ISSN 0005-1055, Automatic Documentation and Mathematical Linguistics, 2011, Vol. 45, No. 5, pp. 218–223. © Allerton Press, Inc., 2011.
Original Russian Text © E.S. Klyshinskii, 2011, published in Nauchno-Tekhnicheskaya Informatsiya, Seriya 2, 2011, No. 9, pp. 6–11.

Analysis of Complex Measures for the Subject Similarity of Documents

E. S. Klyshinskii

Moscow Institute of Electronics and Mathematics, Moscow, Russia

e-mail: klyshinsky@itas.miem.edu.ru

Received March 17, 2011

Abstract—The author proposes a method for document retrieval with the use of a document pattern. For this purpose, a complex measure for assessing the subject similarity of documents is introduced. A numerical evaluation of the results of a search output for different measures is given.

Keywords: documentary database, selection of keywords, significant combination, measure of document similarity, subject similarity of document.

DOI: 10.3103/S0005105511050025

INTRODUCTION

With the advent of highly-developed computing machinery, all current large-scale enterprises turned to electronic documentary system, which can be applied in parallel with the traditional accounting for paper documents. The use of electronic documents allows one to create a large archive containing the information of an enterprise's activity over a long period. A thus-developed documentary database is of great importance in terms of experience storage. A well-assigned system of document circulation stores all documents describing the process of product development, as well as the history of the advantages and disadvantages of such an implementation.

When approaching the design of a new product, developers analyze the relevant solutions. At a large-scale enterprise with a long history, the internal archive should be monitored for the information before searching for other solutions. It is common practice to apply search engines that allow one to store the information in a convenient form and to present the retrieval results. However, the vast majority of engines process queries in the form of several keywords, by which the retrieval is carrying out.

Such an approach results in a series of problems. When entering several words, we cannot guarantee that it is precisely these words that will be found in the documentary database. After a lapse of years from the time of the design of a project, the accepted vocabulary could be changed and new concepts with appropriate terms could replace out-of-use ones. A developer cannot be sure that he will enter precisely the words that were required in the original documentation due to synonyms and similar terms within it. The extension of the list of entered keywords may not lead to the desired result, since wrong words will be used constantly. This is why information retrieval is like an art and it requires special skills to use the right terms during such retrieval.

Although keywords will be different, original documents can contain many similar words. The key vocabulary can turn out to be changed, but the rest of it persists. Moreover, detached words can have multiple meanings with variants in different object domains. In this case, multiword structures are highly stable. Taking them as the markers of special vocabulary, we can improve the retrieval results. However, highlighting multiword structures leads to the same problems.

Notice that the base of the design is a technical statement. Such a document should contain a short but full description of the formulation and solution of the problem. Thus, the technical statement must include a full set of terms that describe the considered task. Using separate terms in a given document results in the above-mentioned problems. In this connection, the task of generating a retrieval request based on a short description in the form of a formal document is required. As this occurs, the vocabulary needs to be maximally conserved, when comparing the entire retrieval patterns instead of the use of typical queries to the retrieval system.

THE CURRENT SOLUTIONS

In such or similar setting, the problem was formulated rather a long time ago. At the present moment, the extraction of a given number of keywords or phrases (from several units to several hundreds), which is further used for the retrieval of similar documents, is a widespread approach. The selection of keywords can be performed in terms of different methods. The most

used measure is tf*idf (or its modification) that shows the "information value" of a term. Terms are discarded if they are met in too large or too small a number of documents. Frequently occurring terms are considered to be less informative, since these are most likely to be stylistic elements. Rarely occurring terms do not connect significant numbers of documents, and are likely to be only noise. Both statements are rather controversial. So the absence of frequently occurring terms in a document from a single subject collection can be a distinctive feature of a document for others. In contrast, the presence of rarely occurring words shows the difference between documents in terms of vocabulary distinctions. However, the problem of determining a degree of document similarity instead of quality separation of documents by the object domains is at issue. Therefore, such arguments are usually ignored, since they lead to noise amplification when defining similarity measures.

For word combinations detecting, the collocations method can be used (e.g. MI, t-score or log-score measures [3]) as well as tf*idf and frequency methods. On the one hand, measures that are developed for phrase extraction allow one to obtain their numerical evaluation, based on which the selection of the most significant combinations can be conducted. On the other, the values of measures for different documents are not comparable with each other, since their calculations are based on occurrence frequencies and document lengths. In spite of this, various measures make it possible to extract vocabularies of various types [4]. Therefore, a t-score extracts the stylistic features of a text, while the MI extracts the vocabulary of the object domain. In this case, the t-score can be used for the detection of search terms, instead of MI used for the definition of the measure of document similarity.

