set x. To denote that there exist distinct values for variables from the list $\overline{x}$ the notation $\exists \overline{x}_{\neq} A_k(\overline{x})$ is used.

The introduced descriptions allow to solve many artificial intelligence problems which may be formulated as follows.

Identification problem. To pick out all parts of the object ω which belong to the class Ω_k.

Classification problem. To find all such class numbers k that $\omega \in \Omega_k$.

Analysis problem. To find and classify all parts τ of the object ω.

These problems are reduced in [1] to the following formulas respectively

$$S(\omega) \Rightarrow (?\overline{x}_k) A_k(\overline{x}_k), \tag{1}$$

$$S(\omega) \Rightarrow (?k) A_k(\overline{x}_k), \tag{2}$$

$$S(\omega) \Rightarrow (?k)(?\overline{x}_k) A_k(\overline{x}_k), \tag{3}$$

where $(?k)$ and $(?\overline{x})$ denote the words "what are the values of k?" and "what are the values of $\overline{x}$?".

It is proved in [6] that the corresponding recognition problems

$$S(\omega) \Rightarrow \exists \overline{x}_{k\neq} A_k(\overline{x}_k), \tag{4}$$

$$S(\omega) \Rightarrow \vee_{k=1}^{K} A_k(\overline{x}_k), \tag{5}$$

$$S(\omega) \Rightarrow \vee_{k=1}^{K} \exists \overline{x}_{k\neq} A_k(\overline{x}_k) \tag{6}$$

are NP-complete. Hence the problems (1), (2), (3) are NP-hard.

3. Methods of proof and upper bounds of their number of steps

If one can solve the problem (1) with $A_k(\overline{x})_k$ be a conjunction of atomic formulas then he can solve the problems (1) with arbitrary $A_k(\overline{x}_k)$, (2) and (3), and the number of steps of their solutions would differ from the first one polynomially. If we solve problems (4), (5), (6) by means of a "constructive" algorithm (i.e. algorithm not only proves the existence but also finds values for variables $\overline{x}$ and parameter k) then we simultaneously solve problems (1), (2), (3). That is why the complexity bounds of algorithms will be done for the problem (4) in the form

$$S(\omega) \Rightarrow \exists \overline{x}_{\neq} A(\overline{x}), \tag{4'}$$

where $A(\overline{x})$ is a conjunction of atomic formulas.

The **exhaustive search method** is one which allows to finds values for variables $\overline{x}$. It is proved in [6] that its number of steps is

$$O(t^m), \tag{7}$$

where t is the number of the elements in ω, m is the number of variables in the formula $A(\overline{x})$. Note that this estimate coincides with the one for simulation of predicate approach to the artificial intelligence problems by boolean formulas [14].

Logical methods (namely logical derivation in a sequent calculus or by resolution method) also allow to finds values for variables $\overline{x}$. Both these methods has the number of steps

$$O(s^a), \tag{8}$$

where s is the maximal number of occurrences of the same predicate in the description $S(\omega)$ and a is the number of atomic formulas in the formula $A(\overline{x})$ [10].

4. Level approach to the decision of problems

To decrease the obtained step number estimates a level description of goal formulas was offered in [8; 11]. Let $A_1(\overline{x}_1)$, ..., $A_K(\overline{x}_K)$ be a set of goal conditions every of which is a conjunction of atomic formulas. Find all subformulas $P_i^1(\overline{y}_i^1)$ with a "small" complexity which "frequently" appear in goal formulas $A_1(\overline{x}_1)$, ..., $A_K(\overline{x}_K)$ and denote them by atomic formulas with new predicates p_i^1 and new first-level arguments z_i^1 for lists $\overline{y}_i^1$ of initial variables. Write down a system of equivalences

$$p_i^1(z_i^1) \Leftrightarrow P_i^1(\overline{y}_i^1), \quad i = 1, \ldots, n_1.$$

What object must be called a "common sub-formula" of two formulas A and B?

For example, let

$$A(x, y, z) = p_1(x)\&p_1(y)\&p_1(z)\&p_2(x, y)\&p_3(x, z),$$

$$B(x, y, z) = p_1(x)\&p_1(y)\&p_1(z)\&p_2(x, z)\&p_3(x, z).$$

If the formula

$$P(u, v) = p_1(u)\&p_1(v)\&p_2(u, v)$$

is their common sub-formula?

The formula $P(u, v)$ is their common up to the names of variables sub-formula with the substitutions $\lambda_{P,A} = |_{x\ y}^{u\ v}$ and $\lambda_{P,B} = |_{x\ z}^{u\ v}$ because

— $P(x, y) = p_1(x)\&p_1(y)\&p_2(x, y)$ is a sub-formula of $A(x, y, z) = p_1(x)\&p_1(y)\&p_1(z)\&p_2(x, y)\&p_3(x, z)$,

— $P(x, z) = p_1(x)\&p_1(z)\&p_2(x, z)$ is a sub-formula of $B(x, y, z) = p_1(x)\&p_1(y)\&p_1(z)\&p_2(x, z)\&p_3(x, z)$.

Definition. *The formula P is called a common up to the names of variables sub-formula of formulas A and B if there are such substitutions $\lambda_{P,A} = |_{\overline{t}_A}^{\overline{x}}$ and $\lambda_{P,B} = |_{\overline{t}_B}^{\overline{x}}$ of the lists of terms $\overline{t}_A$ and $\overline{t}_B$ instead of the list of variables $\overline{x}$ that the formula P turns into a sub-formula of A and B respectively.*

Such substitutions are called unifiers of P with A and B respectively.

Let $A_k^1(\overline{x}_k^1)$ be a formula received from $A_k(\overline{x}_k)$ by substitution of $p_i^1(z_i^1)$ instead of $P_i^1(\overline{y}_i^1)$. Here $\overline{x}_k^1$ is a list of all variables in $A_k(\overline{x}_k^1)$ including both some (may be all) initial variables of $A_k(\overline{x}_k)$ and first-level variables appeared in the formula $A_k^1(\overline{x}_k^1)$.

A set of all atomic formulas of the type $p_i^1(\omega_i^1)$ where ω_i^1 denotes some ordered list $\overline{\tau}_i^1$ of elements from ω for which the formula $P_i^1(\overline{\tau}_i^1)$ is valid is called a first-level object description and denoted by $S^1(\omega)$. Such a way extracted subsets $\overline{\tau}_i^1$ are called first-level objects.

Repeat the above described procedure with formulas $A_k^1(\overline{x}_k^1)$. After L repetitions L-level goal conditions in the following form will be received.

$$\left\{\begin{array}{ccc} & A_k^L(\overline{x}_k^L) & \\ p_1^1(z_1^1) & \Leftrightarrow & P_1^1(\overline{y}_1^1) \\ & \vdots & \\ p_{n_1}^1(z_{n_1}^1) & \Leftrightarrow & P_{n_1}^1(\overline{y}_{n_1}^1) \\ & \vdots & \\ p_i^l(z_i^l) & \Leftrightarrow & P_i^l(\overline{y}_i^l) \\ & \vdots & \\ p_{n_L}^L(z_{n_L}^L) & \Leftrightarrow & P_{n_L}^L(\overline{y}_{n_L}^L) \end{array}\right. .$$

Such L-level goal conditions may be used for efficiency of an algorithm solving a problem formalized in the form of logical sequent (3). To decrease the number of steps of an exhaustive algorithm (for every t greater than some t_0)

with the use of 2-level goal description it is sufficient

$$n_1 \cdot t^r + t^{s_1+n_1} < t^m, \tag{7}$$

where r is a maximal number of arguments in the formulas $P_i^1(\overline{y}_i^1)$, n_1 is the number of first-level predicates, s_1 is the number of atomic formulas in the first-level description, m is the number of variables in the initial goal condition.

Similar condition for decreasing the number of steps of a logical algorithm solving the problem (3) is

$$\sum_{k=1}^{K} s^{1a_k^1} + \sum_{j=1}^{n_1} s^{\rho_j^1} < \sum_{k=1}^{K} s^{a_k}, \tag{8}$$

where a_k and a_k^1 are maximal numbers of atomic formulas in $A_k(\overline{x}_k)$ and $A_k^1(\overline{x}_k^1)$ respectively, s and s^1 are numbers of atomic formulas in $S(\omega)$ and $S^1(\omega)$ respectively, ρ_j^1 is the number of atomic formulas in $P_i^1(\overline{y}_i^1)$ [8].

5. Partial deduction

During the process of partial deduction instead of the proof of $A(\overline{x}) \Rightarrow \exists \overline{y}_{\neq} B(\overline{y})$ we search such a maximal (up to the names of variables) sub-formula $B'(\overline{y}')$ of the formula $B(\overline{y})$ that $A(\overline{x}) \Rightarrow \exists \overline{y}'_{\neq} B'(\overline{y}')$.

Let a and a' be the numbers of atomic formulas in $A(\overline{x})$ and $A'(\overline{x}')$ respectively, m and m' be the numbers of objective variables in $A(\overline{x})$ and $A'(\overline{x}')$ respectively. Parameters q and r are defined as $q = a'/a$ and $r = m'/m$. In such a case sub-formula $A'(\overline{x}')$ is called a (q, r)-fragment of the formula $A(\overline{x})$.

Definition. *The problem of partial deducibility of a formula $B(\overline{y})$ from $A(\overline{x})$*

$$A(\overline{x}) \Rightarrow_P \exists \overline{y}_{\neq} B(\overline{y})$$

is the problem of extraction of such a maximal (upon q) (q, r)-fragment $Q(\overline{u})$ of the formula $B(\overline{y})$ that

$$A(\overline{x}) \Rightarrow \exists \overline{u}_{\neq} Q(\overline{u}).$$

It may be proved that if $Q(\overline{u})$ and $R(\overline{v})$ are two maximal sub-formulas of $A(\overline{x})$ and $B(\overline{y})$ obtained while checking

$$A(\overline{x}) \Rightarrow_P \exists \overline{y}_{\neq} B(\overline{y})$$

and

$$B(\overline{y}) \Rightarrow_P \exists \overline{x}_{\neq} A(\overline{x})$$

then $Q(\overline{u})$ and $R(\overline{v})$ coincide up to the names of variables.

That is there exists their common unifier $\lambda = |_{\overline{z}}^{\overline{u}} {}_{\overline{z}'}^{\overline{v}}$.

6. Algorithm of level description construction

The below described algorithm was offered in [11].

Let $A_1(\overline{x}_1), ..., A_K(\overline{x}_K)$ be elementary conjunctions which are components of class descriptions.

1. For every pair $A_i(\overline{x}_i)$ and $A_j(\overline{x}_j)$ $(i \neq j)$ extract their maximal common up to the names of variables sub-formula $Q_{i,j}^1(\overline{x}_{i,j}^1)$ and find unifiers $\lambda_{(i,j),i}$ and $\lambda_{(i,j),j}$.
2. Repeat the extraction of maximal common up to the names of variables sub-formula for every pair of already extracted sub-formulas $Q^{l-1}_{i_1 \ldots i_{2^{l-1}}}(\overline{x}_{i_1 \ldots i_{2^{l-1}}})$ ï£¡ $Q^{l-1}_{j_1 \ldots j_{2^{l-1}}}(\overline{x}_{j_1 \ldots j_{2^{l-1}}})$ and obtain their common sub-formulas $Q^{l}_{i_1 \ldots i_{2^{l-1}}, j_1 \ldots j_{2^{l-1}}}(\overline{x}_{i_1 \ldots i_{2^{l-1}}, j_1 \ldots j_{2^{l-1}}})$ $(l = 2, , L)$ and the unifiers.

