COMPARISON SOFTWARE SYSTEMS BASED ON INFORMATION QUALITY MEASURING

Krassimir Markov, Krassimira Ivanova, Stefan Karastanev

***Abstract**: The usual analysis of experiments using rank-based multiple comparison was discussed in [Ivanova et al, 2016c]. In this paper we will outline another approach. It is based on the comparison of received results with user's information expectation, i.e. on quality of information about the systems received from experiments. All examples in the paper are based on results from real experiments presented in the [Markov et al, 2015].*

***Keywords**: Quality of information, Evaluation of informational services; Rank-based multiple comparison.*

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

Introduction

In the papers [Ivanova et al, 2016a; 2016b, 2016c] a method for comparison software systems was presented. In this paper we outline an extension of the method. It is a multiple comparison based of computing the quality of information received from the experiments. All examples in the paper are based on results from real experiments presented in the [Markov et al, 2015].

The formula for computing quality of information was published in [Markov et al, 1996a, 2006]. In this paper it will be used for ranging software systems. Firstly we will remember the main definitions concerning quality of information given in [Markov et al, 1996a, 2006]. After that we will outline the experiments and ranging based on Friedman test (ANOVA). Finally, ranging based on quality of information will be shown. A comparison of both approaches will be done in the conclusion.

Subjective information expectation and Quality of Information

Every entity which is active in respect to another entity, called "object" of this activity, is called "Subject" [Markov et all, 2006]. The Subject may reflect (temporary or permanently) a certain relationship from the object, i.e. the subject during its interaction with a particular entity (object) might reflect some of its elements and relations between them.

The reflection in the subject’s consciousness which represents a real object is called “***Mental Information Model***” (MIM). The subject can establish a certain relationships between some of the mental information models in his conscious. In this case, the relationships form a set of interrelated MIM.

On the base of the already existing MIM, the subject forms (actively actual) mental model and turns to “**expect**” the connection of the new originated MIM with it.

The orientation towards (the origination of) inside-defined MIM, which depends on the concrete process of information interaction, is called subjective **“information expectation”** (IE). The types of IE were discussed in the paper [Markov et al, 1996b].

The Subject estimates the incoming information depending on the distance to the information expectation.

If the subject couldn’t generate and include in his conscious such a "virtual" information model, we say there is no IE.

Quality of Information

The Subject combines the characteristics of the information expectation with ones of the incoming MIM. The combining the IE with some other MIM is called ***resolving the information expectation***.

Let "n" is the number of the characteristics of an information expectation. Some of them may be combined as well as the others could not. It is clear that “n” is always positive, i.e. n>0. If “n” is a zero then no IE exits.

When a new MIM is generated the Subject evaluates the distance between the IE and MIM. The more this distance is small, the more the IE is better resolved, i.e. satisfied and the incoming MIM is more qualitative.

Quality of the information (**Q**) is evaluated by the distance between the MIM and the IE (inverse proportional of distance between them).

It is proposed to compute the value of quality **Q** by the normalized formula [Markov et al, 1996a, 2006]:

$$\mathbf{Q = 1 / (1 + D),}$$

where **Q** is quality value; **D** is the distance between IE and MIM. The value of **D** depends on the types of information expectation [Markov et al, 1996b] and needs to be computed by corresponding formulas (see [Deza et al, 2012] or [Deza and Deza, 2016]).

For different types of IE we need different formulas for computing the distance R. For different goals of the subject the distance R may be defined as "linear" distance; as distance between corresponding "curves"; as distance between "subspaces"; etc. In this work we assume the simples case where the distance between IE and MIM is Euclidean.

Let remember that the **Euclidean distance** between points **p** and **q** is the length of the line segment connecting them. In **Cartesian coordinates**, if p = (p_1, p_2,..., p_n) and q = (q_1, q_2,..., q_n) are two points in Euclidean n-space, then the distance (d) from p to q, or from q to p is given by the Pythagorean formula:

$$D(p,q) = D(q,p) =$$

$$= \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + (q_3 - p_3)^2 + \ldots + (q_n - p_n)^2} =$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