In addition, removal for stop words (prepositions, conjunctions, pronouns, and so on) is carried out in documents, since these words are senseless in isolation from their context. As this occurs, the definition of phrases should be conducted before the search, because the phrases can include the words from a given list.

Selected words can be variously used. Therefore, extracted words can be delivered as a query to an existing retrieval engine. Thus, the process of the selection of analogous documents via the expertise of demands and the assessment of their novelty is automated [5]. It should be mentioned that a request can be transmitted by both one of the global retrieval engines, including Yandex and Google, which are meant for document searching on the Internet, and by a local retrieval engine, which is a part of a documentary system. Here, the simplest approach (but not the most effective one) is the statistical selection of detached words by their frequency of occurrence in a document [6].

The second class of solutions includes the methods for determining the subject similarity of documents. In this case, a vector of document features is formed. It contains a list of the most significant terms. Further comparison is performed with the use of different measures of the similarity of specified methods.

As of now, an entire set of measures exist that use small fragments of documents for comparison with the last ones. Thus, the group of algorithms that are based on the Rabin–Karp random polynomial algorithm [7] depends upon the detection of similar groups of sequential words of the k length (shingles) or dactylograms, which are substrings of documents of a fixed length. After the calculation of the fixed number of shingles and dactylograms, such sequences are compared. For faster methods, hash functions are applied to obtained sequences. Documents are considered to be similar if the obtained values of the hash function coincide. Slower but exact methods are compared with these sequences. Then, on the basis of the extent of their coincidence, one can draw a conclusion about the extent and the probability of document similarity. A detailed survey of both of these methods and of others can be found in [8]. Apart from word sequences, the longest and the most informative assumptions in terms of the tf*idf measure can be taken.

To determine documents similarity, cosine measure and Dice coefficient are commonly used [9]. They are introduced as follows. There's a feature vector $\mathbf{w} = \langle w_i \rangle$ of words used to determine the documents similarity, where w_i is a word in a given language. $\mathbf{f} = \{f_i\}$ is a vector of frequencies of occurrence of those words and f_i is a frequency value for word w_i in a given document. $\mathbf{f}_x$ is a term frequency vector for document $\mathbf{x}$. Then the measure of co-occurrence of words from documents $\mathbf{x}$ and $\mathbf{y}$ is represented as a dot product $|\mathbf{xy}|=\mathbf{f}_x \cdot \mathbf{f}_y$. The measure of words occurrence in document x is defined as a dot product of square frequency vector: $|\mathbf{x}|=\mathbf{f}_x \cdot \mathbf{f}_x$.

In this case, the Dice measure is found as twice the ratio of the word co-occurrence in the $\mathbf{x}$ and $\mathbf{y}$ documents to the sum of the measures of word occurrence in these documents.

$$\mathrm{Dice}(\mathbf{x}, \mathbf{y}) = 2^* \frac{|\mathbf{xy}|}{|\mathbf{x}| + |\mathbf{y}|}. \quad (1)$$

The cosine measure is defined as the ratio of the word co-occurrence in the $\mathbf{x}$ and $\mathbf{y}$ documents to the square root of the product of measures of word occurrence in these documents.

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{|\mathbf{xy}|}{\sqrt{|\mathbf{x}| \cdot |\mathbf{y}|}}. \quad (2)$$

In both measures, the value 1 denotes that the documents are the same and value 0 denotes the absence of any common words.

Since the documents belong to some collection, the vocabulary of all documents within the collection can be used as a measure of similarity. In case of large documentary base, the vocabulary is equal to whole

domain lexis. That is why one should use the reduced vocabulary that contains the most valuable terms from the total vocabulary (e.g. [8]). Ilyinsky et al. [10] introduce the method that uses a descriptive set containing N terms. The value of 1 is recorded for the vector of features if the frequency of a given word in the document exceeds some value; a zero value is otherwise recorded. Then, documents with coinciding vectors of features are considered to be original. The value of the threshold is selected so that the minimum numbers of documents have an analog value of tf.

