Some Keyword-Based Characteristics for Evaluation of Thematic Structure of Multidisciplinary Documents

Mikhail Alexandrov
Alexander Gelbukh
Pavel Makagonov

The problem of classification of documents of complex interdisciplinary character with high level of informational noise is considered. The set of classification domains is supposed to be fixed. A domain is defined by an appropriate keyword list. Quantitative and qualitative characteristics, as well as visual presentations used for such classification are discussed. A program Text Recognizer based on these characteristics is presented.

1 INTRODUCTION

1.1 Practical tasks

Let us consider some practical examples. About 40 to 60 thousands letters, appeals, and complains of Moscow dwellers, for example, are received every year by the Reception Office of Moscow Mayor Directorate. Each of them is to be directed for consideration to a corresponding department of the Government. The set of the departments and their topics of competence are fixed or at least change infrequently.

Another example: Every day the director of the Center for Computing Research (CIC) of the Mexican National Polytechnic Institute (IPN) receives dozens of various letters and messages concerning the financial relations, collaboration, and specific work in the field of Computer Sciences. These letters are to be forwarded to the appropriate departments of the Center or – in some difficult cases – are to be personally considered by the director.

In these and similar examples, the documents under consideration have following specific features in common.

First, they have high level of information noise – the information that is useless for classification of the document. For example, a dweller gives naive advices on a "better" city management, reasons about his or her achievements, and so on – which have nothing to do with, say, a municipal housing he or she is asking for. Thus, the thematic structure of the document is to be detected basing on only 10% to 30% of useful information in the text. The usual business

correspondence also often contains information noise related with references to previous letters, description of various difficulties, etc. Though in this case the level of information noise is smaller than that in dweller's letters, it is still significant, up to 10% to 30%.

Second, many such documents are devoted to several themes in almost equal degree. For instance, a dweller asks for a pension and in the same letter complains about police in the district and discusses the personality of region's authorities. Or, correspondence received by CIC simultaneously reflects many themes: administrative, educational, scientific activity in various fields (while even these fields have interdisciplinary character), etc. The classification program should detect this and, say, send the document to the person dealing with such complex cases.

Such a situation is quite usual in many document-processing tasks in government, business, or scientific organizations, information agencies, etc.

1.2 Related work

The present paper deals with document classification applications of a set of dictionaries and with visual representation of the relations between dictionaries and documents, including grouping of texts by thematic structure. Dictionary-based algorithms of document classification similar to the methods we present here were described in [Guzman-Arenas, 1998]. However, in that paper a very large predefined concept tree is used; in contrast, we consider the case of a relatively small set of domains that the users can easily define or change.

There exist effective document classification algorithms relying on the differences in the frequency properties of the words in the general versus specific domain texts [Feldman, 1995, Alexandrov, 1999, Gelbukh, 1999]. In our case, however, no pre-existing knowledge about the general lexicon is used. Also, in the present work it is important that we deal with a set of dictionaries and not with one dictionary. While [Alexandrov, 1999, Makagonov, 1999] discuss mostly the issues of compilation and maintaining of dictionaries, we concentrate on their use.

2 DOCUMENT METRIZATION

2.1 Domain dictionary

A set of domain dictionaries is necessary to obtain a numerical representation of the document, which permits to use the traditional methods of numerical analysis for the task of document classification.