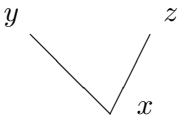
3. Select among the extracted sub-formulas $Q^l_{i_1\dots i_{2^{l-1}},j_1\dots j_{2^{l-1}}l}(\overline{x}_{i_1\dots i_{2^{l-1}},j_1\dots j_{2^{l-1}}})$ minimal ones and denote them by means of $P^1_i(\overline{y}^1_i)$ $(i = 1,,n_1)$. They are elementary conjunctions defining the first-level predicates $p^1_i(y^1_i)$ and the first-level variable y^1_i is the variable for the string of initial variables.
4. Sub-formulas of the higher levels $P^{l+1}_i(\overline{y}^{l+1}_i)$ $(i = 1,,n_{l+1},\ l = 2,,L)$ are constructed from the previously extracted sub-formulas $Q^l_{i_1\dots i_l,j_1\dots j_l}(\overline{x}_{i_1\dots i_l,j_1\dots j_l})$ with the substitution of $p^1_i(y^1_i)$ instead of $P^1_i(\overline{y}^1_i)$. Here y^1_i is the variable for the string of the less level variables.

The found unifiers are used here.

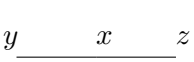
7. Example

The images for this example are taken from [4].

Let we must recognize contour images described by the following predicates.



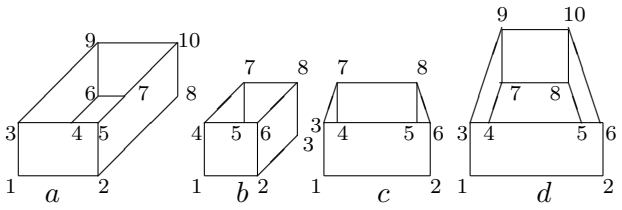
$V(x,y,z) \Longleftrightarrow (\angle yxz < \pi)$



$L(x,y,z,) \Longleftrightarrow x$ belongs to the segment $[y,z]$

Initial predicates.

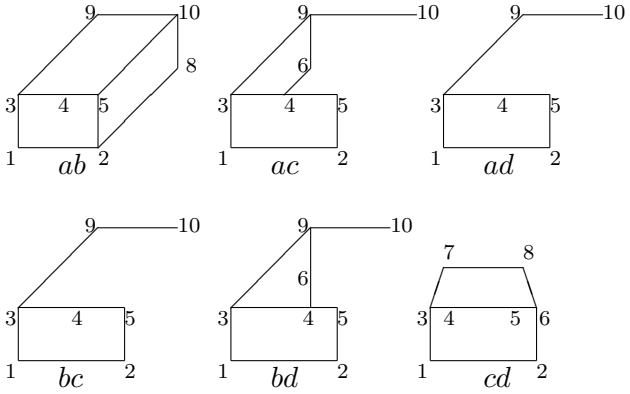
Given a set of contour images of "boxes" presented on the picture one can obtain the description of the class of "boxes" by means of changing the name of a node i by the variable x_i in the description of an object.



Training set

Given a complex image containing t nodes and not more than s occurrences of the same predicate in the image description, the number of steps needed for identification (and extraction) of a "box" is $O(t^{10})$ for an exhaustive algorithm and $O(s^{29})$ for a logical algorithm.

The first extraction of the common up to the names of variables subformulas gives 5 subformulas (subformulas corresponding to the images ad and bc coincide).



Images corresponding
to the extracted sub-formulas.

The second extraction of the common up to the names of variables subformulas gives 1 subformula. It defines the first-level sub-formula.

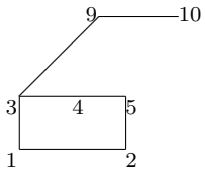


Image corresponding
to the first-level sub-formula.

This subformula contains 7 nodes and 10 relations between them. A new first-level variable x^1 for the string of variables $(x_1, x_2, x_3, x_4, x_5, x_9, x_{10})$ and a new first-level predicate p^1 such that $p^1(1, 2, 3, 4, 5, 9, 10)$ is true for the object a are introduced.

Images corresponding to the second-level sub-formulas are ab, ac, bd and cd. Their formulas have the first-level sub-formula which is changed by the first-level predicate. They have the variable x^1 and the initial variables

x_2, x_5, x_8, x_9, x_{10} for ab,

x_4, x_6, x_9, x_{10} for ac,

x_4, x_6,x_9, x_{10} for bd,

x_4, x_5, x_6, x_7, x_8 for cd.

At the same time the indicating the value of x^1 makes unknown only variables x_8, x_{10} for ab; x_6, x_{10} for ac; x_6, x_9, x_{10} for bd; x_4, x_8 or x_5, x_8 for cd.

Hence, these second-level sub-formulas contain respectively $m_{ab} = 3$, $m_{ac} = 3$, $m_{bd} = 3$ and $m_{cd} = 2$ essential variables (x^1 and some "old" ones).

Every of the second-level sub-formulas contain the first-level subformula $p^1(x^1)$ and some "old" atomic formulas. Their amounts are respectively $s_{ab} = 8$, $s_{ac} = 7$, $s_{bd} = 5$, $s_{cd} = 8$.

Elementary conjunctions corresponding the training set in the three-level descriptions contain one of the second-level subformulas $p^2_k(x^2_k)$ ($k = 1, ..., 6$) and some "old" atomic formulas. Every of these formulas contain respectively $m_a = 4$, $m_b = 3$, $m_c = 2$ and $m_d = 4$ essential variables (x^2_k and some "old" ones).

The amounts of atomic formulas (with a second-level predicate and initial ones) are respectively $s_a = 8$, $s_b = 9$, $s_c = 5$, $s_d = 9$.

So the number of an exhaustive algorithm steps for the tree-level description is $O((t^3 + t^3 + t^3 + t^2) + (t^4 + t^3 + t^2 + t^4)) = O(t^4)$ instead of $O(t^{10})$ for the initial description.

The number of a logical algorithm steps for the tree-level description is $O((s^8 + s^7 + s^5 + s^8) + (s^8 + s^9 + s^5 + s^9)) = O(s^9)$ instead of $O(s^{29})$ for the initial description.

8. Discussion

The open question is ï£¡what extracted formula must be changed by an atomic one if it may be done in different ways?ï£¡ In the example above the formula corresponding to the image d contains both the subformula corresponding to the image bd and the siubformula corresponding to the image cd. What second-level predicate must appear in the tree-level description of d? To answer this question complexity investigation must be done.

While extracting a sub-formula it may happen that it contains several variables of a lower (not initial) level. In such a case the sub-formula defines a relation between parts of an object. If we must regard these parts as informative pair or a new informative part?

9. Conclusion

The use of predicate calculus language seems to be an adequate one for the simulation of Artificial Intelligence problems. But the NP-completeness of the problems appeared while such a simulation does not allow to implement algorithms directly.

The notion of partial deduction for a predicate formula allows to construct such a level description of classes that the exponent in the complexity upper bound of the problem solution decreases very much.

It does not mean that we can solve an NP-hard problem in a polynomial time. Because the construction of a level description is also an NP-hard problem with the almost same exponent in the complexity upper bound. It corresponds to the long time of learning and the quick implementation of the received knowledge.

10. Acknowledgment

The paper is published with financial support of RFBR grant 14-08-01276.

Bibliography

[1] T.M. Artyushkova (Kosovskaya) and A.V. Timofeev, *On one new approach to the forming of logical decision rules.* In: Vestnik of St.Petersburg University. 1985, No 8. P. 22 ï£¡– 29. (In Russian)

[2] D. Bugaychenko and D. Zubarevich, *Fast Pattern Recognition and Deep Learning Using Multi-Rooted Binary Decision Diagrams*, Lecture Notes in Artificial Intelligence, V. 8556, 2014. P. 73 —77.

[3] D.Z. Du and K.I. Ko, *Theory of Computational Complexity,* A Wiley-Interscience Publication. John Wiley & Sons, Inc. 2000.

[4] R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*, A Wiley-Interscience Publication, John Wiley & Sons, New York London Sydney Toronto, 1973.

[5] M.R. Garey and D.S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness,* Freeman, New York (1979)

[6] T.M. Kossovskaya, *Proofs of the number of steps bounds for solving of some pattern recognition problems with logical description*, Vestnik Sankt-Peterburgskogo Universiteta. Seriya 1. 2007, No. 4, pp. 82–90. (In Russian)

[7] T. Kosovskaya, *Discrete Artificial Intelligence Problems and Number of Steps of their Solution*, International Journal on Information Theories and Applications, Vol. 18, Number 1, 2011. P. 93 — 99.

[8] T.M. Kossovskaya, *Level descriptions of classes for decreasing step number of pattern recognition problem solving described by predicate calculus formulas,* Vestnik SanktPeterburgskogo Universiteta. Seriya 10. 2008, No. 1, pp. 64 – 72. (In Russian)

[9] T.M. Kossovskaya. *Partial deduction of predicate formula as an instrument for recognition of an object with incomplete description,* Vestnik SanktPeterburgskogo Universiteta. Seriya 10. 2009, No. 1, pp. 74 – 84. (In Russian)

[10] *T.M. Kossovskaya*, Some artificial intelligence problems permitting formalization by means of predicate calculus language and upper bounds of their solution steps. //SPIIRAS Proceedings, 2010. No. 14, pp. 58 - 75. (In Russian)

[11] T.M. Kosovskaya. *Construction of Class Level Description for Efficient Recognition of a Complex Object,* International Journal ï£¡Information Content and Processingï£¡, Vol. 1,No 1. 2014. P. 92 ï£¡ 99.

[12] M.A. Lejeune, *Pattern definition of the p-efficiency concept*, ANNALS OF OPERATIONS RESEARCH V. 200, Issue 1, 2012. P. 23 — 36.

[13] N.J. Nilsson, *Problem-Solving Methods in Artificial Intelligence*, McGRAW-HILL BOOK COMPANY, NEW YORK, 1971.

[14] S.J. Russel and P. Norvig, *Artificial Intelligence. A Modern Approach.* Pearson Education, Inc. 2003.

Authors' Information

Tatiana M. Kosovskaya *prof. of St. Petersburg State University,*
University av., 28, St. Petersburg, 198504, RUSSIA
Senior researcher of St. Petersburg Institute on Informatics and Automation of Russian Academy of Science, 14 line, 39, St. Petersburg, 199178, RUSSIA
E-mail: kosovtm@gmail.com

Review of some problems on the complexity of simultaneous divisibility of linear polynomials

Nikolay K. Kosovskii, Mikhail Starchak

Abstract: *An introduction to the problems considering complexity of simultaneous divisibilities of values of linear polynomials is presented. Some history facts, recent results and open questions that stimulate further research are discussed.*

Keywords: *NP-completeness, existential Presburger arithmetic with divisibility, systems of divisibilities of values of linear polynomials, quadratic diophantine equations*

ACM Classification Keywords*: F.1.3 Complexity Measures and Classes, Reducibility and completeness; F.2.1 Analysis of Algorithms and Problem Complexity, Numerical Algorithms and Problems, Number-theoretic computations*

Introduction

Being defined, the notion of NP-completeness has become widely known as a synonym of practical intractability of a computational problem. There were found such problems in various fields of applied mathematics. In [Garey, Johnson, 1979] there were presented the most natural ones, among those found during the first decade after the appearance of the notion. While NP-completeness of a particular problem, encountered in algorithmist's practice, is enough for him to be sure it is impossible to solve exactly this problem effectively (in time polynomial in the length of the input), the theory of computation complexity also considers questions inspired by pure mathematics fields, such as number theory and model theory. A tight relationship between computability theory and number theory was stated in Matiyasevich's theorem on equivalence of enumerablility and diophantiness of sets and consequently undecidability of Hilbert's 10th problem ([Matiyasevich, 1970] and a textbook [Matiyasevich, 1993]). Subsequent researches were partly concentrated on the lower bound for the number of variables in diophantine equation which preserved undecidability.

From the point of view of the complexity theory the intriguing question is the complexity of deciding solvability in non-negative integers of diophantine equations in two variables. S.Smale in his "Mathematical problems for the next century" [Smale, 2000] points out on the importance of these questions in his 5th problem, independently of the famous **P**?**NP** problem (number 3 in his list).

