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]:

\[ Q = \frac{1}{1 + D}, \]

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 simplest 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:

$$D(p, q) = D(q, p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + (q_3 - p_3)^2 + ... + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

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:

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

$$= \sqrt{(m_1 - e_1)^2 + (m_2 - e_2)^2 + (m_3 - e_3)^2 + ... + (m_n - e_n)^2} = \sqrt{\sum_{i=1}^{n} (m_i - e_i)^2}$$

**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>
<thead>
<tr>
<th>Name of RDF-dataset:</th>
<th>B50K</th>
<th>B250K</th>
<th>B1M</th>
<th>B5M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact Total Number of Instances:</td>
<td>50,116</td>
<td>250,030</td>
<td>1,000,313</td>
<td>5,000,453</td>
</tr>
<tr>
<td>File Size Turtle (unzipped)</td>
<td>14 MB</td>
<td>22 MB</td>
<td>86 MB</td>
<td>1.4 GB</td>
</tr>
</tbody>
</table>

**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 \( r_{ij} \) be the rank of system \( j \) on dataset \( i \). The average rank for each system is calculated as

\[
R_j = \frac{1}{n} \sum_{i=1}^{k} r_{ij}
\]

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

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

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

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

<table>
<thead>
<tr>
<th>test system</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>3</td>
<td>2272</td>
<td>14.79</td>
<td>3469</td>
<td>60</td>
<td>60</td>
<td>301</td>
<td>13641</td>
<td>1453</td>
<td>5901</td>
</tr>
<tr>
<td>S</td>
<td>3</td>
<td>2404</td>
<td>19</td>
<td>2341</td>
<td>179</td>
<td>213</td>
<td>1988</td>
<td>21896</td>
<td>44225</td>
<td>282455</td>
</tr>
<tr>
<td>V</td>
<td>2</td>
<td>1327</td>
<td>05</td>
<td>1235</td>
<td>23</td>
<td>25</td>
<td>609</td>
<td>7017</td>
<td>1035</td>
<td>3833</td>
</tr>
<tr>
<td>J</td>
<td>5</td>
<td>3557</td>
<td>13</td>
<td>3305</td>
<td>49</td>
<td>41</td>
<td>1053</td>
<td>70851</td>
<td>1013</td>
<td>5654</td>
</tr>
</tbody>
</table>

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

**Table 4.** Ranks of tested systems

<table>
<thead>
<tr>
<th>test system</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>average rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>2.5</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2.85</td>
</tr>
<tr>
<td>S</td>
<td>2.5</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3.35</td>
</tr>
<tr>
<td>V</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1.2</td>
</tr>
<tr>
<td>J</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2.6</td>
</tr>
</tbody>
</table>
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:

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

where \( q \) 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:

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

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>
<thead>
<tr>
<th>place</th>
<th>system</th>
<th>average rank</th>
<th>Distance between average rank of the every system and average rank of the first one</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>V</td>
<td>1.2</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>J</td>
<td>2.6</td>
<td>1.4</td>
</tr>
<tr>
<td>3</td>
<td>R</td>
<td>2.85</td>
<td>1.65</td>
</tr>
<tr>
<td>4</td>
<td>S</td>
<td>3.35</td>
<td>2.15</td>
</tr>
</tbody>
</table>

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.
Comparison based on 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 into 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 gives us possibility to choose IE = (1, 1, ..., 1)
Table 6. Transformed benchmark values and values of IE

<table>
<thead>
<tr>
<th>MIM</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>IE</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>R</td>
<td>0.4</td>
<td>0.36125949</td>
<td>0.21052632</td>
<td>0</td>
<td>0.66480447</td>
<td>0.71630986</td>
<td>0.84859155</td>
<td>0</td>
<td>0.96714528</td>
<td>0.97910818</td>
</tr>
<tr>
<td>S</td>
<td>0.4</td>
<td>0.32414956</td>
<td>0</td>
<td>0.32516575</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.83948626</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>V</td>
<td>0.6</td>
<td>0.62693281</td>
<td>0.73684211</td>
<td>0.64398962</td>
<td>0.87150838</td>
<td>0.69366197</td>
<td>0.94856024</td>
<td>0.97659695</td>
<td>0.9864297</td>
<td></td>
</tr>
<tr>
<td>J</td>
<td>0</td>
<td>0</td>
<td>0.31578947</td>
<td>0.04727587</td>
<td>0.72625698</td>
<td>0.80751174</td>
<td>0.47032193</td>
<td>0.48061021</td>
<td>0.9770944</td>
<td>0.97998265</td>
</tr>
</tbody>
</table>

As we have pointed in previous sections, if sets of characteristics of IE and MIM are assumed as Cartesian coordinates, than we have IE = (e₁, e₂,..., eₙ) and MIM = (m₁, m₂,..., mₙ) and Pythagorean formula:

\[ D(IE, MIM) = \sqrt{(m₁ - e₁)^2 + (m₂ - e₂)^2 + (m₃ - e₃)^2 + ... + (mₙ - eₙ)^2} = \sqrt{\sum_{i=1}^{n}(m_i - e_i)^2} \]

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

<table>
<thead>
<tr>
<th>MIM</th>
<th>Distance between IE and MIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>1.855536774</td>
</tr>
<tr>
<td>R</td>
<td>2.468421719</td>
</tr>
<tr>
<td>J</td>
<td>2.520168491</td>
</tr>
<tr>
<td>S</td>
<td>2.987382987</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>MIM</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>0.350196856</td>
</tr>
<tr>
<td>R</td>
<td>0.288315574</td>
</tr>
<tr>
<td>J</td>
<td>0.284077311</td>
</tr>
<tr>
<td>S</td>
<td>0.250791059</td>
</tr>
</tbody>
</table>

Now we have new order of the systems (1) V, (2) R, (3) J, and (4) S, which takes into 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:

$$CD = \frac{\max_q - \min_q}{2}$$

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


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: kras78@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.