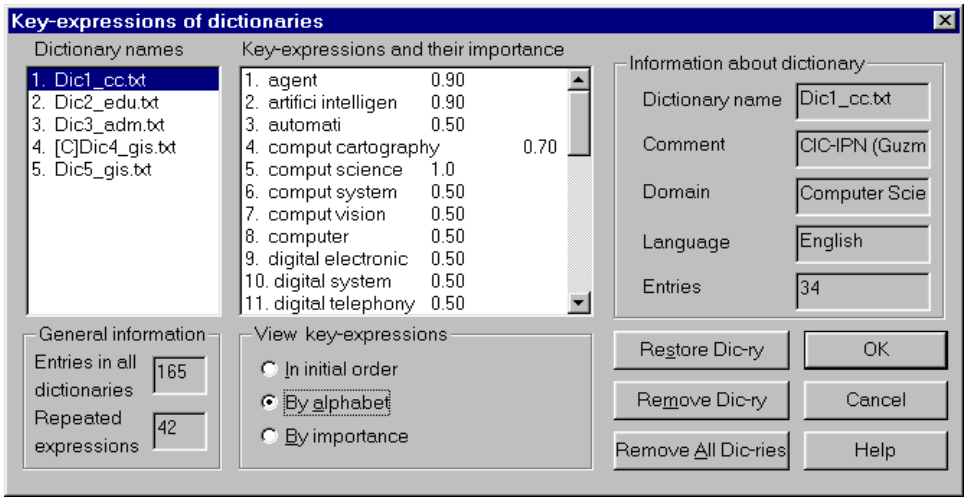

Figure 1. A domain dictionary.

We will use the term *keyword* to refer to any key expression that can be a single word or a word combination. What is more, we represent a keyword by a pattern describing a group of words with equivalent meaning. In such a pattern, the inflection for time, person, gender, number, etc., as well as part of speech distinction, some suffixes, etc., are ignored, e.g.: *obligation*, *obligations*, *obligatory*, *oblige* → oblig-, where oblig- is the pattern representing all these words. For simplicity we call such a pattern a keyword.

A domain dictionary (DD) is a dictionary consisting of such keywords (i.e., patterns) supplied with the coefficients of importance for the given domain. The coefficient of importance is a number between 0 and 1 that reflects the fuzzy nature of the relationship between the keywords and the selected domain, i.e., a DD is a fuzzy set of the keywords.

The methodology of creating domain dictionaries includes analysis of both domain-oriented texts selected by the experts and the frequency list of general lexicon [Alexandrov, 1999, Makagonov, 1999]. In practice, the coefficients are determined basing on an expert's or the user's intuition. The general recommendations for their assignment are: the keyword that is essential for the given domain is assigned the weight > 0.8, an important one 0.6 to 0.8, regular 0.4 to 0.6, important for this and also for some other domains 0.2 to 0.4, typical for many domains < 0.2. If a domain dictionary does not contain these coefficients, they all are considered to be 1. In the simplest case a user can build DDs using her own system of preferences. Figure 1 shows one of the DDs that we use in the CIC for selection the messages related to the topic of Computer Science from the flow of incoming messages.

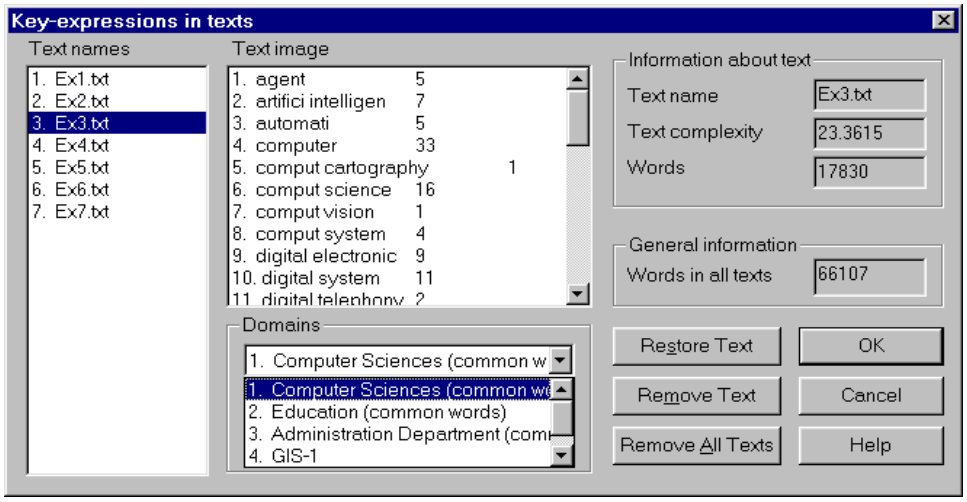

Figure 2. A document image relative to a specific domain.

The domains that define the thematic structure of documents are supposed not to be significantly interrelated. In other words, the DDs under consideration have no significant intersection. If two DDs constructed by the user significantly intersect, they should be joint into one combined domain. The intersection is measured as that of fuzzy sets, i.e., basing on the coefficients of importance.