Further researches on the existential fragments of theories of non-negative integers, weaker than for addition and multiplication (Hilbert's 10th problem), resulted in decidability of the so-called "Diophantine problem for addition and divisibility". This result, obtained independently by A.P.Bel'tyukov [Bel'tyukov, 1976] and L.Lipshitz [Lipshitz, 1976], found various applications in computer science (e.g. in [Degtyarev, etc., 1996], [Bozga, Iosif, 2005]). An expressiveness of the language makes it useful in formulation of particular problems, such as reachability problem for one-counter machines and their modifications ([Haase, 2012], [Bundala, Ouaknine, 2016]), which leads their decidability.

In the following sections we will firstly represent complexity results concerning simultaneous divisibilities of linear polynomials and then consider some sub-problems and related questions.

Definitions and complexity results for simultaneous divisibility of values of linear polynomials

The decidability of the Diophantine problem for addition and divisibility means the existence of an algorithm for recognizing every satisfiable in non-negative integers quantifier-free formula of the first order language in the signature $\langle +, 1, | \rangle$, where $|$ is a predicate symbol for the relation of divisibility of integers and $x|y$ means "x is a divisor of y". The general question is reducible to the satisfiability in non-negative integers of a linear divisibilities system, i.e. formula of the form

$$\&_{j=1}^{m}(f_i(x_1, ..., x_n) \mid g_i(x_1, ...x_n)), \tag{1}$$

where $f_i(x_1, ..., x_n)$ and $g_i(x_1, ..., x_n)$ are linear polynomials with non-negative integer coefficients.

The study of the complexity of the problem was started by L.Lipshitz in [Lipshitz, 1981] and resulted in the proof of its NP-completeness for every fixed (greater than 5) number of divisibilities in a system.

Theorem 1 ([Lipshitz, 1981]) The Diophantine problem for addition and divisibility is NP-hard (and NP-complete for every fixed number of divisibilities $m \geq 5$ in 1).

It is very important that while for every arbitrarily large but fixed number of divisibilities the problem is in the class **NP**, it is only NP-hard in the general case. This situation may be illustrated, for example, with the NP-complete problem of consistency in non-negative integers of a system of linear diophantine equations, which is closer to the practical algorithms. It appears to be in the class **P** for every fixed number of variables, that is, in some extent a tractable problem (see [Schrijver, 1986]).

There should be given some terminology remarks. The almost same problems have several names in different papers. In the decidability proof [Lipshitz, 1976] and in [Degtyarev, etc., 1996] we see "the Diophantine problem for addition and divisibility", in Russian literature "universal theory of natural numbers for addition and divisibility" as in [Bel'tyukov, 1976] or [Mart'yanov, 1977] (since a universal theory is decidable, the corresponding existential theory is also decidable). Furthermore, as "existential theory of $\langle \mathbb{N}; =, +, | \rangle$" in [Bes, 2002] and "existential Presburger arithmetic with divisibility" in the recent papers [Lechner, Ouaknine, Worrell, 2015] and [Bundala, Ouaknine, 2016]. In abbreviated form it is written as $\exists PAD$. Also, when we speak about the problem of consistency in non-negative integers of a system of linear diophantine equations, it is, in other words, the problem of satisfiability of a quantifier-free formula of Presburger arithmetic (abbreviated as $\exists PA$).

Detailed analysis of the decision procedure of L.Lipshitz was performed in [Lechner, Ouaknine, Worrell, 2015]. There was shown that for every satisfiable formula an upper bound for every assignment of variables would be doubly exponential in the length of the input. Thus, the problem belongs to the complexity class **NEXPTIME**, i.e. solvable on non-deterministic Turing machine using $2^{n^{O(1)}}$ number of steps, where n is the length of the input string with binary representation of the values of the input.

Theorem 2 ([Lechner, Ouaknine, Worrell, 2015]) The Diophantine problem for addition and divisibility ($\exists PAD$) is in the complexity class **NEXPTIME**.

With this result we have a complexity upper bound for those problems, which are reducible to the $\exists PAD$, in particular, those presented in the introduction. The exact complexity is not known and the problem of its determination remains open. It should also be noted that the existence of an instance of a problem with minimal solution in binary representation of size exponential in the length of the input, does not mean the problem is not in **NP**. For example, J.C.Lagarias in [Lagarias, 2006] shows that there are instances of the negative Pell equation (or anti-Pellian equation as it is named in the paper)

$$x^2 - dy^2 = -1 \tag{2}$$

with minimal non-negative integer solution of exponential length, while the problem is in **NP** since there exists a succinct certificate to establish solvability of the equation. Therefore, further progress can be quite a difficult task of significant importance for theoretical computer science.

Some sub-problems and related problems

Thus we know that $\exists PAD$ is NP-hard and we even do not know it is in **NP**. The corresponding algorithm is very impracticable. However, it could happen that for some applications it is sufficient to deal with sub-problems in order to state NP-completeness or the existence of a polynomial algorithm. In this section we will consider systems of divisibilities of a number by values of linear polynomials with non-negative coefficients and the opposite problem of systems of divisibilities of values of linear polynomials with non-negative coefficients by a number. There will not be any restrictions on the number of divisibilities, but on the number of non-zero coefficients in every polynomial. Proofs of some results, presented in the section, will be published in [Kosovskii, etc., 2017].

The first one could be considered as validity in positive integers of a formula of the form

$$\exists x_1 ... \exists x_n \, \&_{i=1}^{m} (K \mid f_i(x_1, ..., x_n)). \tag{3}$$

If there is no restriction on values of the variables, it will be a system of linear congruences

$$\exists x_1 ... \exists x_n \, \&_{i=1}^{m} (f_i(x_1, ..., x_n) \equiv 0 (\mathrm{mod}\, K)), \tag{4}$$

which could be solved in polynomial time in accordance with [Cohen, 1993] (section 2.3.4). If the variables will take their values from the interval of positive integers $[D, D']$, $0 < D \leq D' < K$, the problem is obviously in **NP**. This problem is NP-complete for every $K > 2$ (if $K = 2$ the problem is trivially in the class **P**) and exactly three non-zero coefficients of the variables in each polynomial (in [Kosovskii, etc., 2017]).

Since we are interested mainly in the number of non-zero coefficients, it will be convenient to use some abbreviations. Let $\bar{x} = (x_1, ..., x_n)$ be the list of variables of a formula and let in this case the fact that there are not greater than k non-zero coefficients in a polynomial be written down as ${}^{k}f_i(\bar{x})$. Thus, with this notation we have the following theorem.

Theorem 3 ([Kosovskii, etc., 2017]) The problem of satisfiability on the interval of positive integers $[D, D']$, $0 < D \leq D' < K$ of formulas of the form $\&_{i=1}^{m}(K \mid {}^{3}f_i(\bar{x}))$ is NP-complete for every $K \geq 3$.

From the point of view of the number of non-zero coefficients, we have the following result.

Theorem 4 ([Kosovskii, etc., 2017]) The problem of satisfiability on the interval of positive integers $[D, D']$, $0 < D \leq D' < K$ of formulas of the form $\&_{i=1}^{m}(K \mid {}^{k}f_i(\bar{x}))$ is NP-complete for every $k \geq 2$ and is in the class **P** for $k = 1$.

The case $k \geq 3$ in theorem 4 is a corollary of the Theorem 3, while for only two non-zero coefficients in each polynomial there is a polynomial reduction from GOOD SIMULTANEOUS APPROXIMATION ([Lagarias, 1982]). In his proof, J.C.Lagarias has used constructions, as he writes "inspired by Manders and Adleman" ([Manders, Adleman, 1978]). This method was introduced by K.L.Manders and L.Adleman for encoding an instance of a special case of KNAPSACK problem to an instance of solvability in non-negative integers of a quadratic diophantine equation of the form $ax^2 + by = c$ with positive integer coefficients. Possibly, the proof in [Kosovskii, etc., 2017] could be made more natural by means of polynomial reduction from 3-SAT with "conversion lemma" separately formulated in [Manders, Adleman, 1978].

For a system of divisibilities of a number on linear polynomials, the first result was achieved in [Adleman, Manders, 1977]. The problem LINEAR DIVISIBILITY (abbreviated as LD) of solvability in positive integers of one divisibility of the form $ax+1 \mid K$ was shown to be γ-complete. A problem is γ-reducible to another problem if there is a reduction procedure that can be performed in polynomial time on a non-deterministic Turing machine. Thus, a polynomial reduction is a special case of a γ-reduction. A problem is called γ-complete if it is in **NP** and every problem in **NP** is γ-reducible to it. These problems most likely are not in the class **P** nor are NP-complete; for the discussion of the notion see pages 158-160 in [Garey, Johnson, 1979].

The primary object of interest for K.L.Manders and L.Adleman was the complexity of an equivalent problem of solvability in non-negative integers of a binary quadratic equation of the form $axy + by = c$ and the study of the diophantine complexity. From this result one can conclude that simultaneous divisibility of a number by values of linear polynomials is γ-complete. By polynomial reduction from ONE-IN-THREE 3-SAT (in [Garey, Johnson, 1979]), the problem of validity in positive integers of formulas of the form

$$\exists \bar{x} \,\&_{i=1}^{m}({}^{3}f_i(\bar{x}) \mid K) \tag{5}$$

is NP-complete for every $K \geq 4$ and is in the class P for $0 < K < 4$. Though there is much more freedom for polynomial reductions for the problem

$$\exists \bar{x} \,\&_{i=1}^{m}({}^{2}f_i(\bar{x}) \mid K) \tag{6}$$

in comparison with LD, the proof of its NP-completeness does not look like obvious.

Among the references on the decidability proof of $\exists PAD$, the paper [Mart'yanov, 1977] on the decidability of the universal theory of non-negative integers for addition and D(x,y,z) predicate, true for each triplet (x,y,z) such that z=GCD(x,y), is sometimes mentioned (in [Degtyarev, etc., 1996]). As it was mentioned above, the decidability of a universal theory is equivalent to the decidability of the corresponding existential theory. This decision problem is equivalent to $\exists PAD$ because of the mutual existential definability of the predicates (see remarks in [Belyakov, Mart'yanov, 1983]). We have

$$x \mid y \;\Leftrightarrow D(x,y,x) \tag{7}$$

and in other direction

$$D(x,y,z) = \exists u(z \mid x \,\&\, z \mid y \,\&\, x \mid u \,\&\, y \mid z+u), \tag{8}$$

$$\neg D(x,y,z) = \exists u(\neg(z \mid x) \;\vee\; \neg(z \mid y) \;\vee\; (u \mid x \,\&\, u \mid y \,\&\, \neg(u \mid z)). \tag{9}$$

NP-completeness of the problem

$$\exists \bar{x}(GCD({}^{3}f_1(\bar{x}), ..., {}^{3}f_m(\bar{x})) = K) \tag{10}$$

on every non-empty and non-trivial integer interval could be proved in the same manner as in Theorem 3 by polynomial reduction of ONE-IN-THREE 3-SAT ([Kosovskii, Starchak, 2016]). Corresponding question for only two non-zero coefficients does not look like as an evident consequence of Theorem 4.

Conclusion

One of the aims of the paper was to show the importance and actuality of problems concerning divisibilities of values of linear polynomials. Although the general problem has high complexity lower bound, some sub-problems could appear sufficient in applications for determining complexity of various problems, arising, for example, in counter automata theory.

The other purpose was to point to a number-theoretical interest in the problem and its possible relations with complexity of solvability of quadratic diophantine equations.

Bibliography

Adleman L., Manders K.L., "Reducibility, randomness and intractability" // Proceedings of the 9th Annual ACM Symposium on Theory of Computing, 1977, pp. 151-163.