Instead of the detached words in these measures, combinations of different lengths can be used [11]. However, this does not always result in increasing the relevance of delivered documents, as shown below. Moreover, in different cases both normal word forms, which result from morphological analysis, and invariable word forms can be applied for retrieval. However, the last variant is rarely used.

PROBLEM STATEMENT

In the above-mentioned examples, the relative values of word occurrence in documents are applied. But such an approach is not always correct. Upon searching the measure of fuzzy similarity of documents, which is used, e.g., in antiplagiarism problems, a similar value will indicate the extent of changes that entered the original document. For searching thematically relative documents, a simple measure of vocabulary similarity can be used. Let us have two documents. As it happens, 30% of the first one is devoted to hardware and 70% is devoted to software; 70% of the second document is conversely devoted to hardware and 30% to software. In this case, the documents will be significantly different (the cosine measure will give a value of 0.72) in terms of the antiplagiarism system, while both documents belong to a single subject and we can find the second document using the first one.

In this connection, there is a need to design new methods for assessing the subject similarity of documents based on the current methods. This method will use information from the vocabulary of a given document without significant influence of the relative frequency of word occurrence. In addition, since multiword terms can characterize the object domain to a greater extent, they should be used for the subject similarity of documents.

RETRIEVAL STRATEGY

Unless the relative word occurrence is used, we can do the following; 1 can replace nonzero values of frequency and 0 values can be constant. In this case, words that occur in the text once will act as words that occur in the text many times. Because of this, the threshold of the cutoff is introduced. Zeros are ascribed to all words with an occurrence that is below the cutoff threshold, whereas other words are assigned to the unit.

$$x_i' = \begin{cases} 1, & \text{at } x_i \geq \varepsilon \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

Note that the cutoff threshold can be taken in both relative and in absolute units.

Within such approach modified Dice coefficient and cosine are calculated in the same way as traditional ones. In practice, the number of coinciding words can be counted instead of the dot product and divided into the root of the product of the number of unique words in each document in the case of the cosine measure and into the sum of similar values in the case of the Dice measure. In addition, the classic approach assumes that the vector length of features for each document is the same and equal, as an example, to the total volume of the vocabulary of all documents. Such a method is used to prevent the retrieval of words from one document in the feature vector of another. However, if we store the identificators of the initial form instead, the problem of comparing two vectors will become trivial. In this case, the vocabulary can be stored for each document, which significantly reduces memory consumption and increases the productivity. We find a rate that is proportional to the sum of the dictionaries of two documents instead of the calculation rate that is proportional to the volume of the common vocabulary for the collection (i.e., properly to the volume of the morphological dictionary). For short document, using such an approach, we can improve by many or 10–100 times (compared to methods that don't use reduction of the space of features). Furthermore, the algorithm leads to saving the most significant words that are included in the space of features. Note that there should be no difference between the rate of comparison against the methods that use the reduction of the space of features, although we save the calculation algorithm and the reduction of the dictionary that is consulted for the addition of each document.

All above-mentioned assumptions can be applied equally to word combinations, i.e., w_i is not a detached word but a word combination that is selected according to specified features.

For the sake of convenience, let us denote the modified cosine similarity measure (*simplified cosine*) as s_cos(**x**, **y**, *n*). This measure is calculated for *n*-grams without regard for the frequencies of their occurrence. The modified Dice measure is denoted as s_Dice(**x**, **y**, *n*). Let $\|\mathbf{xy}\|_n$ be the number of a word combinations with length $n > 0$, contained in documents **x** and **y**. Let $\|\mathbf{x}\|_n$ be the number of collected word combinations

with length n in the **x** document. Then, based on (1) and (2) we find

$$\text{s\_cos}(\mathbf{x}, \mathbf{y}, n) = \frac{\|\mathbf{xy}\|_n}{\sqrt{\|\mathbf{x}\|_n \cdot \|\mathbf{y}\|_n}} \quad (4)$$

and

$$\text{s\_Dice}(\mathbf{x}, \mathbf{y}, n) = 2^* \frac{\|\mathbf{xy}\|_n}{\|\mathbf{x}\|_n + \|\mathbf{y}\|_n}. \quad (5)$$

Let us denote two measures in addition to the cosine measure and the Dice measure, as well as their modifications. The first is the symmetric simplified measure of similarity.

$$\text{NSL}(\mathbf{x}, \mathbf{y}, n) = \frac{\|\mathbf{xy}\|_n}{\|\mathbf{x}\|_n}. \quad (6)$$