2.2 Document image

Given a DD, for every document its so-called document image relative to the domain defined by the given DD can be built. Such an image is a list of the domain keywords with their corresponding numbers of occurrences in this document. Given several DDs, several images for a document are built, one for each domain. Figure 2 shows an example of a document image.

Thus, each document is represented with a set of numerical vectors (X_{1j}, X_{2j}, ..., X_{kj}, ...,), one for each domain j. Note that such a vector representation does not imply that any of the traditional vector operations can be used, since they do not represent any real vectors. In particular, the zero vector represents a document that has no relation to the selected theme. Consequently, no binary operations usually used for vectors in mathematics can be applied to such document images.

3 EVALUATION OF THEMATIC STRUCTURE OF ONE DOCUMENT

3.1 Qualitative characteristics

A set of document images represents the raw absolute measure of the "intersection" between a set of DDs and a set of documents. However, the corresponding "densities" rather than absolute amounts are more interesting and informative characteristics. These can be obtained by a suitable normalization of the absolute data. In theory, there are four parameters to normalize the corresponding data by:

- By the size of each individual document,
- By the total importance of each individual dictionary (see below),
- By the set of the documents,
- By the set of the dictionaries.

This amounts to a total of 16 possible combinations of different normalizations. However, not of them are equally useful. We use the following quantitative characteristics for evaluation of the relation between a document and the set of domains:

1. The absolute (not normalized) weight of a specific domain, that reflects the total amount of the information related to this domain in a document.
2. The relative weight (density) of a specific domain, that reflects the amount of the information related to this domain per page of the document (more precisely, per a fixed number of words).
3. The thematic structure of the document, which is a normalized vector of relative weights.

Calculation of these characteristics involves all document images, the size of each document, and the importance coefficients of individual expressions for every domain.

Let us denote (X_{1j}, X_{2j}, ..., X_{kj}, ...,) the image of some document for j-th domain, and (A_{1j}, A_{2j}, ..., A_{kj}, ...,) the coefficients of importance for the corresponding keywords. A naïve way to calculate the weight of the domain in a given document could be $W'_j = \sum_{k=1}^{L_j} A_{kj} \times X_{kj}$, where L_j is the size of DD_j. However, the domains are usually not in equal conditions: their DDs can have different sizes and very different importance coefficients. Thus, the weights of the texts are to be corrected taking into account the "power of the dictionary" $P_j = \sum_{k=1}^{L_j} A_{kj}$, the correct weight being $W_j = W'_j / P_j$. Such an absolute weight of the domain for the document reflects the total amount of the information concerning the given domain in the given document.

Alternatively, if we want to evaluate the correspondence of the document to the domain or to compare several documents, then the domain weights are normalized by the document size. For this, we consider a 1000 word document as a standard size document. If our real document contains M words then the normalizing coefficient is $1000/M$, and the final relative weight of j-th domain is $W_j = (1000/M) \times W_j$. With this, if one concatenates two copies of the same document into a new document, the relative weight will not change. On the other hand, if one concatenates a document with another document which has the same length and which has nothing to do with the given domain, the relative weight decreases twice. These examples reflect the intuition of the share of the document occupied by a given domain.

When discussing the thematic structure of a document, one should take into consideration the relation between the themes reflected in this document. The documents having similar thematic structure must have similar relations between their themes, i.e., these documents must have similar thematic vectors (W_1. W_2,... W_j, ...). For the convenience of further comparison of the document thematic structures, the thematic vectors are to be normalized. Such normalization may be realized in several ways depending on the task.

If the user wants to emphasize the most relevant domain for the document, the weights are normalized by the maximal weight: $W'_j = W_j / W_M$, $j = 1, ..., N$, where N is the number of DDs and $W_M = max \{W_j\}$. However, this operation has some deficiency: The most relevant domain always has relative weight 1, which creates an illusion of that the document is very closely related with this domain.

On the other hand, if the user wants to emphasize the relation between the domains in the document, the weights are normalized by the total weight: $W'_j = W_j / W_S$, $j = 1, ..., N$, where N is the number of DDs, $W_S = \sum_j W_j$. This operation has another deficiency: When more domains are added to the current set of domains (i.e., new DDs are attached to the program), the relative weight of every domain decreases. This creates the illusion that the document becomes less and less connected with each of the domains. And vice versa, when any domain is eliminated (detached from the program) the relative weight of the other domains increases, which creates the illusion of that the removed domain was some source of noise. The former way seems to be more preferable if the thematic structure of a document set rather an individual document is considered. In Text Recognizer, it is this characteristic that is used.

