An Approach to Clustering Abstracts*

Mikhail Alexandrov[1,2], Alexander Gelbukh[1], and Paolo Rosso[2]

[1] Center for Computing Research, National Polytechnic Institute, Mexico
dyner1950@mail.ru, gelbukh@gelbukh.com
www.gelbukh.com
[2] Polytechnic University of Valencia, Spain
prosso@dsic.upv.es

Abstract. Free access to full-text scientific papers in major digital libraries and other web repositories is limited to only their abstracts consisting of no more than several dozens of words. Current keyword-based techniques allow for clustering such type of short texts only when the data set is multi-category, e.g., some documents are devoted to sport, others to medicine, others to politics, etc. However, they fail on narrow domain-oriented libraries, e.g., those containing all documents only on physics, or all on geology, or all on computational linguistics, etc. Nevertheless, just such data sets are the most frequent and most interesting ones. We propose simple procedure to cluster abstracts, which consists in grouping keywords and using more adequate document similarity measure. We use Stein's MajorClust method for clustering both keywords and documents. We illustrate our approach on the texts from the Proceedings of a narrow-topic conference. Limitations of our approach are also discussed. Our preliminary experiments show that abstracts cannot be clustered with the same quality as full texts, though the achieved quality is adequate for many applications; accordingly, we suggest Makagonov's proposal that digital libraries should provide document images of full texts of the papers (and not only abstracts) for open access via Internet, in order to help in search, classification, clustering, selection, and proper referencing of the papers.

1 Introduction

1.1 Difficulties in Clustering Short Documents

In Information Retrieval, clustering algorithms are used to analyze large collections of documents by means of subdividing them into groups of similar documents. A typical approach to document clustering in a given domain is to transform the textual documents into vector form basing on a list of index keywords and then use well-known numerical procedures of cluster analysis [13]. The list of keywords is constructed from a training document set belonging to the same domain. If the domain is unknown then this list is constructed directly from the document collection itself. The keyword list used for document representation is weighted using, for example, the well-known *tf-idf* technique [15].

* Work done under partial support of the Government of Valencia, Mexican Government (CONACyT, SNI, CGPI, COFAA-IPN), R2D2 CICYT (TIC2003-07158-C04-03), and ICT EU-India (ALA/95/23/2003/077-054).

A. Montoyo et al. (Eds.): NLDB 2005, LNCS 3513, pp. 275–285, 2005.

Currently such an approach is used for clustering not only full-text documents, but also for clustering short documents containing 50–100 words – as, for example, news, brief historical or advertising information, etc. However, the fact that good results have been obtained in these particular cases is not a reason for optimism: this approach gives very *unstable* or *imprecise* results when clustering abstracts of scientific papers, technical reports, patents, etc. Nevertheless, just these cases are the most interesting ones: most digital libraries and other web-based repositories of scientific and technical information nowadays provide free access only to abstracts and not to the full texts of the documents. Therefore, the results prove to greatly depend on the type of short documents being clustered. Let us consider the following two different cases:

1. Document collection containing the documents that belong to essentially different domains, such as sport, culture, politics, etc.
2. Document collection containing the documents from one narrow domain, such as physics, linguistics, urbanistics, etc.

This partition of document collections may seem very subjective. For example, in the domain of *physics* we can distinguish *nuclear physics*, *optics*, *chemical physics*, *experimental physics*, etc. In fact by different domains we mean the domains whose keyword vocabularies have no or very few words in common. In this case the size of the documents is not important for clustering in the keyword space: any clustering procedure will divide such documents into clusters - which are just the domains - well enough, since the documents are mapped to completely different keyword subspaces of the whole keyword space of the document collection.

A weak intersection of domain vocabularies can slightly disfigure the results: some documents can contain keywords only from this intersection, which is much more probable for short documents than for large ones. To avoid such an effect, one can remove common words from the list of index keywords and work only with the "pure" domains without keywords in common, as described above. Some short documents may in this case prove not to have any index keywords; they can be collected together and classified manually. The words in common can be found by a simple two-step procedure: (1) construct the list of keywords for the whole document collection and use it for clustering all documents; (2) construct the keyword lists for every document cluster and select the words in common.