Bel'tyukov A.P., "Decidability of the universal theory of the natural numbers with addition and divisibility" // Zapiski Nauchnyh Seminarov LOMI, Vol. 60, 1976, pp. 15-28. (in Russian) English translation, Journal of Soviet Mathematics, Vol. 14, No. 5, 1981, pp. 1436-1444.

Belyakov E.B., Mart'yanov V.I., "Universal theories of integers and the extended Bliznetsov hypothesis" // Algebra i Logika, Vol. 22, 1983, pp. 26-34. (in Russian) English translation, Algebra and Logik, Vol. 22, 1983, pp. 19-26.

Bes A., "A survey of arithmetical definability. A Tribute to Maurice Boffa" // Bulletin de la Societe Mathematique de Belgique, 2002, pp. 1-54.

Bozga A., Iosif R., "On decidability within the arithmetic of addition and divisibility" // Proceedings of FoSSaCS, ser. Lecture Notes in Computer Science, Springer, Vol. 3441, 2005, pp. 425-439.

Bundala D., Ouaknine J., "On parametric timed automata and one-counter machines" // Inf. Comput., 2016, http://dx.doi.org/10.1016/j.ic.2016.07.011

Cohen H., "A Course in Computational Algebraic Number Theory", ser. Graduate Texts in Mathematics. Springer-Verlag, Vol. 138, 1993.

Degtyarev A., Matiyasevich Y., Voronkov A., "Simultaneous rigid E-unification and related algorithmic problems" // Proceedings 11th Annual IEEE Symposium on Logic in Computer Science, 1996, pp. 494-502.

Garey M.R., Johnson D.S., "Computers and Intractability: A Guide to the Theory of NP-Completeness", Freeman, New York, 1979.

Haase C., "On the complexity of model checking counter automata" // Thesis, University of Oxford, 2012.

Kosovskii N.K., Starchak M.R., "NP-complete problems for greatest common divisor of values of linear polynomials" // Proceedings of the 9th conference ITU-2016, St. Petersburg, 2016, pp. 71-72. (in Russian)

Kosovskii N.K., Kosovskaya T.M., Kosovskii N.N., Starchak M.R., "NP-complete problems for systems of divisibilities of values of linear polynomials" // Vestn. St. Petersburg Univ.: Math., 2017, to be published. (in Russian)

Lagarias J.C., "The computational complexity of simultaneous diophantine approximation problems" // 23th Annual Symposium on Foundations of Computer Science, IEEE, New York, 1982, pp. 32-39.

Lagarias J.C., "Succinct certificates for the solvability of binary quadratic diophantine equations" // e-print arXiv:math/0611209v1, 2006. Extended and updated version of a 1979 FOCS paper in Proceedings of the 20th IEEE Symposium on Foundations of Computer Science , IEEE Press, 1979, pp. 47-54.

Lechner A., Ouaknine J., Worrell J., "On the Complexity of Linear Arithmetic with Divisibility" // Proceedings of the 30th Annual ACM/IEEE Symposium on Logic in Computer Science(LICS), 2015, pp. 667-676.

Lipshitz L., "The Diophantine problem for addition and divisibility" // Transactions of the American Mathematical Society, Vol. 235, 1976, pp. 271-283.

Lipshitz L., "Some remarks on the Diophantine problem for addition and divisibility" // Bull. Soc. Math. Belg. Ser. B, Vol. 33, No. 1, 1981, pp. 41-52.

Manders K.L., Adleman L., "NP-Complete decision problems for binary quadratics" // Journal of Computer and System Sciences, Vol. 16, No. 2, 1978, pp. 168-184.

Mart'yanov V.I., "Universal extended theories of integers" // Algebra i Logika, Vol. 16, No. 5, 1977, pp. 588-602. (in Russian) English translation, Algebra and Logik, Vol. 16, No. 5, 1977, pp 395-405.

Matiyasevich Y.V., "Enumerable sets are diophantine" // Doklady Akademii Nauk SSSR, Vol. 191, 1970, pp. 279-282. (in Russian) English translation, Journal of Soviet Mathematics, Doklady, Vol. 11, No. 2, 1970, pp 354-358.

Matiyasevich Y.V., "Hilbert's 10th problem", MIT Press, 1993.

Schrijver A., "Theory of Linear and Integer Programming", John Wiley and Sons, New York, 1986.

Smale S., "Mathematical problems for the next century" // Mathematics: frontiers and perspectives, Amer. Math. Soc., 2000, pp. 271-294.

Authors' Information

Nikolay K. Kosovskii *- Dr., Professor of Computer Science Chair of St. Petersburg State University, University av., 28, Stary Petergof, St. Petersburg, 198504, Russia; e-mail: kosov@NK1022.spb.edu*
Major Fields of Scientific Research: Mathematical Logic, Theory of Computational Complexity of Algorithms

Mikhail Starchak *- PhD student of Computer Science Chair of St. Petersburg State University, University av., 28, Stary Petergof, St. Petersburg, 198504, Russia; e-mail: mikhstark@gmail.com*
Major Fields of Scientific Research: Theory of Computational Complexity of Algorithms

ТАКСОНОМИЗАЦИЯ ЕСТЕСТВЕННО-ЯЗЫКОВЫХ ТЕКСТОВ

Виталий Приходнюк

***Аннотация**: В статье описывается подход формирования таксономий на основе семантического анализа текстовых массивов. Представлен алгоритм и описаны основные этапы его работы, определена спецификация входных и выходных данных в виде бестиповых термов ламбда-исчисления. Приведены так же вспомогательные алгоритмы выделения различных типов (в частности, географической) информации. Дается оценка эффективности предложенного алгоритма, полученная с помощью вычислительных экспериментов.*

***Ключевые слова**: таксономия, гиперотношение, структуризация, инженерия знаний*

***ACM Classification Keywords**: I.2 ARTIFICIAL INTELLIGENCE - I.2.4 Knowledge Representation Formalisms and Methods, H. Information Systems*

Введение

Быстрый рост тематических объемов информации, необходимость ее более качественной обработки и усвоения требуют использования методов, средств получения информации и преобразования ее в такую форму, с которой будет удобнее работать на всех этапах решения задач. Главная цель такого преобразования заключается в нахождении документов по нужной тематике, обработке и анализе текстовых (естественно-языковых) документов с помощью определенных инструментов, которые позволяют выявлять свойства описанных объектов и логические закономерности, существующие между ними. По мнению ученых, предложенная должным образом информация позволяет увидеть те дополнительные скрытые закономерности, которые не удается обнаружить другими методами [Величко, 2009; Гаврилова, 2001; Валькман, 2012; Гладун, 1994; Van Rijsbergen, 1979; Helbig, 2006].

Таким образом, актуальной является задача идентификации терминов, которые совместно с контекстами определяют содержание документа и семантические связи между ними. Это дает возможность в дальнейшем сформировать топологическую структуру текста в виде таксономии, отображения и создания новых объектов, связей, увязки новых атрибутов, которые могут быть

использованы при аналитической обработке информации [Стрижак, 2014; Величко, 2015]. Такие топологические структуры могут использоваться при формировании информационной среды корпоративных систем (КС), сетевые инструменты которых обеспечивают поиск, формулировки, формирования, структурирования и представления информации и сообщений, из которых в дальнейшем формируются знания и принимаются соответствующие решения.

Процесс выделения информации

Эффективная обработка неструктурированных (в частности, написанных естественно-языковым текстом) документов может достигаться с помощью их таксономизации [Стрижак, 2014; Шаталкин, 2012] с последующим представлением в виде онтологического графа [Величко, 2009].

На процесс взаимодействия с текстовыми информационными ресурсами, особенно в сетевой среде влияют такие три аспекта, как:

а) синтаксический, который касается формальной правильности сообщений с точки зрения синтаксических правил языка, используемого безотносительно к его содержанию;

б) семантический, который отражает уровень понятийного взаимодействия;

в) прагматический, который определяет операциональные аспекты их использования.

Первичная обработка информационных ресурсов, особенно при их использовании, требует решения целого звена проблем, которые также характеризуют процессы взаимодействия. К этим проблемам специалисты относят следующие: распределенность; гетерогенность; интероперабельность информации только на синтаксическом и структурном уровнях; неполную ответственность за информацию, передаваемую при интеграции; дублирование информации; потерю полноты контроля доступа к информации; технологические трудности, связанные с разнообразием форматов представления данных; содержательность конфликтов между информационными единицами на понятийном уровне; информационная энтропия источника информации. И каждая из этих проблем имеет свои определенные проблемные вопросы с точки зрения технологии ее решения.

В рамках данного процесса необходимо пройти три ключевых *этапа*:

- идентификацию множества терминов концептов X, принадлежащих заданному тексту терминополя;
- идентификацию множества семантических отношений R_{sem} между концептами;

— идентификацию множества атрибутов концептов A (таких, как географическая или темпоральная информация).

Идентификация возможна с помощью последовательного преобразования входного множества лексем естественно-языкового текста L с помощью последовательного применения правил из множества Rul.

На множестве лексем $l \in L$ с помощью оператора «$\prec$» (предшествует) определено линейный порядок, и, таким образом, L является линейно упорядоченной множеством $l_1 \prec l_2 \prec ... \prec l_n$. Также лексемы разбиты на предложения S_i, на множестве которых аналогичным образом задано линейный порядок $S = \{S_1 \prec S_2 \prec ... \prec S_m\}$.

Каждое предложение также представляет собой линейно упорядоченное множество: $S_i = \{l_1^i \prec l_2^i \prec ... \prec l_{n_i}^i\}$. Правила применяются отдельно к каждому из предложений S_i и действуют исключительно в рамках предложения. Сами правила имеют аппликативную форму и могут быть представлены в виде бестиповых выражений [Барендрегт, 1985; Стрижак, 2014]:

$$f_a = (\lambda x.t(x))a = t(a) \tag{1}$$

где:

— λ-теория – ламбда-исчисления; запись λx подразумевает, что это λ-терм;
— x – переменная, принимающая значения на множестве лексем L или концептов X;
— t – выражение, содержащее переменную;
— a – аргумент функции, определяющей возможные значения переменной x;
— f_a – функция, которая может быть применена к аргументу a.

Каждое такое правило задает преобразование одного из видов (2) – (4).

$$L \xrightarrow{Rul} X \tag{2}$$

$$X \xrightarrow{Rul} \langle X, R_{sem} \rangle \tag{3}$$

$$L \xrightarrow{Rul} A \tag{4}$$

Кроме того, возможны другие преобразования:

$$L \xrightarrow{Rul} L \quad (5)$$

$$L \xrightarrow{Rul} L^* \quad (6)$$

где множество L^* – множество конструктов.

Конструкт объединяет в себе несколько лексем, которые в дальнейшем обрабатываются как одна. Конструкты могут иметь такие же характеристики, как и лексемы, и могут быть связаны с другими лексемами или конструктами синтаксическими связями R_{syn} .

Любое правило вида (2), (4), а в некоторых случаях – и вида (5) может быть применено не только к множеству L , но и к множествам L^* или $L \cup L^*$.

Предварительная структуризация

Первым и наиболее очевидным источником структуры текста является его содержание. Оно представляет собой набор предложений $S_{toc} \subset S$, которые определенным образом выделены из основного текста. Чаще всего под содержание отводится несколько страниц в начале или в конце текста. Тогда содержание достаточно легко идентифицировать, задав его пределы, и воспользовавшись гиперотношением множественности порядка S [Клини, 1957; Малишевский, 1998]. Применение гиперотношения S , является необходимым условием и обеспечивает идентификацию конкретных мест в тексте, в которых позиционируются конкретные понятия, и на которые ссылаются элементы содержания. Описываемая процедура разметки текста реализуется на основе следующих двух правил:

$$T = \lambda l_1, l_2, ..., l_n.t \quad (7)$$

$$t \equiv \exists i, \forall j \in [1, n_i], S_i \in S_{toc} \cup l_j^i \in S_i \quad (8)$$

Конструктивность правил (7) и (8) позволяет их применять даже без предварительного использования процедур оригинальной разметки текста, что характеризует процессы формирования лингвистических корпусов [Широков, 2005].