Respectively, if sets of characteristics of IE and MIM are assumed as Cartesian coordinates, than we will have IE = (e_1, e_2,..., e_n) and MIM = (m_1, m_2,..., m_n) and Pythagorean formula:

$$D(IE, MIM) = D(MIM, IE) =$$

$$= \sqrt{(m_1 - e_1)^2 + (m_2 - e_2)^2 + (m_3 - e_3)^2 + \ldots + (m_n - e_n)^2} =$$

$$= \sqrt{\sum_{i=1}^{n} (m_i - e_i)^2}$$

Experiments

We had compared four real RDF-data storing systems: **R** [RDFArM, 2015], **V** [Virtuoso, 2013], **J** [Jena, 2016], and **S** [Sesame, 2015]. Systems **V**, **J**, and **S** are tested by Berlin SPARQL Bench Mark (BSBM) team and connected to it research groups [Becker, 2008; BSBMv2, 2008; BSBMv3, 2009]. System **R** was tested directly with the same data sets.

The experiments with middle-size RDF-datasets were based on selected real datasets from DBpedia [DBpedia, 2007a; 2007b] and artificial datasets created by BSBM Data Generator [BSBM DG, 2013; Bizer & Schultz, 2009]. The real middle-size RDF-datasets used consist of DBpedia's homepages and geocoordinates datasets with minor corrections [Becker, 2008]:

The artificial middle-size RDF-datasets, generated by BSBM Data Generator [BSBM DG, 2013], are published in N-triple as well as in Turtle format [BSBMv1, 2008; BSBMv2, 2008; BSBMv3, 2009]. We converted Turtle format in N-triple format using “rdf2rdf” program developed by Enrico Minack [Minack, 2010].

We have used four BSBM datasets – 50K, 250K, 1M, and 5M. Details about these datasets are summarized in following Table 1.

Table 1. Details about used artificial middle-size RDF-datasets

Name of RDF-dataset:	B50K	B250K	B1M	B5M
Exact Total Number of Instances:	50,116	250,030	1,000,313	5,000,453
File Size Turtle (unzipped)	14 MB	22 MB	86 MB	1,4 GB

Analysis of experiments: Rank-based multiple comparison

In [Ivanova et al, 2016c] we have presented experiments with middle-size and large RDF data sets, based on selected datasets from DBpedia's homepages and Berlin SPARQL Bench Mark (BSBM). The result from Rank-based multiple comparison is remembered below.

We had used the Friedman test to detect statistically significant differences between the systems [Friedman, 1940]. The Friedman test is a non-parametric test, based on the ranking of the systems on each dataset. It is equivalent of the repeated-measures ANOVA [Fisher, 1973]. We used Average Ranks ranking method, which is a simple ranking method, inspired by Friedman's statistic [Neave & Worthington, 1992]. For each dataset the systems are ordered according to the storing time measures and are assigned ranks accordingly. The best system receives rank 1, the second – 2, etc. If two or more systems have equal value, they receive equal rank which is mean of the virtual positions that had to receive such number of systems if they were ordered consecutively each by other.

Let **n** is the number of observed datasets; **k** is the number of systems.

Let r^i_j be the rank of system j on dataset i. The average rank for each system is calculated as

$$R_j = \frac{1}{n}\sum_{i=1}^{k} r_j^i .$$

The null-hypothesis states that if all the systems are equivalent than their ranks R_j should be equal. When null-hypothesis is rejected, we can proceed with the Nemenyi test [Nemenyi, 1963] which is used when all systems are compared to each other. The performance of two systems is significantly different if the corresponding average ranks differ by at least the critical difference

$$CD = q_{\alpha}\sqrt{\frac{k(k+1)}{6N}}$$

where critical values q_{α} are based on the Studentized range statistic divided by $\sqrt{2}$. Some of the values of q_{α} are given in Table 2 [Demsar, 2006].

Table 2. Critical values for the two-tailed Nemenyi test

quantity of systems	2	3	4	5	6	7	8	9	10
$q_{0.05}$	1.960	2.343	2.569	2.728	2.850	2.949	3.031	3.102	3.164
$q_{0.10}$	1.645	2.052	2.291	2.459	2.589	2.693	2.780	2.855	2.920

The results of the Nemenyi test are shown by means of critical difference diagrams.

Benchmark values from experiments are given in Table 3.