When we deal with documents from one given domain, the situation is cardinally different. All clusters to be revealed have strong intersections of their vocabularies and the difference between them consists not in the set of index keywords but in their proportion. This causes very unstable and thus very imprecise results when one works with short documents, because of very low absolute frequency of occurrence of the keywords in the texts. Usually only 10%–20% of the keywords from the complete keyword list occur in every document and their absolute frequency usually is 1 or 2, sometimes 3 or 4. In this situation, changing a keyword's frequency by 1 can significantly change the clustering results. Thus the first difficulty consists in:

– very low absolute frequencies of occurrences of the keywords from the general keyword list in the text, which leads to unstable results.

Thus our first problem is to assure stability and thus correctness of the results of clustering.

Consider now another classification of short documents:

1. Document collection containing news or other self-contained information;
2. Document collection containing abstracts of full-text scientific or technical documents not included in the collection.

When clustering a new data collection, we must assure first of all validity of clustering results, i.e., good quality of grouping according to the given internal criteria. When we work with abstracts, we always have an additional specific criterion of closeness between abstracts being clustered and the corresponding full-text documents. Consider manual clustering of full texts as the gold standard, we can assess usability of our clustering results. Some researchers evaluate the usability comparing automatic versus manual clustering of only abstracts; however, we use a more rigorous gold standard consisting in dealing with full texts.

Our experiments with clustering abstracts and full-text papers show that it is practically impossible to obtain coincident or at least very close clusters. In fact, abstracts and full-text documents have different contents: Indeed, the abstracts explain the goals of the research reported in the paper (the problem), while the paper itself explains the methods used to achieve these goals (e.g., the algorithms). In consequence, a collection of abstracts and a collection of full-texts documents have significantly different keyword lists; at least they use the lexicon in different ways. Thus, the second difficulty consists in:

– significant difference between the use of keywords in abstracts and their full-text counterparts, which leads to imprecise results.

Consequently, our second goal is to provide more exact results with respect to the closeness between clustering abstracts and full papers. The problem of clustering abstracts arises when one works with documents from one narrow domain. Abstracts belonging to significantly different domains are clustered well, but this case is not interesting; any search engine can easy classify such abstracts.

1.2 Related Work

Though there exists extensive literature on information retrieval [3, 21], the problem of clustering narrow-domain short documents is not well-studied. One of the reasons for this consists in that when clustering algorithms are applied to multi-domain document collections no problems arise.

The only works concerning categorization of short documents we are aware of use supervised methods, i.e., are based on prior training [8; 22]. These works obtained excellent results, but for a different situation, because we deal in an unsupervised manner with clusters that are unknown beforehand, rather than with predefined categories. Makagonov *et al.* [10] considered the problem of clustering abstracts. However, the document collection contained the texts from easily distinguishable domains, and the number of domains was known beforehand.

Makagonov *et al.* [12] used stronger criteria of keyword selection and a combined measure of closeness between documents (cosine and polynomial ones). These criteria can give more confidence to the low absolute frequencies of keyword occurrences in abstracts; the combined measure can make the results closer to expert opinion. However, both techniques received some critics because they were not justified well enough and were not tested in more complex situation (the number of clusters was

known in advance). Alexandrov *et al*. [2] considered two procedures for clustering abstracts from a given narrow domain using clustering keywords. Those results were very preliminary, and no discussion of limitations of the method was provided. Besides, both works used well-known clustering methods - *k*-means and the nearest neighbor method - while there exists very promised method MajorClust [17], which have been demonstrated to give excellent results on clustering textual documents in comparison with the two mentioned methods [18].

In this paper we suggest the following modifications of the traditional approach, which significantly improve clustering results:

- Grouping words from the word frequency list, to make the results more stable and correct;
- Transformation of usual cosine measure of document similarity, to take into account the difference between abstracts and full-text documents.