При отсутствии содержания необходимо сформировать предикат q для анализа разметки и заменить условие (8) на (9).

$$t \equiv \exists i, \forall j \in [1, n_i], q\left(l_j^i\right) \quad (9)$$

После применения предиката идентификации выделенные им последовательности лексем формируют множество категорий $\{X_{cat}\}$. Благодаря линейному порядку лексем и предложений можно разбить оригинальный текст на части:

$$L_i^{cat} \equiv \left\{ l \middle| \forall l^{i-1} \in S_{i-1}^{toc}, \forall l^{i+1} \in S_{i+1}^{toc}, l^{i-1} \prec l \prec l^{i+1} \right\} \tag{10}$$

Каждую из множеств L_i^{cat} можно обрабатывать как отдельный текст.

Категории $\{X_{cat}\}$ формируют верхний уровень онтографа: все выделенные из фрагмента текста L_i^{cat} категории являются подкатегориям соответствующей категории X_i^{cat}.

Выделение концептов и связей

Выделение концептов и связей является сложным процессом через большую вариативность языковых конструкций возможных в тексте. Анализатор должен иметь формальное описание таких конструкций, а качество анализа напрямую зависит от полноты этого описания.

Описание предложений в виде правил вида (1). Конкретный вид правил зависит от типа правила и входящего подмножества лексем, для обработки которых предназначено это правило. Составляющими правилами есть предикаты идентификации вида (11) и (12), которые предназначены для обработки отдельной лексемы:

$$c_{a,b} = (\lambda x, y.t(x, y))a, b \equiv \langle a, b \rangle \in LP \tag{11}$$

$$r_{a,b,c} = (\lambda x, y, z.t(x, y, z))a, b, c \equiv \langle a, b, c \rangle \in LS \tag{12}$$

Для каждого предиката определенным образом формируется множество LP или LS. Так LP представляет собой множество лексем и может быть определена двумя способами: простым перечнем допустимых лексем или определением определенного признака, что формирует категорию таких лексем. А LS представляет собой множество пар лексем, связанных определенным видом синтаксической связи. Таким образом, каждый предикат определяется на основе выделения соответствующего ему множества допустимых лексем.

На основе таких предикатов формируется правило вида (13):

$$rul = c_{x_1 p_1} \wedge c_{x_2 p_2} \wedge c_{x_n p_n} \wedge r_{x_1 x_2 k_{12}} \wedge r_{x_2 x_3 k_{23}} \wedge r_{x_{n-1} x_n k_{n-1n}} \quad (13)$$

Применение правила заключается в нахождении упорядоченного множества лексем (14), для которых выполняется условие (15).

$$L_{rul} \subset L, l_1^{rul} \prec l_2^{rul} \prec \ldots \prec l_n^{rul} \quad (14)$$

$$(\lambda x_1, x_2, \ldots, x_n . rul) l_1^{rul}, l_2^{rul}, \ldots, l_n^{rul} \quad (15)$$

Правила вида (2), (4), (5), (6) в дальнейшем выполняют преобразование (16) – (19) соразмерно:

$$\left\{l_1 \ldots l_k, l_1^{rul} \ldots l_n^{rul}, l_{k+n} \ldots l_m\right\} \xrightarrow{rul} \left\{l_1 \ldots l_k, x, l_{k+n} \ldots l_m\right\} \quad (16)$$

$$\left\{l_1 \ldots l_k, l_1^{rul} \ldots l_n^{rul}, l_{k+n} \ldots l_m\right\} \xrightarrow{rul} \left\{l_1 \ldots l_k, \mathrm{a}, l_{k+n} \ldots l_m\right\} \quad (17)$$

$$\left\{l_1 \ldots l_k, l_1^{rul} \ldots l_n^{rul}, l_{k+n} \ldots l_m\right\} \xrightarrow{rul} \left\{l_1 \ldots l_k, l, l_{k+n} \ldots l_m\right\} \quad (18)$$

$$\left\{l_1 \ldots l_k, l_1^{rul} \ldots l_n^{rul}, l_{k+n} \ldots l_m\right\} \xrightarrow{rul} \left\{l_1 \ldots l_k, l^*, l_{k+n} \ldots l_m\right\} \quad (19)$$

Правила формата (3) имеют другую структуру и выполняют преобразование (20):

$$\left\{l_1^{rul}, l_2 \ldots l_{n-1} l_2^{rul}\right\} \xrightarrow{rul} \left\langle \left\{l_1^{rul}, l_2 \ldots l_{n-1} l_2^{rul}\right\}, \left\{\left\langle l_1^{rul}, l_n^{rul}, R_{sem} \right\rangle\right\} \right\rangle \quad (20)$$

Выделение атрибутов

Выделение кандидатов в атрибуты происходит при выделении концептов правила (4). В результате их применения преобразованиями (17) формируется множество A^* . Для формирования множества собственно атрибутов A необходимо осуществить процедуру валидации элементов $a \in A^*$ и отбросить те, которые не пройдут валидацию.

Для каждого из возможных типов атрибутов создается отдельный предикат валидации q , что и определяет, должна ли лексема входить в итоговое множество атрибутов. Предикаты валидации зависят от многих факторов, в частности, типа текста, подмножества языка, обрабатываемого предметной области. Например, для географических координат условием валидности может

быть принадлежность координат определенной рабочей области. Образующий предикат – правило будет выглядеть (21), а сам предикат – вид (22):

$$f_q = x_{\min} < a_x < x_{\max} \wedge y_{\min} < a_y < y_{\max} \tag{21}$$

$$q(a_x, a_y) = \begin{cases} 1, f_q(a_x, a_y) \\ 0, \neg f_q(a_x, a_y) \end{cases} \tag{22}$$

На основе таких предикатов формируются множества атрибутов по типам (23) и обобщающее множество (24):

$$A^i = \left\{ a^i \mid q_i(a^i) \right\} \tag{23}$$

$$A = \bigcup_i A^i \tag{24}$$

Все элементы множества вида (24) могут быть использованы как атрибуты отображения различных массивов обрабатываемых текстов. За счет применения гиперотношения S в виде правил (7), (8), (20) они обеспечивают уникальность представления множественности их смыслов, и могут быть использованы в процедурах поиска и идентификации необходимых текстовых массивов. Также, указанные атрибуты, составляющие множество вида (24), могут быть использованы в процедурах интеграции распределенных текстовых массивов, которые имеют определенную степень смысловой эквивалентности. Более того конструктивность правил (2) – (8) и предиката (22) позволяет формировать процедуры расширения смыслов текстовых массивов при их интеграции, на основе связности их элементов гиперотношением множественности порядка S для всех элементов обрабатываемых текстов.

Построение таксономии текста

Сформированное множество атрибутов вида (24), характеризуется тем, что над всеми его элементами задается гиперотношение множественности порядка S . Тогда на основании

применения предикативного выражения вида (22), в нем всегда можно выделить непустое множество элементов, образующих бинарные пары вида:

$$\lambda((x_i)rul)S\lambda(y_i)rul) \tag{25}$$

каждый терм которого представляет определенную лексему обрабатываемых текстов. Конструктивной особенностью выражения (25) является представимость каждой бинарной пары в виде тематической тавтологии [Стрижак, 2014].

Процесс построения таксономии текста теперь, на основе применения правил (1) – (24), может быть представлен в виде следующей продукции:

$$\lambda((x_i)rul)S\lambda(y_i)rul) \Rightarrow \breve{T} = (\lambda(x.t(x), S, \prec) \tag{26}$$

Правило (26) задает индуктивность процесса формирования упорядоченных множеств концептов вида (24), между элементами которых устанавливаются гиперотношение множественности порядка, и фактически конструируется таксономия. Необходимым условием индуктивности является определение над концептами текстов предикативного выражения (22). Предикативные выражения, формулируются на основе концептов таксономической категории с заданным множественным отношением упорядоченности и принимают только значения истинности. Это позволяет формировать на основе терминов концептов таксономической системы, лингвистические выражения, которые содержательно отражают смысловые состояния текстовых массивов, как пассивной системы тематических знаний.

Так, для множества таксономических категорий = {тип (phylum) подтип (subphylum) класс (classis) подкласс (subclassis) ряд (у растений - порядок) (ordo) подряд (subordo) семья (familia) подсемейство (subfamilia) род (genus) подрод (subgenus) вид (species) подвид (subspecies) разновидность (varietas) форма (forma)}, гиперотношение бинарной множественности порядка, обеспечивает сохранение всех типов взаимодействия между термами, определяющих тематики текстов. Это также позволяет формировать все множества таксономий из концептов

сложившейся онтологической системы. Однако задав над указанными категориями отношение линейной упорядоченности, мы сужаем их перечень, так как ряд категорий, таких, как: класс (classis) форма (forma) вид (species) могут занимать в выражениях (24) и (25), представляющий бинарное отношение - «быть элементом категории», как левую, так и правую часть выражений вида (11), (12) и (26).

В заключение отметим, что согласно [Малишевский, 1998; Стрижак, 2014], бинарные выражения, составляющие правила (24) – (26) обладают свойствами агиперцикличности, иррефлексивности, гипертранзитивности и регулярности:

– агиперцикличность – если для S не существует гиперцикличного множества концептов $X \subseteq U$ такого, когда:

$$\forall x \in X \exists Y \subseteq X : YSx \tag{27}$$

– иррефлексивность:

$$YSx \Rightarrow (Y / \{x\})Sx \tag{28}$$

– гипертранзитивность:

$$YSx, x \in X, XSz \Rightarrow ((Y \cup X) / \{x\})Sz \tag{29}$$

– регулярность:

$$YSx, Y' \supseteq Y \Rightarrow Y'Sx \tag{30}$$

Указанные свойства позволяют в дальнейшем реализовывать процедуры, которые обеспечивают формирование на основе выделенных таксономий, тематических онтологических систем [Гаврилова, 2001; Валькман, 2012; Гладун, 1994; Стрижак, 2014; Guarino, 1994].

Оценка эффективности

Из-за чрезвычайно большой вариативности предикатов и правил оценка эффективности работы алгоритмов вполне может быть выполнена только экспериментальными методами.

Очень чувствительна эффективность к качеству предварительной обработки текста (лексического анализа) и соответствия базы правил языка.

Для оценки качества исчисляются такие параметры:

- *TP (True Positive)* – количество объектов, которые правильно идентифицированы системой;
- *FP (False Positive)* – количество последовательностей лексем, которые не определяют объекты, но были идентифицированы системой;
- *FN (False Negative)* – количество объектов, которые не были идентифицированы системой.

Оценка качества работы алгоритма осуществляется по следующим параметрам [Helbig, 2006]:

- *Precision (точность)* представляет собой отношение корректно идентифицированных объектов количества всех идентифицированных системой объектов;
- *Recall (полнота)* представляет собой отношение количества правильно выделенных идентифицированных системой объектов с количеством всех объектов в тексте;
- *F (мера)* представляет собой интегральный показатель точности и полноты, а также вычисляется как их среднее гармоничное.