3.2 Qualitative characteristics

As it was mentioned before, the latter form of the normalization operation emphasizes the main theme in a document that thus always has weight 1. These

Table 1. Subjective estimations of domain representativity by experts.

Number of keywords from DD occurring in the document	Qualitative estimations	Numerical estimations	Density of keywords in the document (per mille)	Qualitative estimations	Numerical estimations
More than 75 %	High	1	More than 50 ‰	High	1
25 % to 75 %	Mean	0	10 ‰ to 50 ‰	Mean	0
Less than 25 %	Low	–1	Less than 10	Low	–1

operations in essence actually remove information noise from the document image. However, at the same time the information about the real contribution of domains to the document is lost. To indicate the real representativity of various domains in the documents, two characteristics should be taken into account simultaneously:

1. Density of keywords in the document.
2. Coverage of the dictionary.

The expert's opinions on the correspondence between these characteristics and domain representativity are presented in Table 1. Using numerical equivalents of mentioned characteristics, a *generalized characteristic of domain representativity* can be constructed as

$$D=D1 + D2$$

where *D1* is the numerical estimation of domain representativity on the basis of density of keywords in the document, *D2* is the numerical estimation of domain representativity on the basis of coverage of domain dictionary. The domain representativity is a heuristic reflecting equal role of these characteristics.

The sum gives us some integer value in the interval in the interval (–2, 2), with the correspondent qualitative estimations in the interval (*Very low*, *Very high*). Text Recognizer presented this ordered qualitative scale using color transparency.

Figure 3 shows the thematic structure of a document EX4.TXT. As one can see, the relative contributions of the five domains are approximately (0.1, 0.2, 0.6, 0.95, 1.0). However, the second dominating theme (domain 4) is represented very weak and this theme must be excluded from consideration. High relative weight of this theme was caused by repetition of very limited list of keywords from the appropriate dictionary. It means that some sub-domain of the given domain (or just an unrelated theme) was essentially represented in the document, rather than the whole given domain. This visual representation allows compensating of the information loss caused by normalization.