The first suggestion is close to that proposed in [9], where clusters of keywords were used for constructing semantic space for information retrieval problems. Our procedure is simpler: we do not use any linguistic information on the relation betweens words; we rely only on links extracted with statistical methods from a document corpus. Our second suggestion continues that proposed in [12], where combined measure of document similarity was used. Again, our procedure is simpler: we use logarithmic transformation of term frequencies, inspired by the information-theoretic aspect of the problem.

For clustering keywords, we used MajorClust method. In the paper we pay attention to the limitations caused by grouping keywords and their stronger selection.

2 Main Algorithm

2.1 Indexing

Preliminary keyword selection. In the framework of the traditional approach, random character of document presentation (as considered in the vector space model) is not taken into account: all word occurrences are considered fixed values; this is justified for long documents. The corresponding procedure finds all words in the whole document collection, filters out stopwords, and joins the words having the same base meaning using, for example, Porter stemming algorithm [14].

In case of abstracts, we have a different situation. Namely, one or two occurrences of any low-frequency word in a text double its frequency count. Because of the random nature of such occurrences, the error of the frequency estimation becomes comparable to the frequency itself. To increase the confidence for low-frequency words, we have to perform word selection. For this, we use the following criterion introduced in [11]: only those words W are included in the index keyword list that satisfy the following inequality:

$$F_{Dom}(W) >> F_{Gen}(W); \text{ namely, } F_{Dom}(W) / F_{Gen}(W) > k, \quad (1)$$

where $F_{Dom}(W)$ and $F_{Gen}(W)$ are the frequencies of the word W in the given document collection and in a general balanced corpus of the given language (a corpus of general use), respectively.

The parameter k is determined empirically. Its value is related to the statistical estimation of the mean error in the measuring of the frequencies due to the limited size

of the sample texts. A reasonable value for k must be greater than 3 or 4 for low-frequency words in short texts; in our practical work we used $k = 7$ in order to obtain stable enough results.

On the other hand, a stable result does not mean a good result: taking only one keyword we would obtain the most stable but absolutely useless result. Indeed, eliminating words we lose certain information; extremely strong filtering leads to absurd results. Thus in our experiments we used $k = 2$ while stability of results was achieved by grouping keywords as described below (which allows for a lower threshold k).

To conflate the words having the same base meaning, we used empirical formulas for testing word similarity [1].

Grouping keywords and weighting. Grouping keywords is an efficient way to compensate for the effect of low frequencies of these keywords. In this case, every group of keywords can be considered a new coordinate in the index space, equal to the sum of the occurrences of all keywords in a given group. We will call this sum *cluster frequency*. One can suggest several variants of how to group the words having close semantics. For example:

- use synsets of an appropriate ontology (e.g., WordNet or a synonym dictionary);
- use a thesaurus related to a given domain;
- cluster the words in the space of documents themselves.

Two former variants use external information. The latter variant is the simplest; we discuss just this variant in this paper. However, here we have to answer on two important questions:

- How many keyword clusters should one take for clustering documents?
- How to evaluate semantic significance of each keyword cluster?

Obviously, the result of clustering documents crucially depends on the number of keyword clusters. To solve such a problem, we suggest using one of density-based methods. These methods automatically determine the number of clusters; one of them, MajorClust, constructs very natural clusters (in the sense that the revealed clusters are the closest ones to those selected by human experts). This method was suggested in [17] and investigated in [4, 18, 19, 20].

When grouping keywords, we suppose a certain semantic closeness between them. Indeed, we suppose that the absence of some words from a cluster may be compensated for by the presence of the others. To evaluate the semantic significance of a cluster, we use the average distance between all words included in the cluster:

$$S_k = \sum_{i,j} d_{i,j} / N_k, \tag{2}$$

where k is the number of the cluster, i and j are the elements of this clusters ($i \neq j$), N_k is the number of links in the cluster k .