Данные параметры определяются по формулам (25) – (27):

$$Precision = \frac{TP}{TP + FP} \tag{31}$$

$$Recall = \frac{TP}{TP + FN} \tag{32}$$

$$F = (\beta^2 + 1)\frac{Precision \times Recall}{\beta^2 Precision + Recall} \tag{33}$$

Поскольку трудно оценить относительную важность точности и полноты в процессе выделения текстов, поэтому есть смысл использовать сбалансированную F-меру (28):

$$F = \frac{Precision \times Recall}{Precision + Recall} \tag{34}$$

Результаты вычисления эффективности работы алгоритма для простого текста, описывающего географическое размещение различных объектов, показано в таблице:

	Точность	Полнота	F-мера
Имена	0,723404	0,85	0,781609
Географическая информация	0,92	0,741935	0,821429
Всего	0,824742	0,784314	0,80402

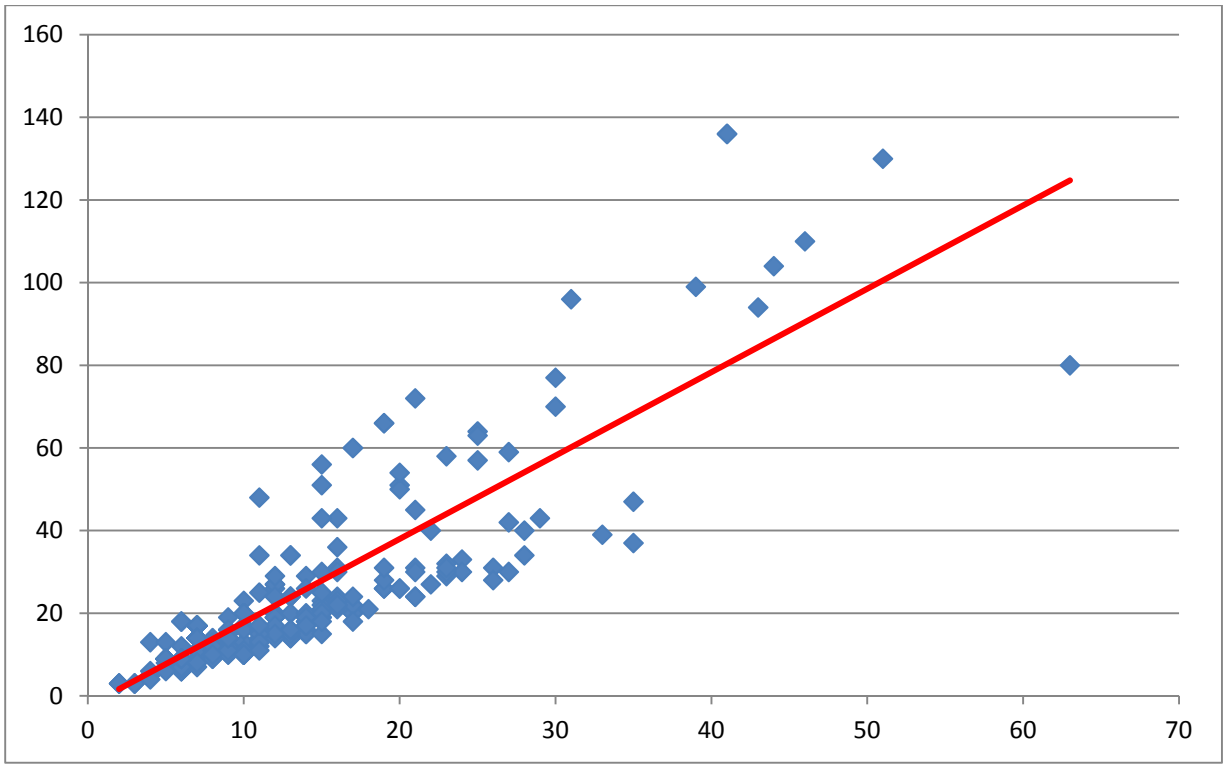
Как видно из таблицы, качество выделения имен (что, по сути, являются концептами) и географической информации (атрибуты концептов) кардинально отличаются.

Географическая информация после идентификации проходит процедуру валидации, которая позволяет добиться чрезвычайно высокой точности. Однако в ходе такой процедуры некоторые выделенные элементы данных отбрасываются, что значительно снижает полноту.

Для имен ситуация прямо противоположная – для них не существует эффективных алгоритмов валидации, поэтому точность их идентификации относительно низкая. Но благодаря тому, что не происходит отвержение элементов данных, повышается их полнота.

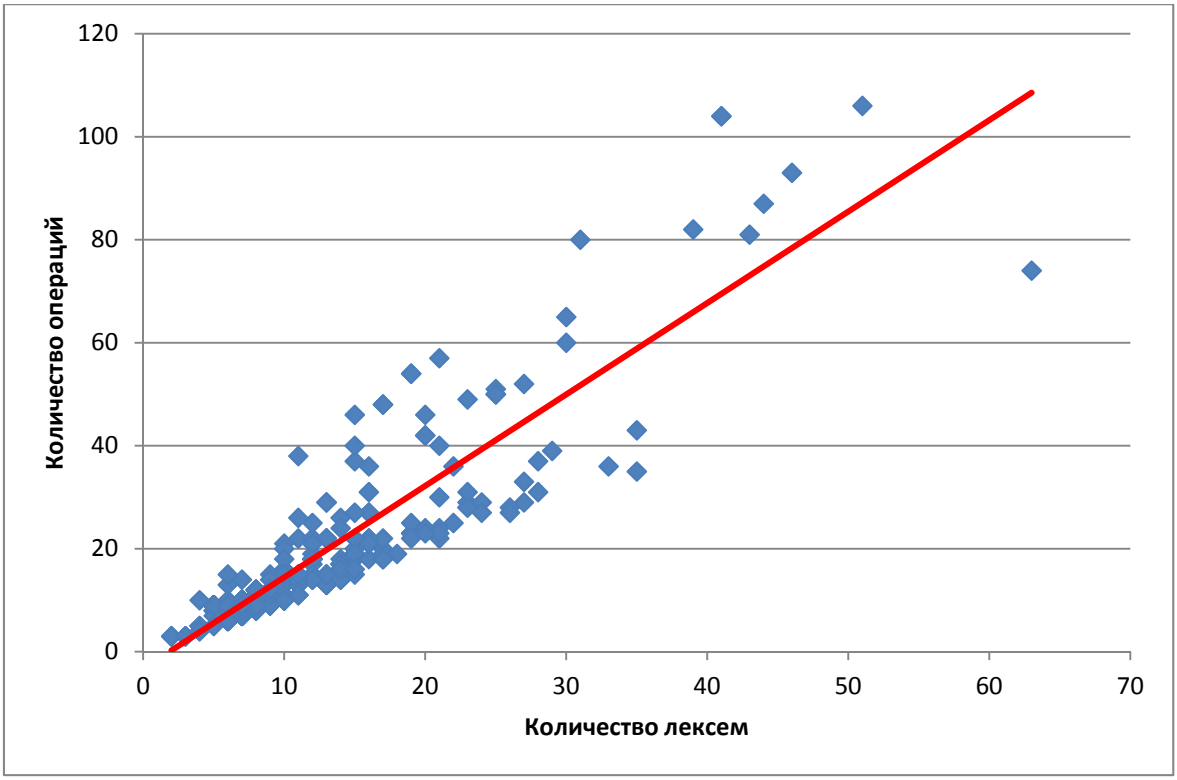
В целом эффективность выделения имен несколько ниже, поскольку в способах написания имен значительно больше вариативности.

Быстродействие работы алгоритма напрямую зависит от количества вызова операций применения предиката к входной лексеме. Поскольку каждое правило работает в рамках одного предложения, то зависимость быстродействия от количества предложений является линейной. Зависимость скорости от длины предложения показано ниже на диаграмме:



Как видно, зависимость количества вызовов операций (и, соответственно, производительности) от длины предложения, близка к линейной. При этом эффективность обработки более распространенных коротких предложений больше, чем менее распространенных длинных.

Также быстродействие зависит от размеров базы правил. Например, если отбросить около половины правил, быстродействие алгоритма несколько повысится, и будет выглядеть следующим образом:



Как видно, основным фактором, влияющим на скорость работы алгоритма, является размер входного текста. Зависимость скорости работы алгоритма от длины текста (в лексемах) показано на графике:

В целом алгоритм имеет достаточно высокое быстродействие и пригоден для использования в процессах поддержки принятия решений, требующих оперативного анализа данных. Алгоритм также пригоден для структуризации больших объемов данных, таких, как книги.

Выводы

Таким образом, применение гиперотношения множественного порядка к множеству понятий, составляющих терминополе текстов, позволяет их представлять как в виде бестиповых термов, описываемых при помощи λ-теории ламбда-исчисления. Бестиповые процедуры обеспечивают процесс идентификации мест позиционирования конкретных понятий текста, выделяют бинарные структуры, которые могут быть также представлены тематическими тавтологиями. За счет этого реализуется структуризация текстовых массивов и формирование их таксономических систем. В общем случае таксономию текстового массива можно определять как архитектуру соответствующего документа. Другими словами таксономия есть обобщающее покрытие тематического многообразия текстового документа.

Конструктивной характеристикой указанного многообразия является ее интегративность и выявление смысловых связностей, представляемых в виде продуцируемых гиперотношением множественности порядка тематических тавтологий и установлением между ними бинарного отношения порядка.

Более того бестиповая множественность порядка определяет достаточно эффективные процедуры разметки и идентификации понятий терминополя текстовых документов и как следствие достаточно эффективное многообразие их таксономических систем .

Bibliography

[Guarino, 1994] Guarino N. The Ontological Level [Текст] / N. Guarino, R. Casati, N. Smith, G. White // Philosophy and the Cognitive Sciences. – Vienna : Holder-Pichler-Tempsky, 1994. – p. 443-456.

[Helbig, 2006] Hermann Helbig: Knowledge Representation and the Semantics of Natural Language [Text]. – Berlin : Springer, 2006. – 651 p.

[Van Rijsbergen, 1979] Van Rijsbergen, C. J. Information Retrieval (2nd ed.). – Butterworth. – 1979. – 208 p.

[Барендрегт, 1985] Барендрегт Х. Ламбда-исчисление. Его синтаксис и семантика: Пер. с англ. – М. : Мир, 1985. – 606 с.

[Валькман, 2012] Валькман Ю. Р. Модельно-параметрическое пространство: теория и применение : [монография] / Ю. Р. Валькман, В. И. Гриценко, А. Ю. Рыхальский. – К. : Наукова думка, 2012. – 192 с.

[Величко, 2009] Величко В. Автоматизированное создание тезауруса терминов предметной области для локальных поисковых систем / В. Величко, П. Волошин, С. Свитла // «Knowledge – Dialogue – Solution» International Book Series «INFORMATION SCIENCE & COMPUTING», Number 15. – FOI ITHEA Sofia, Bulgaria. – 2009. – p. 24–31.

[Величко, 2015] Построение таксономии документов для формирования иерархических слоев в геоинформационных системах [Текст] / Виталий Величко, Виталий Приходнюк, Александр Стрижак, Крассимир Марков, Крассимира Иванова, Стефан Карастанев // International Journal "Information Content and Processing", 2015. – Volume 2. – Number 2. – p.181-199.

[Гаврилова, 2001] Гаврилова Т. А. Базы знаний интеллектуальных систем [Текст] / Т. А. Гаврилова, В. Ф. Хорошевский. – СПб. : Питер, 2001. – 384 с.

[Гладун, 1994] Гладун В. П. Процессы формирования новых знаний [Текст] / В. П. Гладун. – София : СД «Педагог 6», 1994. – 189 с.

[Клини, 1957] Клини, С. К. Введение в метаматематику [Текст] / С. К. Клини. – М. : Иностранная литература, 1957. – 526 с.

[Малишевский, 1998] Малишевский А. В. Качественные модели в теории сложных систем / А. В. Малишевський. – М. : Наука. Физматлит, 1998. – 528 с.

[Стрижак, 2014] Стрижак А.Е. Таксономические характеристики онтологических систем [Текст] / А. Е. Стрижак // Бионика интеллекта, 2014. – № 2(83). – с. 24-29.

[Стрижак, 2014] Стрижак О. Є. Трансдисциплінарна інтеграція інформаційних ресурсів [Текст] : автореф. дис. ... д-ра техн. наук : 05.13.06 / Стрижак Олександр Євгенійович ; Нац. акад. наук України, Ін-т телекомунікацій і глобал. інформ. простору. Київ, 2014. 47 с.