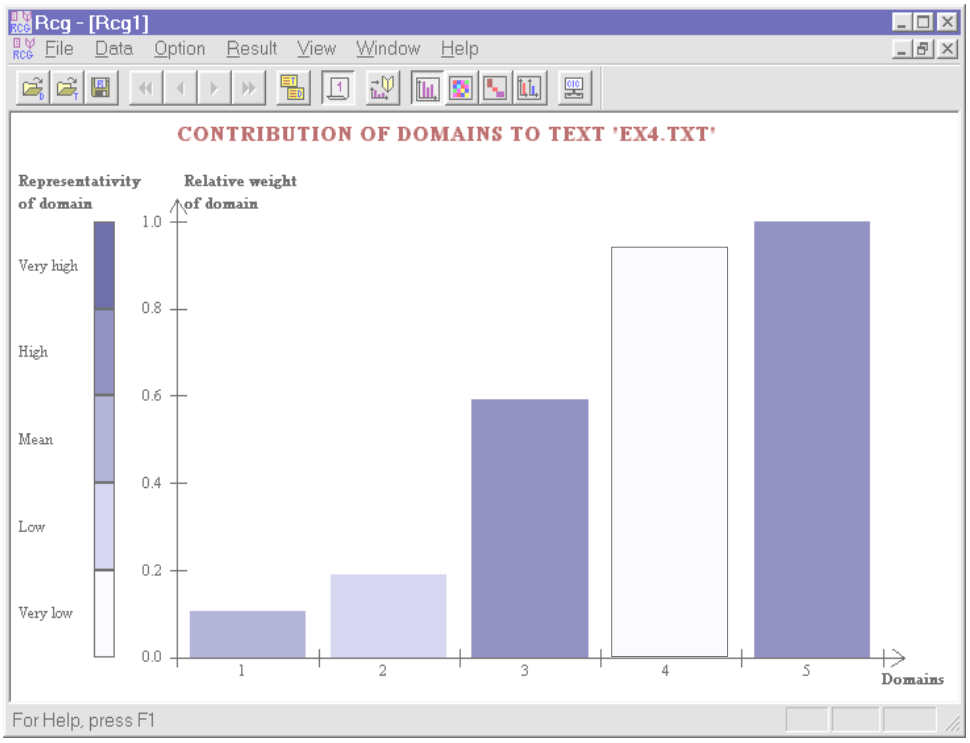

Figure 3. Thematic structure of a document and representivity of various domains

4 EVALUATION OF THEMATIC STRUCTURE OF DOCUMENT SET

4.1 Quantitative characteristics

When a thematic structure of a document set is considered, it implies first of all looking for documents with similar structures. As it was mentioned before, the thematic structure of a document is represented by a thematic vector. Consequently, similar documents have similar direction of these vectors in the space of domains. Thus the correspondent measure of closeness reflects just the angle between two thematic vectors (W'_{11}, W'_{12}, ..., W'_{1N}) and (W'_{21}, W'_{22},, ..., W'_{2N}). For this, a correlative measure is used, which is the inverse value to the correlation between the two vectors, i.e., the normalized scalar product:

$$R_{12} = \frac{\sum_j \left(W'_{1j} \times W'_{2j}\right)^2}{\|W_1\| \times \|W_2\|}.$$

Similar vectors have $R_{12} \approx 1$; this means that the distance $D_{12} = 1 - R_{12} \approx 0$, that corresponds to the intuitive notion of the distance.

Besides correlative measure, it can be used linear and quadratic measures that are calculated in a usual mathematical way:

$$D_{12} = \sum_j \left| W'_{1j} - W'_{2j} \right|, \; D_{12} = \sqrt{\sum_j \left(W'_{1j} - W'_{2j} \right)^2}$$

In Text Recognizer program, the following three measures are used:

1. Correlative,
2. Linear
3. Quadratic.

These different measures are used in different circumstances, in spite of that if the distance between two documents is close to 0 in the correlative measure then the distance between them will be also close to 0 in the other measure, which is a consequence of normalization of the thematic vectors. Namely, if the distance between two documents is close to 1 in the correlative measure, i.e., the vectors are orthogonal, then the distance between them in the other measures can be any arbitrary value greater than 1. Thus, linear and quadratic measures are useful for verifying stability of the results obtaining with the correlative measure.

4.2 Qualitative characteristics

The absence of good formal criteria for grouping documents according to their thematic structure motivates the use of subjective qualitative characteristics that reflect the distribution of the themes by the documents. These characteristics are determined by the prevailing opinions of experts based on their experience; in the program, the user can the default values or tune them according to his or her own experience. Such characteristics are used to compare sets of documents.

In Text Recognizer, the distance between documents in the space of domains, or between domains in the space of documents is measured, and then the program groups them according to their closeness. Then the user can select and to view some clusters.

To represent these characteristics visually, we use an *ordered colored document/domain matrix* that is a convenient representation for supervised evaluation of thematic structure of document set. Color matrices have been used for a long time for a fast informal evaluation of high-dimensional data [Grishin 1982]. In our practice, with such matrices the experts can quickly solve the problem of comparison of various document sets at a glance.