2.2 Clustering

Document similarity measure. We consider the vector model of document representation. To evaluate the closeness between two documents, we use the well-known cosine measure:

$$C_{1,2} = \frac{\sum_k (x_{k1}, x_{k2})}{\|x_1\| \|x_2\|}, \tag{3}$$

where x_{k1} and x_{k2} are the vectors corresponding to the documents 1 and 2. In our case, x_{k1} and x_{k2} are relative cluster frequencies. The difference from the traditional cosine measure consists in the following:

- the coordinates in (3) correspond to the clusters of keywords as described above,
- these coordinates are weighted using the semantic coefficients from (4) below.

We have already mentioned that in case of clustering abstracts one should take into account the difference between contents of abstracts and their full-text papers. Namely, the direct ratio of the coordinates should be changed taking into account the information-theoretic aspect of the problem. Indeed, the abstracts usually introduce the reader to the possibilities of a suggested approach or method, while the full papers give its more or less detailed explanation. This leads to the necessity to change the document representation in the document similarity measure to use logarithm of the vector coordinates:

$$x_k = \log(1 + f_k), \tag{4}$$

where f_k are the relative cluster frequencies. It is these coordinates that are included in (3). The experiments described in the next section support this hypothesis.

Clustering methods. In the suggested approach clustering is applied twice: for grouping keywords and for grouping abstracts. It is well-known that the number of methods and their modifications used in cluster analysis are more than authors working in this area. [7]. Extensive literature is devoted to such methods and their applications in text processing [13, 19].

For clustering abstracts, we used the *K*-medoid method, the nearest neighbor method, and MajorClust method. These are the simplest implementations of the exemplar-based, hierarchy-based, and density-based approaches, respectively. Our goal was not selecting the best clustering method for abstracts; from the previous discussion it is clear that this is separate and difficult task. We used the former two methods, since they are in a sense the most "contradictory" ones: they give the least coincident results on various data sets as compared with other pairs of clustering methods. This was noted by Solomon [16], where instead of *K*-medoid method the usual *K*-means was tested. This was also investigated in detail by Stein *et al*. [4, 18]. Thus closeness of the results of these methods would be a strong indication of stability of obtained clusters.

MajorCluster is a new method firstly desribed in [17]. We slightly modified it in order to avoid circling related with weak connections between the documents due to their small size.

The idea of the method is very simple: it distributes objects to clusters in such a way that the similarity of an object to the assigned cluster exceeds its similarity to any other cluster. Neither *K*-means (*K*-medoid) nor the nearest neighbor (NN) methods possess such optimization property: the former one provides the maximum closeness of objects to the centers of clusters to be constructed, while the latter one provides maximum connectivity of objects inside clusters, independently of their similarity to

other clusters. MajorCluster method works as follows: first, every object is considered a separate cluster. Then the objects are joined to the nearest cluster. In the process of cluster construction, the objects can change their cluster – in contrast to the NN-method. This algorithm is an implementation of the algorithm for graph clustering based on the notion of weighted edge connectivity [17].

For clustering words we used only MajorClust method, because we have no a priori information concerning the number of keyword clusters. Numerous experiments show that this method outperforms both other methods, independently of index selection for the given document collection, such as RCV1-Reuters Corpus, vol. 1 [18–20].

3 Experiments with Web-Retrieved Information

3.1 Data Source and Clustering Quality

CICLing-2002 conference collection. In our experiments, we used a document collection consisting of the abstracts of the CICLing-2002 Conference (Conference on Computational Linguistics and Intelligent Text Processing; www.CICLing.org), which is a narrow domain-oriented conference [5]. The document collection consisted of 48 abstracts (40 KB of text). After indexing, the domain dictionary contained approximately 390 words. Though this is a small collection, our research is preliminary, our goal being to attract attention to the problem and to the possible ways of its solution. For this, one does not need a very large collection. On the other hand, this allowed for careful manual classification and detailed evaluation of the results.

A human expert manually classified the papers of the Conference into 4 and 11 classes, which were, according to the expert's judgment, natural. This can be explained as follows. Classification into 2 classes was not interesting: the binary situation may be effectively analyzed by the corresponding methods. Classification into 3, 5, ..., 10 classes was not considered by the expert to be balanced. This means that in these cases the classes proved to belong to different levels of hierarchy if the hierarchy-based clustering would be applied. And classification into more than 11 classes gave very small groups.