[Шаталкин, 2012] Шаталкин, А.И. Таксономия. Основания, принципы и правила [Текст] / А. И. Шаталкин. – М. : Товарищество научных изданий КМК, 2012. – 600 с.

[Широков, 2005] Широков В. А., Булгаков О. В., Грязнухіна Т. О. та ін. Корпусна лінгвістика [Текст] / В. А. Широков, О. В. Булгаков, Т. О. Грязнухіна та ін. – К.: Довіра, 2005. – 471 с.

Authors' Information

Виталий Приходнюк – аспирант, Институт телекоммуникаций и глобального информационного пространства НАН Украины, Киев-186, 03186, Чоколовский бульвар, 13; e-mail: vitalik1700@yandex.ru

Основные области научных исследований: Data Mining, геоинформационные системы, онтологический инжиниринг

Taxonomyzation of Natural Language Texts

Vitaly Prihodnyuk

Abstract*: An approach for forming taxonomies based on semantic analysis of text arrays is presented in this paper. The algorithm and the main stages of its work are described. The specification of input and output data is defined as lambda calculus terms without types. Auxiliary algorithms for detecting different types (in particular, geographic) information are outlined. The evaluation of the effectiveness of the proposed algorithm, obtained by computational experiments, is given.*

Keywords*: Taxonomy, Hyper-relation, Structuring, Knowledge Engineering*

ACM Classification Keywords*: I.2 ARTIFICIAL INTELLIGENCE - I.2.4 Knowledge Representation Formalisms and Methods, H. Information Systems*

ITHEA SAMPLE SHEET FOR PREPARING PAPERS

Krassimir Markov

Abstract: *The new rules for preparing the manuscripts for the ITHEA International Journals (IJ), International Conferences, and International Book Series (IBS) are given. These rules will be obligatory from the 2017 year. The form for the papers is shown by this sheet.*

Keywords: *ITHEA formatting rules.*

ITHEA Keywords: *Please use keywords from http://idr.ithea.org/tiki-browse_categories.php .*

Introduction

We ask authors to follow some simple guidelines.

In essence, we ask authors to make papers look exactly like this document.

This text is a sample for preparing the manuscripts for publishing in ITHEA International Journals, Conferences, and Book Series. All styles needed for formatting the papers are included.

The easiest way to prepare your manuscript in accordance of these rules is simply to replace the content of this sample sheet with your own material.

Responsibility for papers published in ITHEA International Journals and Book Series belongs to authors.

Please *get permission to reprint* any copyrighted material.

The camera-ready copy of the paper should be received by the ITHEA Journal Submission System (*http://ij.ithea.org*) or respectively by the ITHEA Conference Submission System (*http://ita.ithea.org*); e-mail for questions: *info@foibg.com*.

Instructions for Preparation of Manuscripts

The authors are hoped to prepare manuscripts in close accordance with the instructions given below.

This text is a sample for preparing the articles. All styles needed for formatting the papers are included. Do not include any new styles. Please, *do not use automatic numbering anyway*, because of losing the information during the assembling the journals or books.

Name the file of the manuscript beginning with the journal or conference name, following with the family names of the authors or if they are more than 2 authors – name of the first author, followed by "_et_al".

For instance if the manuscript will be submitted to:

- IJ ITK 2017 from Jackson and Williams, than the file needs to be named: "**IJITK10-Jackson_Williams.doc**";
- IJ ITA 2017 from Jackson, Williams, and Davis, than the file needs to be named: "**IJITA10-Jackson_et_al.doc**";
- i.TECH 2017 from Jackson and Williams, than the file needs to be named: "**iTECH10-Jackson_Williams.doc**";
- i.TECH 2017 from Jackson, Williams, and Davis, than the file needs to be named: " **iTECH10-Jackson_et_al.doc**".

Manuscripts will be peer reviewed and evaluated for originality, significance, clarity, and soundness according to criteria of peer review procedure.

The authors of the accepted manuscripts will be allowed to make a correction in accordance with the suggestions of the reviewers and to submit final camera-ready manuscripts within the stipulated deadline.

ITHEA Manuscript Preparing Styles (ITHEA MPS) are obligatory for manuscripts submitted to ITHEA.

In order to see ITHEA MPS please press ALT+CTRL+SHIFT+S in this ITHEA template document. You will see the window shown on Figure 1. To apply style to some part of text, please press on name of the style.

How to apply ITHEA MPS styles to your document:

- Open the file you would like to format and click *Office Button→Word Options* (Word 2007) /*File→Options* (Word 2010).
- Choose *Add-Ins* on the left side, and then select *Templates* in the drop-down list at the bottom of the dialog. Click *Attach* in the dialog box that opens, navigate to your working directory, select the ***ITHEA.dotm*** template,
- Click *Open*. Check the option *Automatically update document styles* and click *OK*.

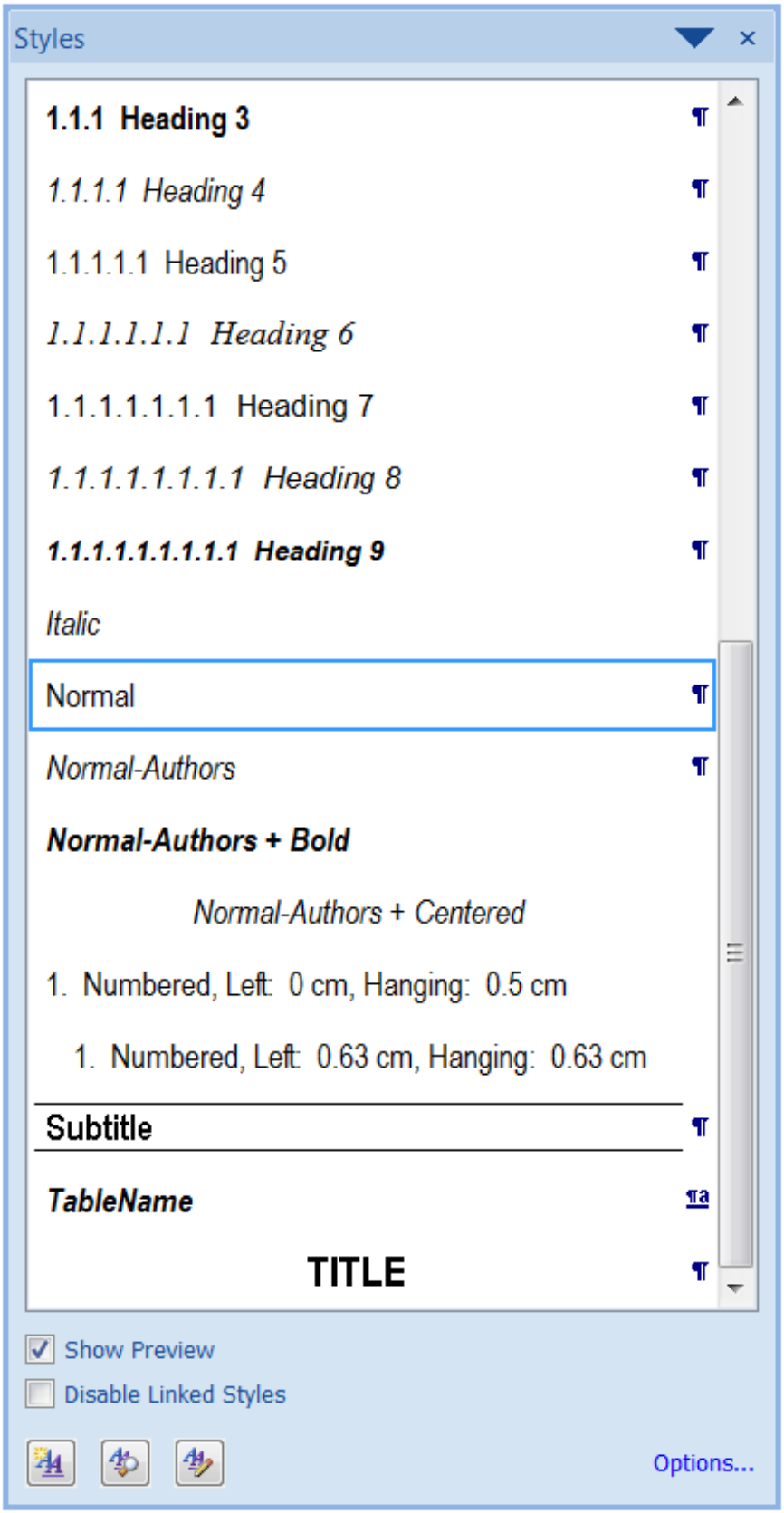


Figure 1. Part of the list of the ITHEA MPS

Main ITHEA formats for manuscripts

Format of the pages is **A4** paper (210 x 297mm).

Margins of the paper sheet are: top - 30mm; bottom, left, right - 25mm.

Plain text has to be formatted by ITHEA MPS ***"Normal"*** style.

"**Normal**" style features: Arial Narrow; 12pt; 1.4-spaced text; 3pt before and after each paragraph; without special indents; left and right justification.

Use the ITHEA MPS style ***"Subtitle"*** for the titles of the separated parts of the article.

Figures and tables should be positioned in the body of the text, as close as possible to the relevant text.

Figures should be properly numbered, centered and should always have a caption positioned under it. Captions should be centered. To caption apply ITHEA MPS ***"Normal"*** style with *Italic* font. Before figure one black space should be left. After caption of table also one black space should be left.

Number manually all figures and tables. Use these numbers to point them in the text. Note that the position of figures and tables may be changed during the assembling the journals or books. Color figures are good for electronic variant but they will be printed in grayscale and some colors may look as equal.

The size of picture should not exceed sixteen centimeters width (16 cm.) and twenty centimeters high (20 cm.). Check it doing the next: wright mouse button click on picture -> "Picture format"->Size. The picture below (Figure 2) is formatted to maximum width - 16 cm.

Figure 2. Example of figures placing, signing, and formatting

Caption of the table should be formatted by ITHEA MPS style ***"Table"***.

"**Table**" style features: Arial Narrow; 12pt; font size 12-point. Spacing before and after should be of 18-point and 3-point, respectively. Captions should be set to justify.

Tables must appear inside the designated page margins. Tables should be properly numbered, and should always have a caption positioned above it. After table one black space should be left.

Titles of the columns should be centered. Text information should be aligned to left, numbers - to right.

Table 1. Two columns table

Title1	Title2
text 1	text 2
number 1	number 2

Please prepare your ***figures*** electronically, and integrate them into your document.

Check that in line drawings, lines are not interrupted and have a constant width. Grids and details within the figures must be clearly readable and may not be written one on top of the other.

Figure resolution should be ***at least 300 dpi.***

Program Code: Program listing or program commands in text should be set in ITHEA MPS ***"Code"***. Example of a Computer Program in C#: Before and after code section left black lines. "**Code**" style features: Arial Narrow; 12pt. Code lines should be justified to left.

```
string s = "456-435-2318";
Regex regex = new Regex(@"\d{3}-\d{3}-\d{4 }");
```

Lists are formulated by ITHEA MPS style "***List***". To mark list points you may use numbers or dashes (-) "**List**" style features: Arial Narrow; 12pt; Spacing between marker (number or dash) and text is 1.4.

Example one:

- One;
- Two;
- Three
- …
- One hundred.

All strings, except the last one, in "example one" should start with capital letter and finished by ";".

Example two:

1. One;
2. Two;
3. Three;
4. …
5. One hundred.

All strings, except the last one, in "example two" should start with capital letter and finished by ";"

Try to avoid nested lists, contained more than three levels of nesting.

Formulas should be positioned in the body of the text, as close as possible to the relevant text.