Table 3. Benchmark values

test / system	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
R	3	2272	14.79	3469	60	60	301	136412	1453	5901
S	3	2404	19	2341	179	213	1988	21896	44225	282455
V	2	1327	05	1235	23	25	609	7017	1035	3833
J	5	3557	13	3305	49	41	1053	70851	1013	5654

The ranking of the tested systems is given in Table 4.

Table 4. Ranks of tested systems

test / system	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	average rank
R	2.5	2	3	4	3	3	1	4	3	3	**2.85**
S	2.5	3	4	2	4	4	4	2	4	4	**3.35**
V	1	1	1	1	1	1	2	1	2	1	**1.2**
J	4	4	2	3	2	2	3	3	1	2	**2.6**

All average ranks are different. The null-hypothesis is rejected and we can proceed with the Nemenyi test. Following [Demsar, 2006], we may compute the critical difference by formula:

$$CD = q_{\alpha}\sqrt{\frac{k(k+1)}{6N}}$$

where $q\alpha$ we take as $q_{0.10} = 2.291$ (from Table 1 [Demsar, 2006; Table 5a]); k is the number of systems compared, i.e. k=4; N is the number of datasets used in benchmarks, i.e. N=10. This way we have:

$$CD_{0.10} = 2.291 * \sqrt{\frac{4*5}{6*10}} = 2.291 * \sqrt{\frac{20}{60}} = 2.291 * 0.577 = 1.322$$

This way, we will use for critical difference $CD_{0.10}$ the value **1.322**.

At the end, average ranks of the systems and distance to average rank of the first one are shown in Table 5.

Table 5. Average ranks of systems and distance to average rank of the first one

place	system	average rank	Distance between average rank of the every system and average rank of the first one
1	**V**	1.2	**0**
2	**J**	2.6	**1.4**
3	**R**	2.85	**1.65**
4	**S**	3.35	**2.15**

The visualization of Nemenyi test results for tested systems is shown on Figure 1.

The order of the systems is (1) **V**, (2) **J**, (3) **R**, and (4) **S**.

Analyzing these experiments we may conclude that **R** is at critical distances to **J** and **S**.

R is nearer to **J** than to **S**.

R, **J**, and **S** are significantly different from **V**.

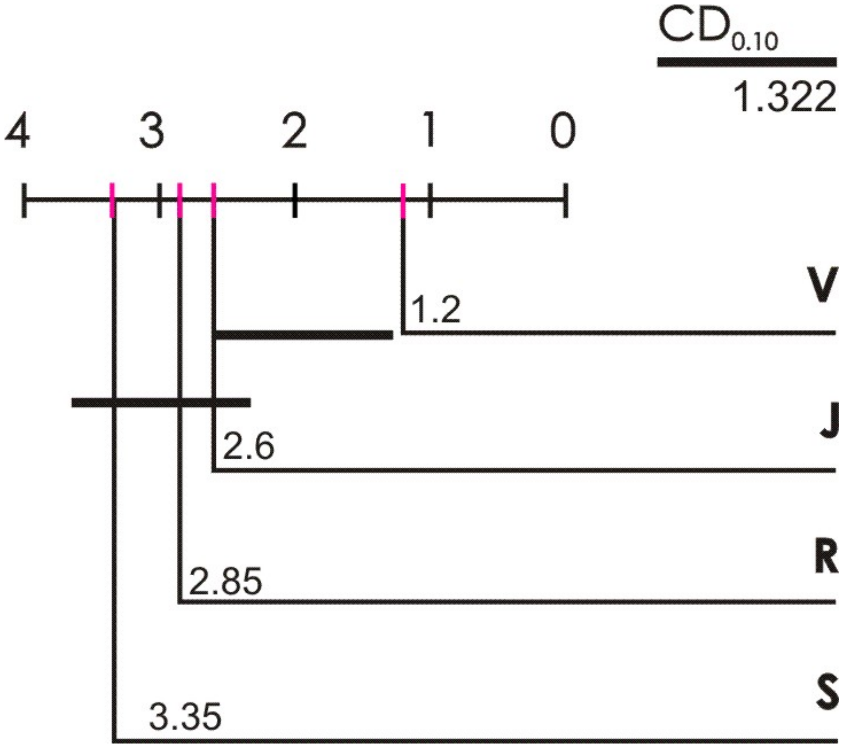

Figure 1. Visualization of Nemenyi test results

Comparison based on of the distance of experiments' information to the information expectation

What is important is that the Friedman test [Friedman, 1940] and ANOVA [Fisher, 1973] conceal the proportions and great differences between received data and this way the ranking does not take in account the distribution of data values. For instance, (see Table 3), in test 9 **S** is 42 times slower than **V**, and in test 10 **S** is 73 times slower than **V**, but in both cases it is on 4 place (see Table 4).