Here are the four categories of papers selected by the expert (the titles are given only for the reader's convenience):

- Linguistic (semantics, syntax, morphology, parsing),
- Ambiguity (word sense disambiguation, anaphora, tagging, spelling),
- Lexicon (lexicon and corpus, text generation),
- Text processing (information retrieval, summarization, text classification).

The selected categories (classes) are rather fuzzy. For example, the intersection of vocabulary for the documents from the most different second and forth groups was about 70%. This implies that the selected domain is rather narrow.

Validity and usability of clustering. Quality of clustering is evaluated using various objective criteria based in the relations within and between the clusters. The procedure of evaluating the clustering quality is called cluster validation. For testing cluster validity we used so-called index of expected density of clustering [20]:

$$\bar{\rho}(C) = \sum_{i=1}^{k} \frac{|V_i|}{|V|} \cdot \frac{w(G_i)}{|V_i|^{\theta}}, \quad \textit{where} \quad |V|^{\theta} = w(G), \tag{5}$$

where G is a given graph of connections between the objects, G_i is its subgraph, V is the total number of the objects in the graph, V_i is the number of the objects in i-th subgraph, $w(\cdot)$ is the total weight of the edges in the graphs or subgraphs, C stands for a given clustering, and G_i corresponds to the i-th cluster. Higher values of $\bar{\rho}$ mean better clustering.

The correspondence between the results of automatic and human clustering can be evaluated by various measures. The procedure of evaluating is called cluster usability estimation. We use the F-measure in the form presented in [4]:

$$F = \sum_{i=1,\dots l} \frac{|C_i^*|}{V} \max_{j=1,\dots,k} \{F_{i,j}\}; \quad F_{i,j} = \frac{2\, prec\,(i,j)\, rec\,(i,j)}{prec\,(i,j) + rec\,(i,j)}, \tag{6}$$

where $\{C^*_1, \dots, C^*_l\}$ stand for the human classification, $\{C_1, \dots, C_k\}$ for the clusters obtained by the algorithm, $prec(i,j)$ is the precision of the cluster j with respect to the class i: $|C_j \cap C^*_i| \,/\, |C_j|$; $rec(i,j)$ is the recall of cluster j with respect to the class i: $|C_j \cap C^*_i| \,/\, |C^*_i|$.

Stein *et al.* [20] showed that $\bar{\rho}$-index correlates well with F-measure. Thus to provide a good usability of results, $\bar{\rho}$-index can be calculated without participation of human experts.

3.2 Experiments

Experimental setting. Following the traditional approach, we indexed all the words according to well-known *tf* and *tf-idf* measures, where *tf* stands for term frequency and *idf* for inverse document frequency [15].

In the suggested approach, we used all words selected by the rule (1) discussed above, and then clustered them with MajorClust method. This gave 14 clusters of keywords, which were semantically weighted by the formula (2). Before this procedure, the weakest connections were eliminated from the connection graph.

In all our experiments, the standard cosine similarity measure was used to evaluate both the similarity between documents and the similarity between keywords.

Testing the grouping of keywords. The goal of the first series of experiments was to compare the traditional and suggested approaches to indexing. We also checked the sensibility of the results to the change of the document set. Since we had the reference classification provided by a human expert, we could evaluate the quality of clustering using F-measure.

The number of document clusters to be constructed was fixed at 4. The MajorClust method revealed confirmed that this is the natural number of documents clusters.

We conducted our experiments with two document sets: the original one, containing all 48 abstracts, and a reduced one, containing 75% of the whole set, i.e., 36 documents. Because of reducing the number of documents, the vocabulary used for clustering procedure changed. The results are presented in Table 1. Scaling stands for the correction of document coordinates according to (4).

These results show that the suggested approach gives better and more stable results with respect to changing the document set. Besides, one can see that using *idf* factor reduces the stability of the results. The possible explanation of this is a high level of randomness of word occurrences in texts.