Put formula and its number in a table row without borders. Align the formula to the center and its number to the right as follow:

$$D = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \tag{1}$$

The MathType INI v1: Equation Preferences are:

[Styles]

Text=Arial
Function=Arial
Variable=Arial,I
LCGreek=Symbol,I
UCGreek=Symbol
Symbol=Symbol
Vector=Times New Roman,B
Number=Times New Roman
User1=Arial Narrow
User2=Arial Narrow
MTExtra=MT Extra

[Sizes]

Full=11 pt
Script=58 %
ScriptScript=42 %
Symbol=150 %
SubSymbol=100 %
User1=75 %
User2=150 %
SmallLargeIncr=1 pt

[Spacing]

LineSpacing=150 %	DenomDepth=100 %	VertRadGap=17 %
MatrixRowSpacing=150 %	FractBarOver=8 %	HorizRadGap=8 %
MatrixColSpacing=100 %	FractBarThick=5 %	RadWidth=100 %
SuperscriptHeight=45 %	SubFractBarThick=2.5 %	EmbellGap=12.5 %
SubscriptDepth=25 %	FractGap=8 %	PrimeHeight=45 %
SubSupGap=8 %	FenceOver=8 %	BoxStrokeThick=5 %
LimHeight=25 %	OperSpacing=100 %	StikeThruThick=5 %
LimDepth=100 %	NonOperSpacing=100 %	MatrixLineThick=5 %
LimLineSpacing=100 %	CharWidth=0 %	RadStrokeThick=5 %
NumerHeight=35 %	MinGap=8 %	HorizFenceGap=10 %

References in the text should be keyed with the name(s) and year of the referred material - for instance [Shannon, 1949].

Put list of **bibliography** after the text of the article using the ITHEA MPS style "**Bibliography**".

Code" style features: Arial Narrow; 12pt, Italic; justified to left; hanging 0.5 cm.

Author's Information: Finish the article with the personal information for every author separately: photo, name of the author, position, organization(s), post and e-mail address(es), major fields of scientific research (keywords). For this information use style "**Normal-Authors**".

Note that the only way to contact the authors is pointed e-mail address in the author's information. Be sure that the addresses are written correctly. If you (or your internet provider) use anti-spam protector write the way to access the e-mail address.

For papers written in Russian, the title, authors, abstract, and keywords in English are obligated. Insert them at the end of paper after authors' information.

Recommended structure of the manuscripts to be published by ITHEA

The papers should contain: Title; Authors; Abstract, Author's keywords and ITHEA keywords; Introduction (objective, used methodology and terminology); Short survey of related papers; Task and challenges; Proposed approach; Case study or implementation of results; Conclusion; Further researches; Acknowledgements; Bibliography; Appendices; Author's information, Annex for. papers written in Russian (for papers written in Russian the title, authors, abstract, and keywords in English are obligated. Insert them at the end of paper after authors' information).

Some comments to the parts of the manuscripts are given below:

- **Title:** Title of the paper should be formatted by the style "**Title**" of ITHEA MPS.
- **Authors:** The name(s) of the author(s),are formatted by the style **"Authors"** of ITHEA MPS. Please, write the whole ***first name and surname*** of the authors.
- **Abstract:** The abstract of the paper is formatted by the style "**Abstract**" of ITHEA MPS. The abstract needs to be from 100 to 350 words long. Note that the abstract is very important for directing the paper to the right reviewers. State the problem, your approach and solution, and the main contributions of the paper. Include little if any background and motivation. Be factual but comprehensive. The material in the abstract should not be repeated later word for word in the paper.
- **Keywords:** Keywords are applied by ITHEA MPS style **"Keyword"** respectively. Papers should contain up to 5 keywords. Keywords are authors designated keywords.

  Also please see ITHEA conference topics given in the Call for papers available at the ITHEA International Conferences page: http://www.ithea.org/conferences/conferences.html, as well as ITHEA journal topics and sub-topics given at the ITHEA International Journals pages:

  - "Information models and analyses": http://www.foibg.com/ijima/ijima-finfo.htm ;
  - "Information theories and applications": http://www.foibg.com/ijita/ijita-finfo.htm ;
  - "Information technologies and knowledge": http://www.foibg.com/ijitk/ijitk-finfo.htm ;
  - "Information content and processing": http://www.foibg.com/ijicp/ijicp-finfo.htm .

— **ITHEA keywords:** Please use keywords from ITHEA Classification Structure (Figure 3), given at: http://idr.ithea.org/tiki-browse_categories.php .

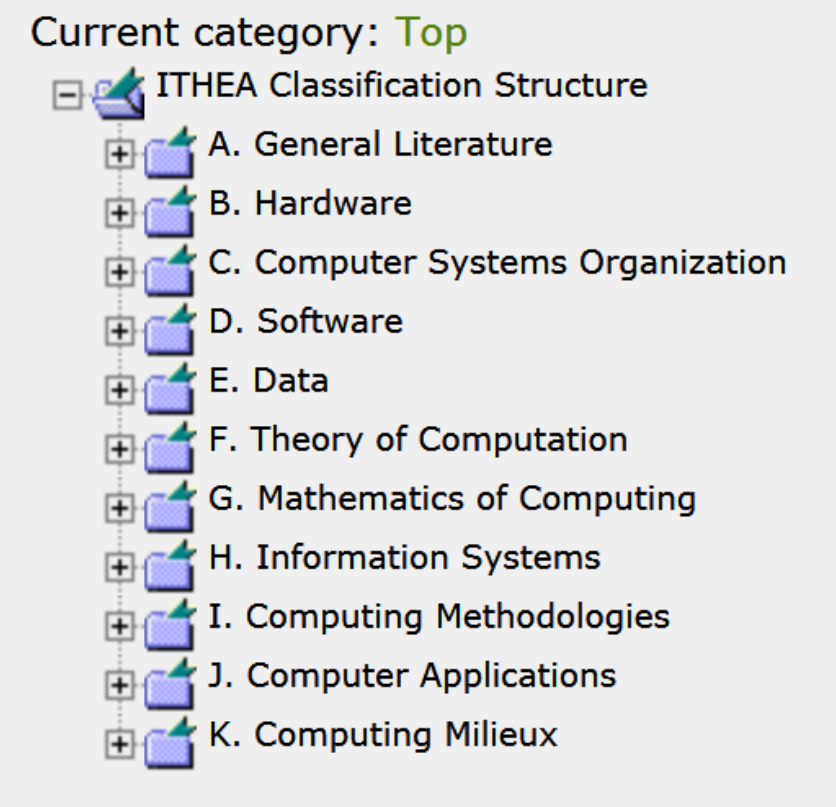


Figure 3. ITHEA Classification Structure (based on ACM Classification System)

— **Introduction:** It is advisable to represent description of research objective, used methodology and terminology. The Introduction is crucially important. Here is the Stanford InfoLab's patented five point structure for Introductions [Widom, 2006]. Unless there's a good argument against it, the Introduction should consist of five paragraphs answering the following five questions: 1. What is the problem? 2. Why is it interesting and important? 3. Why is it hard? (E.g., why do naive approaches fail?) 4. Why hasn't it been solved before? (Or, what's wrong with previous proposed solutions? How does mine differ?) 5. What are the key components of my approach and results? Also include any specific limitations.

Then have a final paragraph or subsection: "Summary of Contributions". It should list the major contributions in bullet form, mentioning in which sections they can be found. This material doubles as an outline of the rest of the paper, saving space and eliminating redundancy.

- **Related work:** Authors are advised to provide wide review, reflecting state of art for latest achievements in investigated area. Describe solutions close to paper topics, represented in leading periodical issues (it may be journals, proceedings, books, etc.). Also it is advisable to explain the place of task, investigated in current paper, in common authors' researches. You may reference to your previous papers to distinguish unanswered questions or aspects needed to be improved.

- **Task and challenges:** Formulate task clearly and laconically. Challenges are ground why the results of your investigation should look like as you formulate in task.

- **The Body:** It is aimed to present the main results, proposed approach, etc. The structure of the body varies a lot depending on content but important components are [Widom, 2006]: (1) *Running Example*: When possible, use a running example throughout the paper. It can be introduced either as a subsection at the end of the Introduction, or its own Section 2 or 3 (depending on Related Work). (2) *Preliminaries*: This section, which follows the Introduction and possibly Related Work and/or Running Example, sets up notation and terminology that is not part of the technical contribution. One important function of this section is to delineate material that's not original but is needed for the paper. (3) *Content*: The meat of the paper includes algorithms, system descriptions, new language constructs, analyses, etc. Whenever possible use a "top down" description: readers should be able to see where the material is going, and they should be able to skip ahead and still get the idea.

- **Case study or implementation of results:** In this section an evaluation of the presented results has to be done. Usually, it is a practical implementation and/or a theoretical analysis. A comparison with leading results in the same scientific area is very important to be given.

- **Conclusion:** Obtained results and comparing them with similar results in the world, and references. In general a short summarizing paragraph will do, and under no circumstances should the paragraph simply repeat material from the Abstract or Introduction. In some cases it's possible to now make the original claims more concrete, e.g., by referring to quantitative performance results [Widom, 2006].

- **Further work:** Interconnect results presented in the paper with general task of your research. This material is important part of the value of a paper is showing how the work sets new research directions.

- **Acknowledgements:** This is the right place to point the Project or any other source for partial support of your publication. Also, acknowledge anyone who contributed in any way: through discussions, feedback on drafts, implementation, etc. If in doubt about whether to include someone, include them. If your paper is supported by ITHEA ISS to be published with reduced fee, please include the obligatory acknowledgement: *"The paper is published with partial support by the ITHEA ISS (www.ithea.org) and the ADUIS (www.aduis.com.ua)"*

- **Bibliography:** Please use the following template for references formatting:

[Family name of the first author, publication year] First author family name, name, Second author family name, name,...,N-th author family name, name. Title of the publication. Name of the journal, proceedings, book. Volume, Issue, Publisher, year, pages, link to resource (if exists).

Examples for referring book, journal, and conference proceeding:

Reference to book:

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

Reference to journal:

[Smith and Brown, 2003] Smith, V., Brown, A., To be or not to be. Journal ABCDEF, Vol.2, Issue 3. Springer-Verlag, 2003. pp. 187-210. ISSN: 0123-0123 (print version), ISSN: 1619-1374 (on-line), doi:00099994342342 http://article-link.com

Reference to proceedings:

[Jackson et al, 2016] Jackson J., Williams D., Davis Y., Software engineering In: i.Tech 2016, Proceedings of the 19th International Conference information technologies. Edited by A. Test, B. Test. ITHEA Sofia, Bulgaria, 2016. ISBN: 123 456 789 012-X, pp. 87-98. DOI xxxx-yyyyyyy, http://proceeding-link.com

- **Appendices:** Appendices should contain detailed proofs and algorithms only. Appendices should not contain any material necessary for understanding the contributions of the paper as well as all material that most readers would not be interested in.

- **Author's information:** This information is very important for contact with all authors to be contacted during the publishing process. Do not forget to check if the e-mail addresses are correct. Photo and Major Fields of Scientific Research are important, too. This information is useful for readers for further collaboration with authors especially during the conferences time.

Conclusion

This exemplar is meant to be a model for manuscript format. Please make your manuscript look as much like this exemplar as possible. In case of serious deviations from the format, the paper will be returned for reformatting.

Bibliography

[Widom, 2006] Widom Jennifer. Tips for Writing Technical Papers. Stanford InfoLab, 2006. https://cs.stanford.edu/people/widom/paper-writing.html

Authors' Information

Krassimir Markov *– Institute of Information Theories and Applications; ITHEA Editor in chief. P.O. Box: 775, Sofia-1090, Bulgaria; e-mail: markov@foibg.com*

Major Fields of Scientific Research: General theoretical information research, Multi-dimensional information systems

Annex for papers written in Russian

ITHEA IJ and IBS Sample Sheet for Preparing the Manuscripts

Krassimir Markov

Abstract*: For papers written in Russian the title, authors, abstract, and keywords in English are obligated. Insert them at the end of paper after authors' information, i.e. just here.*

Keywords*: (Keywords are your own designated keywords).*

TABLE OF CONTENTS