Below we will show another approach based on the distance to IE.

Firstly, we will transform data from Table 1 to be in the interval [0, 1] using transformation formula:

$$X_{new} = 1 - \frac{X_{old}}{MAX\_value\_of\_the\_test}$$

For instance, the results from Test 1 (second column of Table 6) will be transformed by formula:

$$X_{new} = 1 - \frac{X_{old}}{5}$$

because the worst storing time is 5 for the system **J**.

This transformation give us possibility to chose IE = (1, 1,..., 1)

Table 6. Transformed benchmark values and values of IE

MIM \ test	1	2	3	4	5	6	7	8	9	10
IE	1	1	1	1	1	1	1	1	1	1
R	0.4	0.36125949	0.21052632	0	0.66480447	0.71830986	0.84859155	0	0.96714528	0.97910818
S	0.4	0.32414956	0	0.32516575	0	0	0	0.83948626	0	0
V	0.6	0.62693281	0.73684211	0.64398962	0.87150838	0.88262911	0.69366197	0.94856024	0.97659695	0.9864297
J	0	0	0.31578947	0.04727587	0.72625698	0.80751174	0.47032193	0.48061021	0.9770944	0.97998265

As we have pointed in previous sections, if sets of characteristics of IE and MIM are assumed as Cartesian coordinates, than we have IE = (e_1, e_2,..., e_n) and MIM = (m_1, m_2,..., m_n) and Pythagorean formula:

$$D(IE, MIM) =$$

$$= \sqrt{(m_1 - e_1)^2 + (m_2 - e_2)^2 + (m_3 - e_3)^2 + ... + (m_n - e_n)^2} =$$

$$= \sqrt{\sum_{i=1}^{n} (m_i - e_i)^2}$$

Using this formula, we compute distance between IE and MIM of every system (Table 7):

Table 7. Distance between IE and MIM of every system

MIM	Distance between IE and MIM
V	1.855536774
R	2.468421719
J	2.520168491
S	2.987382987

Finally, we compute the quality of information using formula:

$$Q = \frac{1}{1 + D(IE, MIM)}$$

Ranking of the systems based on quality of information for MIM of every system is given in Table 8.

Table 8. Ranking of the systems based on quality of information for MIM of every system

MIM	Q
V	0.350196856
R	0.288315574
J	0.284077311
S	0.250791059

Now we have new order of the systems (1) **V**, (2) **R**, (3) **J**, and (4) **S,** which takes in account data proportions.

The visualization of new results for tested systems is shown on Figure 2. The Critical Distance now is 0.049702899 or rounded off to 0.050 It is computed using formula:

$$CD = \frac{\max_Q - min_Q}{2}$$

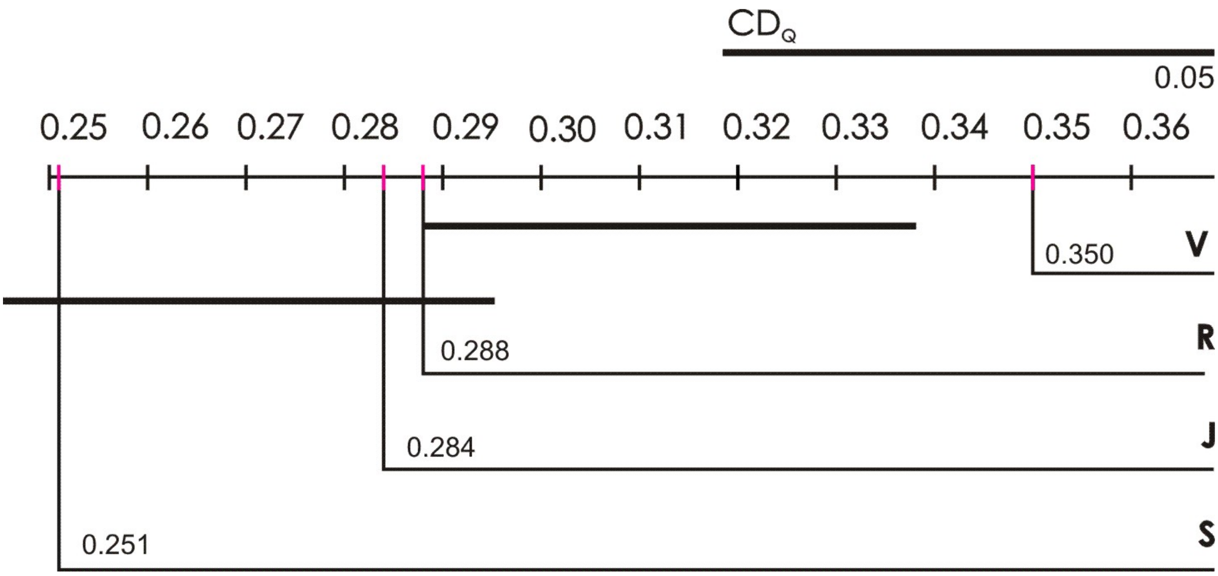

Figure 2. Visualization of quality test results

Analyzing these experiments we may conclude that R is at critical distances to J and S. R is much nearer to J than to S. R, J, and S are significantly different from V.

It is important that **R** and **J** change their places. Now **R** is at the second place.

Conclusion

We have presented results from series of experiments which were needed to estimate the storing time of four systems for middle-size and very large RDF-datasets. Experiments were provided with both real and artificial datasets. Experimental results were systematized in corresponded tables.

The main goal of this work was to propose a new ranking approach based on ***quality of received information***. We have remembered the main theoretical results from [Markov et al, 1996a, 2006] and using examples from real experiments we have shown the new approach is more reliable because it takes in account the distribution of the data values.

Acknowledgement

This paper is published with partial support by the ITHEA ISS (www.ithea.org).

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)

[Bizer & Schultz, 2009] Christian Bizer, Andreas Schultz, "The Berlin SPARQL Benchmark", In: International Journal on Semantic Web & Information Systems, Vol. 5, Issue 2, Pages 1-24, 2009, http://wifo5-03.informatik.uni-mannheim.de/bizer/pub/Bizer-Schultz-Berlin-SPARQL-Benchmark-IJSWIS.pdf; see also http://wifo5-03.informatik.uni-mannheim.de/bizer/berlinsparqlbenchmark/ (accessed: 31.07.2013)

[BSBM DG, 2013] Data Generator and Test Driver, In: Berlin SPARQL Benchmark (BSBM) - Benchmark Rules, http://wifo5-03.informatik.uni-mannheim.de/bizer/berlinsparqlbenchmark/spec/BenchmarkRules/index.html#datagenerator (accessed: 31.07.2013)

[BSBMv1, 2008] Berlin SPARQL Benchmark Results, V1, 2008, http://wifo5-03.informatik.uni-mannheim.de/bizer/berlinsparqlbenchmark/V1/results/index.html (accessed: 31.07.2013)

[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)

[BSBMv3, 2009] Berlin SPARQL Benchmark Results, V3, 2009, http://wifo5-03.informatik.uni-mannheim.de/bizer/berlinsparqlbenchmark/results/V3/index.html (accessed: 31.07.2013)

[DBpedia, 2007a] DBpedia dataset "homepages.nt" dated 2007-08-30, http://wifo5-03.informatik.uni-mannheim.de/benchmarks-200801/H.nt.gz (accessed: 31.07.2013)

[DBpedia, 2007b] DBpedia dataset “geocoordinates.nt” dated 2007-08-30, http://wifo5-03.informatik.uni-mannheim.de/benchmarks-200801/G.nt.gz (accessed: 31.07.2013)

[Demsar, 2006] Demsar J, ”Statistical comparisons of classifiers over multiple data sets”, J. Mach. Learn. Res., 7, 2006, pp. 1 30.