Table 1. F-measure for comparison of ways of indexing.

Indexing	Scaling	48 abstracts	36 abstracts
Tf-idf	No	0.56	0.49
Tf	No	0.56	0.51
Grouping	Yes	0.64	0.58

Testing logarithmic measure. The goal of the second series of experiments was to demonstrate the advantage of using logarithmic scaling in cosine similarity measure. Here we used the whole document set.

Table 2. *F*- measure for evaluation of scaling.

Indexing	Scaling	Without scaling
tf-idf	0.64	0.56
Grouping	0.64	0.58

The results show that logarithmic scaling improves clustering both for traditional approach to indexing and for the suggested one.

Testing methods and model complexity. In our last series of experiments, we tested the clustering methods under different number of expected classes. We considered the *K*-medoid, NN- and MajorCluster methods. The number of classes was equal 4 and 11. Since the MajorCluster determines the number of clusters automatically, it was used only one time. The quality of clustering was evaluated by *F*-measure and $\bar{\rho}$ index. The results are presented in Table 3. Note that the values for *F* and $\bar{\rho}$ vary in the range between 0 and 1.

Table 3. *F*- measure and $\bar{\rho}$ -index for different cluster number.

	4 clusters		11 clusters	
Method	*F*	$\bar{\rho}$	*F*	$\bar{\rho}$
NN method	0.64	0.71	0.46	0.42
MajorClust	0.64	0.71	–	–
K-medoid	0.56	0.67	0.49	0.39

The results demonstrate a quite good stability of clustering with 4 clusters: both contradictory methods gave the same results. Worse quality of clustering on 11 clusters can be explained as follows. By joining keywords, we improve stability of the results, but simultaneously lose some information related to the semantics of individual keywords. This means that by clustering keywords we limit the structure complexity (number of classes), which can be revealed from the data using given set of keyword clusters. Reduced value of the validity index supports such a hypothesis.

It is easy to see that changing of expected density index reflects the change of *F*-measure. However, to consider their correlation, one should conduct experiments on different data sets and with different number of clusters. For RCV1 document collection, this was elaborated by Stein *et al*. [20].

4 Conclusions and Future Work

Conclusions of the experiments. Nowadays the problem of clustering abstracts is important for handling documents in digital libraries of scientific information, where the most part of the data is presented in the form of abstracts. We have suggested a technique for clustering such information, which consists in grouping indexes and using logarithmic scaling in document similarity measure. Our experiments with abstracts show its advantage in comparison with traditional approaches.

Proposal on open access to full text document images. Clustering only abstracts one cannot achieve as good results as when clustering full text papers. To facilitate the work of search engines, both in search and in clustering the search results, especially in context of the Semantic Web effort, we propose that digital libraries and Internet repositories *provide open access to document images of the papers.* A document image is a vector of word frequencies, which can be restricted to a small list of keywords extracted from the whole document collection. This does not violate the copyright laws because it is impossible to recover the full text of the paper from such a document image. This proposal was originally suggested by Makagonov [12] and is now under consideration in the Library of the Mixteca University.

Future work. This is a preliminary paper. In the future, we plan to use WordNet and other ontology-related techniques for grouping semantically similar keywords and compare the results with those obtained in this paper. We plan to consider the hypervolume clustering criterion [6] to improve cluster validity. We also plan to apply our techniques for very large medical database of Czech Ministry of Healthcare, in cooperation with our Czech colleagues from Masaryk university. Finally, we plan to evaluate our methods on the 20newsgrous collection.

We will also test the methods of assessing the cluster quality taking into account the possibilities for the division of a given test set into classes. This depends on the intersection of the vocabulaty of the classes.