Such a measure will represent the relative volume of similar vocabulary of the **x** and **y** documents in the **x** document. Obviously, such a measure will be single-ended, i.e., $\text{NSL}(\mathbf{x}, \mathbf{y}, n) \neq \text{NSL}(\mathbf{y}, \mathbf{x}, n)$. To correct such a deficiency, one should sum both measures, which results in a simplified measure of similarity.

$$\text{SSL}(\mathbf{x}, \mathbf{y}, n) = \frac{\|\mathbf{xy}\|_n}{\|\mathbf{x}\|_n} + \frac{\|\mathbf{xy}\|_n}{\|\mathbf{y}\|_n}. \quad (7)$$

By using the simple arithmetical arguments from (7) we find that

$$\text{SSL}(\mathbf{x}, \mathbf{y}, n) = \|\mathbf{xy}\|_n \cdot \frac{\|\mathbf{x}\|_n + \|\mathbf{y}\|_n}{\|\mathbf{x}\|_n \cdot \|\mathbf{y}\|_n}. \quad (8)$$

Since the right member of the product in (8) shows the probability of crossing the lists, the total measure represents the probability of finding $\|\mathbf{xy}\|_n$ combinations in both lists (except the normalizing factor 1/2).

METHOD OF EVALUATION

In order to define the operation quality of the above measure and other measures, a set of numerical tests was carried out. A collection of 450 documents on different subjects was gathered for the test. Scientific articles, dissertations, and abstracts of dissertations from different branches of science were taken as the base of the collection. To put some information noise into the collection, several tens of the works of T. Pratchett were added. All documents were divided into clusters. In some cases, the division was already given by the location of the abstract in the rubricator of the Higher Attestation Commission of the Russian Federation [12]. An expert clustered the remaining part of the documents by hand according to his views on the document subject and the similarity between documents. As a result, 26 clusters of different volumes were obtained. Several clusters included only one document for verifying the minimum of the output of such document in response to the request. Some clusters were marked like the original ones.

Table 1

Measure	Completeness	Evaluation
Cos(n =1)	0.402	–0.07
Cos(n =2)	0.201	–0.97
Cos(n =3)	0.278	–2.97
s_cos (n =1)	0.479	2.3
s_cos (n =2)	0.354	–0.3
s_cos (n =3)	0.111	–6.17
Dice(n = 1)	0.396	–0.3
Dice(n = 2)	0.34	–1.07
Dice(n = 3)	0.25	–3.8
s_Dice(n = 1)	0.451	1.67
s_Dice(n = 2)	0.472	1.67
s_Dice(n = 3)	0.389	–0.43
SSL(n = 1)	0.389	–1.23
SSL(n = 2)	0.486	1.23
SSL(n = 3)	0.431	0.96

The experiment was carried out as follows. A list of documents that were received as the request to the system was formed in advance. For each document, the output contained the ten most relevant documents according to the chosen measure. Then, two following evaluations were performed on the obtained output: the percentage in the output of documents from the same cluster and the complex evaluation. The complex evaluation was added to 1 for each output document of the same cluster and to 0.5 for the documents of similar clusters. For the rest of the documents from the complex measure, 1 was subtracted. Further, the complex measure was divided into the numbers of requested documents. When calculating the percentage of documents, the size of the output was equal to ten (under the numbers of documents in a cluster that is ten) if it was not equal to the size of the cluster.

Thus, the percentage of relevant documents shows the completeness of the output and the complex evaluation shows the average relevance of the documents in the output. Such an approach bears some resemblance to the evaluation technique of the results that was presented in the Russian Information Retrieval Evaluation Seminar [13], although there are some differences between them.

For verification, the results were tested for the random selection of a file. A file was selected from the entire collection and sent as the request to the system. The described method served as an estimate for this problem. Then, the best results out of ten were selected to exclude the effect of files that did not contain relevant documents.