[Deza and Deza, 2016] Michel Marie Deza, Elena Deza. Encyclopedia of Distances. Springer-Verlag Berlin Heidelberg, 2016. eBook ISBN: 978-3-662-52844-0; Hardcover ISBN: 978-3-662-52843-3; DOI: 10.1007/978-3-662-52844-0, Edition: 4, pp: 756

[Deza et al, 2012] Michel Deza, Michel Petitjean, Krassimir Markov. Mathematics of distances. ITHEA® 2012, Sofia, Bulgaria, ITHEA IBS ISC No.: 25. ISBN: 978-954-16-0063-4 (printed), ISBN: 978-954-16-0064-1 (online). 189 pp.

[Fisher, 1973] R. A. Fisher, “Statistical methods and scientific inference”, (3rd edition) Hafner Press, New York, 1973, ISBN 978-002-844740-7.

[Friedman, 1940] Friedman, M., “A comparison of alternative tests of significance for the problem of m rankings”, Annals of Mathematical Statistics, Vol. 11, 1940, pp.86-92

[Ivanova et al, 2016a] Krassimira Ivanova, Emiliya Saranova, Krassimir Markov, Stefan Karastanev, “A Method for Evaluation of Informational Services - Step 1: Computing the Hardware Proportionality Constants”, International Journal "Information Technologies & Knowledge" Volume 10, Number 2, 2016, ISSN 1313-0455 (printed), ISSN 1313-048X (online). pp. 103- 110.

[Ivanova et al, 2016b] Krassimira Ivanova, Ivan Ivanov, Mariyana Dimitrova, Krassimir Markov, Stefan Karastanev, “A Method for Evaluation of Informational Services - Step 2: Computing the Informational Services’ Performance Proportionality Constants”, International Journal "Information Models and Analyses" Volume 5, Number 3, 2016, ISSN 1314-6416 (printed), ISSN 1314-6432 (Online)pp. 203 – 214.

[Ivanova et al, 2016c] Krassimira Ivanova, Krassimir Markov, Stefan Karastanev. A Method for Evaluation of Informational Services - Step 3: Rank-Based Multiple Comparison. International Journal “Information Content and Processing”, Vol.3, Number 4, 2016, ISSN 2367-5128 (printed), ISSN 2367-5152 (online), pp. 303-321. (in print)

[Jena, 2016] Apache Jena, https://jena.apache.org/about_jena/about.html (accessed: 23.02.2016)

[Markov et al, 1996a] K.Markov, K.Ivanova, I.Mitov. Mental Information Measure. IJ ITA, 1996; v.4, n.1; pp. 11-16.

[Markov et al, 1996b] K.Markov, K.Ivanova, I.Mitov. Types of Information Expectation. IJ ITA, 1996; v.4, n.3, pp .21-24.

[Markov et al, 2006] Kr. Markov, Kr. Ivanova, I. Mitov. Basic Structure of the General Information Theory., pp. 19 32. IJ ITA, Vol.14, No.: 1, 2006. pp. 5-19.

[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

[Minack, 2010] Enrico Minack, "RDF2RDF converter", http://www.l3s.de/~minack/rdf2rdf/ 2010, (accessed: 31.07.2013).

[Neave & Worthington, 1992] Neave, H., Worthington, P., "Distribution Free Tests", Routledge, 1992

[Nemenyi, 1963] Peter Nemenyi, "Distribution-free multiple comparisons Unpublished", PhD thesis; Princeton University Princeton, NJ, 1963

[RDFArM, 2015] Krassimira Ivanova. RDFArM - A System For Storing Large Sets Of Rdf Triples And Quadruples By Means Of Natural Language Addressing. International Journal "Information Models and Analyses", Vol. 3, Number 4, 2014, ISSN 1314-6416 (printed), 1314-6432 (online), pp. 303 - 322.

[Sesame, 2015] Sesame, OpenRDF, https://bitbucket.org/openrdf/sesame (accessed: 01.12.2015)

[Virtuoso, 2013] OpenLink Virtuoso Universal Server: Documentation, http://Virtuoso.openlinksw.com/ (accessed: 23.11.2015)

Authors' Information

Krassimir Markov *– Institute of Mathematics and Informatics, BAS; ITHEA Institute of Information Theories and Applications, Bulgaria; e-mail: markov@foibg.com*

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

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*

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.*