References

1. Alexandrov, M., X. Blanco, P. Makagonov. Testing Word Similarity: Language Independent Approach with Examples from Romance. In: F. Meziane *et al*. (Eds.) *Natural Language Processing and Information Systems*, Springer, LNCS N 3136, 2004, pp. 223–234.
2. Alexandrov, M., A. Gelbukh, P. Rosso. Clustering Very Short Documents based on Grouping Keywords. *Abstracts of the 30-th Latin-American Conf. on Informatics*, Univ. Edition, Peru, 2004, p. 133.
3. Baeza-Yates, R., B. Ribero-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.
4. Eissen, S.,M., B. Stein. Analysis of Clustering Algorithms for Web-based Search. In: *Practical Aspects of Knowledge Management*, LNAI N 2569, Springer, 2002, pp.168–178.
5. Gelbukh, A., (ed.). *CICLing-2002*, *Comput.Linguistics and Intelligent Text Processing*. LNCS N 2276, Springer-Verlag, 2002; www.CICLing.org.
6. Hardy, A., P. Andre. An investigation of nine procedures for detecting the structure in a data set. In: *Advances in data science and classification*, Springer, "Studies in Classification, Data Analysis and Knowledge Organization," 1998, pp. 29–36.
7. Hartigan, J. *Clustering Algorithms.* Wiley, 1975.
8. Hynek, J, K. Jezek, O. Rohlikm. Short Document Categorization – Itemsets Method. In: *PKDD-2000*, Springer, LNCS N 1910, 2000, 6 pp.

9. Kang Bo-Y., H-J. Kim, S-j. Lee. Performance Analysis of Semantic Indexing in Text Retrieval. In: A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing*, CICLing-2004, LNCS N 2945, Springer-Verlag, 2004, pp. 433–436.
10. Makagonov, P., M. Alexandrov, K. Sboychakov. Keyword-based technology for clustering short documents. In: *Selected Papers. Computing Research*, CIC-IPN, Mexico, 2000, pp. 105–114.
11. Makagonov, P., M. Alexandrov, K. Sboychakov. A toolkit for development of the domain-oriented dictionaries for structuring document flows. In: *Data Analysis, Classification, and Related Methods*, Studies in classification, data analysis, and knowledge organization, Springer, 2000, pp. 83–88.
12. Makagonov, P., M. Alexandrov, A. Gelbukh. Clustering Abstracts instead of Full Texts. In: *Text, Speech, Dialog*, LNAI N 3206, Springer, 2004, pp. 129–135.
13. Manning, D., C. and H. Schutze. *Foundations of statistical natural language processing*. MIT Press, 1999.
14. Porter, M. An algorithm for suffix stripping. *Program*, 14, 1980, pp. 130–137.
15. Salton, G., C. Buckley. *Term-weighted approaches in autometic retrieval*. Information Processing in Management, v.24, 1988, N 5, pp. 513–523.
16. Solomon, G. Data dependent methods of cluster analysis. In: *Classification and Clustering*, Academic Press, 1977, pp. 129–147 (Russian version).
17. Stein, B., O. Niggemann. On the Nature of Structure and its Identification. In: *Graph-Theoretic Concepts in Computer Science*. LNCS, N 1665, Springer, 1999, pp.122–134.
18. Stein, B., S. M. Eissen. Document Categorization with MajorClust. In: *Proc. 12th Workshop on Information Technology and Systems*, Tech. Univ. of Barcelona, Spain, 2002, 6 pp.
19. Stein, B., S. M. Eissen. Automatic Document Categorization: Interpreting the Performance of Clustering Algorithms. In: *Proc. 26th German Conf. on Artificial Intelligence*, LNCS N 2821, Springer, 2003, pp. 254–266.
20. Stein, B., S. M. Eissen, F. Wissbrock. On Cluster Validity and the Information Need of Users. In: *Proc. 3-rd IASTED Intern. Conf. on Artificial Intelligence and Applications* (AIA'03), Acta Press, 2003, pp. 216–221.
21. Strzalkowski, T. (Ed.). *Natural Language and Information Retrieval*. Kluwer Academic Publishers, 1999.
22. Zizka, J., A. Bourek. *Automated Selection of Interesting Medical Text Documents by the TEA Text Analyzer*. In: A. Gelbukh (Ed.), Computational Linguistics and Intelligent Text Processing, CICLing-2002, LNCS N 2276, Springer-Verlag, 2002, pp. 402–404.