Table 2

Measure	Completeness	Evaluation
s_cos(1) + SSL(2)	0.583	3.8
s_cos(1) + SSL(2) + SSL(3)	0.569	3.63
s_cos(1) + SSL(1) + SSL(2)	0.569	2.93
s_cos(1) * SSL(2)	0.549	2.43
s_cos(1) * SSL(2) * SSL(3)	0.444	0.63
s_cos(1) * NSL(**x**, **y**, 1) * NSL(**y**, **x**, 1) * NSL(**x**, **y**, 2) * NSL(**y**, **x**, 2)	0.569	2.93
s_cos(1) + s_Dice(1) + s_Dice(2) + SSL(2)	0.549	2.73
s_cos(1) + s_Dice(1) + s_Dice(2) + SSL(2) +SSL(3)	0.549	3.03
SSL(2) +SSL(3)	0.500	2.1

Table 3

Measure	Completeness	Evaluation
s_cos(1) + SSL(2)	0.518	2.63
s_cos(1) + SSL(2) + SSL(3)	0.543	3
s_cos(1) + SSL(1) + SSL(2)	0.482	2.2
s_cos(1) * SSL(2)	0.513	2.8
s_cos(1) * SSL(2) * SSL(3)	0.503	2.53
s_cos(1) * NSL(**x**, **y**, 1) * NSL(**y**, **x**, 1) * NSL(**x**, **y**, 2) * NSL(**y**, **x**, 2)	0.498	1.7
s_cos(1) + s_Dice(1) + s_Dice(2) + SSL(2)	0.508	2.07
s_cos(1) + s_Dice(1) + s_Dice(2) + SSL(2) +SSL(3)	0.487	1.43
SSL(2) +SSL(3)	0.472	0.83

EXPERIMENTAL RESULTS

Using the described method, the evaluations for measures (1), (2), (4), and (8) were calculated at different values of n. Table 1 gives the experimental results. By the "completeness" we mean the percentage of documents in the output; by the "evaluation," we denote the complex measure of the output.

It is seen from the data that the modified measure shows results that exceed the typical cosine measure.[1] The same is related to Dices' measure. It is easy to verify that the growth of n impairs the cosine measure and the Dice measure, i.e., the relevance of required documents is reduced. Thus, the modified cosine measure can be used instead of classic one at $n = 1$. The modified Dice measure varies at the n change from 1 to 3. However, at any values of the parameter it gives the best results on the problem of document retrieval with the use of another document as the request. Finally, the SSL measure does not give worse results in comparison with the cosine measure and the relevance of the response increased by 15–20% at $n = 2,3$.

[1] Note that searching for word combinations from the feature vector was not used in this method. The verification of the effect of feature selection on the results of calculating different measures will be performed on its own.

However, the combination of different measures gave the best results. Therefore, the use of combinations

$$\text{s\_cos}(\mathbf{x}, \mathbf{y}, 1) + \text{SSL}(\mathbf{x}, \mathbf{y}, 2) + \text{SSL}(\mathbf{x}, \mathbf{y}, 3), \quad (9)$$

SSL(**x**, **y**, *2*) * SSL(**x**, **y**, *3*) or s_cos(**x**, **y**, *1*) * SSL(**x**, **y**, *2*) * SSL(**x**, **y**, *3*) allows one to improve the relevance in comparison with the cos(**x**, **y**, *1*) measure by more than 30%. The use of combinations with the s_Dice measure makes it possible to attain approximately the same but slightly smaller effect. Table 2 presents the results for these measures and the others. Notice that the table gives the values for the measures that showed better results in the course of the performed experiments with these combinations and the others.

After changing the numbers of documents in the output to 15, the evaluations were reduced, as expected.

Table 3 shows that all the additive criteria impaired the evaluations of values, while the multiplicative criterion s_cos(1) * SSL(2) * SSL(3) improved the results. The same situation can be explained as follows: the additive criteria mostly place the relevant documents at the head of the output, while these documents tend to be centrally located in the s_cos(1) * SSL(2) * SSL(3) criterion.

The output by random files were estimated for the following measures: s_cos, SSL(**x**, **y**, *2*) + SSL(**x**, **y**, *3*), s_cos(**x**, **y**, *1*) + SSL(**x**, **y**, *2*), and s_cos(**x**, **y**, *1*) + SSL(**x**, **y**, *2*) + SSL(**x**, **y**, *3*). For the cosine measure, the values of completeness, which are from 0.333 to 0.523 (the average value is 0.408), and the values of complex measure, which are from −3.2 to 0 (the average value is −1.75), were obtained. For the SSL(**x**, **y**, *2*) + SSL(**x**, **y**, *3*) measure, the values of completeness, which are from 0.353 to 0.606 (the average value is 0.468), and the values of complex measure, which are from −2.85 to 1.1 (the average value is −1.89), were obtained. For the s_cos(**x**, **y**, *1*) + SSL(**x**, **y**, *2*) measure, the values of completeness, which are from 0.434 to 0.666 (the average value is 0.55), and the values of complex measure, which are from −1.5 to 4.75 (the average value is 2.14), were obtained. Finally, for the s_cos(**x**, **y**, *1*) + SSL(**x**, **y**, *2*) + SSL(**x**, **y**, *3*) measure, the values of completeness, which are from 0.512 to 0.762 (the average value is 0.628), and the values of complex measure, which are from 0.25 to 4.5 (the average value is 2.14), were obtained.

Thus, measure (9) also gives the best results under random selection, while the others depend largely on the files of the request.

CONCLUSIONS

According to the experimental results, the proposed measure for the determination of subject similarity leads to a significant increase in the relevance of the response in information retrieval problems in the documentary databases. We can also expect analog improvement in the problems of clustering and classification of documents.

It should be mentioned that the complex measure will not work for antiplagiarism problems since it does not take the ratio of the frequencies of term occurrence in a text into account.

Currently, there are no papers that are devoted to the reduction of the space of document features. The most prospective method is to search the style component of documents using the t-score measure; combinations with a high evaluation by the MI method are favored.

ACKNOWLEDGMENTS

This work was supported by the Federal targeted program "Scientific and Scientific-Pedagogical Staff for Innovative Russia for 2009–2013".

REFERENCES

1. Salton, G. and Buckley, C., Term-Weighting Approaches in Automatic Text Retrieval, *Int. J. Inf. Process. Manag.,* 1988, vol. 24, no. 5, pp. 513–523.
2. Robertson, S.E., Understanding Inverse Document Frequency: On Theoretical Arguments for IDF, *J. Doc.*, 2004, vol. 60, no. 5, pp. 503–520.
3. Zakharov, V.P. and Khokhlova, M.V., Study of Effectiveness of Statistical Measures for Collocation Extraction on Russian Texts, *Trudy ezhegodnoi mezhdunarodnoi konferentsii "Dialog"* (Proc. Annual Int. Conference "Dialogue"), *J. Comput. Ling. Intel. Technol.*, 2010, vol. 9, no. 16, pp. 137–143.
4. Yagunova, E.V. and Pivovarova, L.M., From Collocations to Constructions, *Trudy vtoroi vserossiiskoi konferentsii "Russkii yazyk: konstruktivnye i leksiko-semanticheskie podkhody"* (Proc. 2nd All-Russian Confence "Constructional and Lexical Semantic Approaches to Russian"), Sai, S.S., Ed., St. Petersburg, 2011.
5. Makarov, S.L., Methods for Operation Quality Improvement of Information Retrieval Systems, *J. Qual. Innov. Educ.*, 2008, no. 1, pp. 35–39.
6. Dubinskii, A.G., Application of Vector Model of Document Representation in Information Retrieval, *J. Contr. Syst. Comput.,* 2001, no. 4.
7. Gusfield, D. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK. 1997.
8. Zelenkov, U.G. and Segalovich, I.V., Comparative Analysis of Near-Duplicate Detection Methods of Web documents, *Vseros. nauch. konf. "Electronnye biblioteki: perpektivnye metody i tekhnologii, electronnye kollectsii"* (All-Russian Science Conf. "Electronic Libraries: Prospect Methods and Technologies, Electronic Collections"), Pereslavl-Zalessky: 2007, vol. 1, pp. 166–174.
9. Manning, C.D., Raghavan, and P., Schutze, H., *Introduction to Information Retrieval*, Cambridge University, Press, 2008.
10. Ilyinsky, S., Kuzmin, M., Melkov, A., and Segalovich, I., An Efficient Method to Detect Duplicates of Web Documents with the Use of Inverted Index, *Proc. 11th Int. Conf. on World Wide Web*, 2002.
11. Milios, E., et al., Automatic Term Extraction and Document Similarity in Special Text Corpora, *Proc. 6th Conference of the Pacific Association for Computational Linguistics*, 2003, pp. 275–284.
12. Announcements of Defence of Doctoral Dissertations, Higher Attestation Comission. http://vak.ed.gov.ru/ru/dissertation. Cited February 20, 2011.
13. Ageev, M., Kuralenok, I., and Nekrest'yanov, I., Official RIRES Metrics, *Trudy ROMIP 2010* (Proc. RIRES 2010), Kazan, 2010.