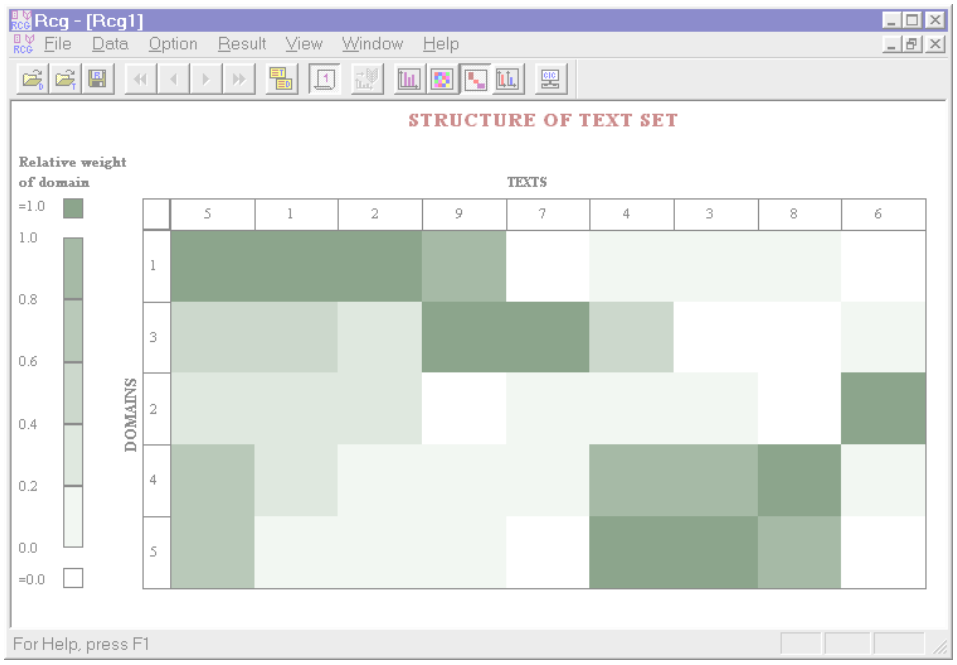

Figure 4. Document set after clustering using the *correlative* measure.

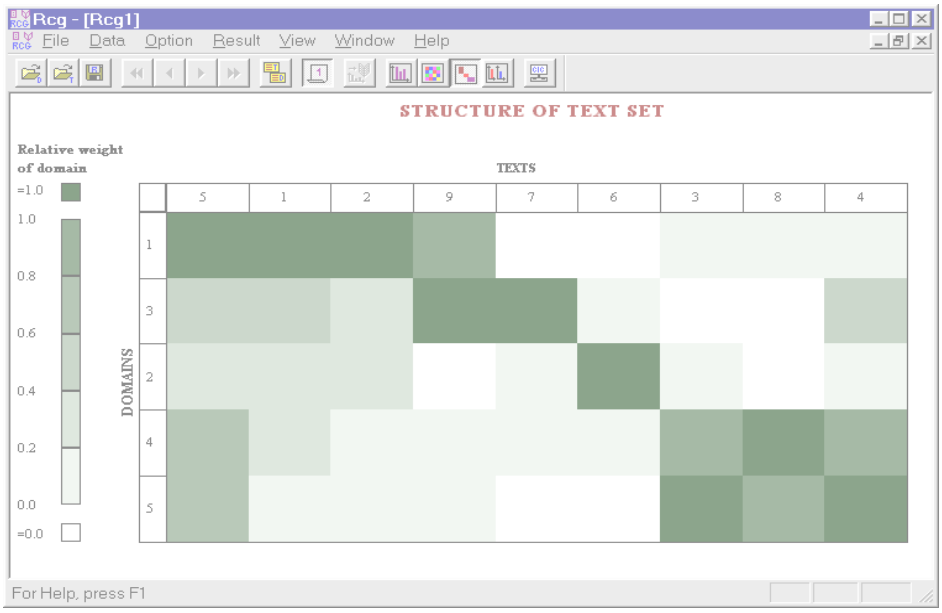

Figure 5. Document set after clustering using the *quadratic* measure.

In Figure 4 and Figure 5, an example for a set of real documents received in CIC, such as official papers of the Directorate, letters from other institutions, electronic books, etc., is shown. The figures show one cluster formed by the documents (5, 1, 2) and another one formed by the documents (4, 3, 8). These figures demonstrate the stability of the obtained results.

5 CONCLUSIONS AND FUTURE WORK

In this paper, the problems of evaluation of the thematic structure for one document and a set of documents has been considered under condition of high level of information noise and the presence of several domains in almost equal degree. The solution of this problem is possible on the basis of domain dictionaries containing domain-oriented sets of keywords. These dictionaries allow to built the numerical images of textual documents and then calculate various estimations of the thematic structure.

Formal quantitative characteristics for evaluation of document thematic structure are suggested. However, they have limited possibilities and do not solve the problem completely. Some qualitative characteristics compensating for these limitations have been suggested. These characteristics rely on the possibilities of visual analysis.

A program Text Recognizer realizing the technology being discussed has been presented. It was extensively tested on real-world texts, including the materials of large international Conference 'APORS-97' [Makagonov, 1999]. It is currently used in the Mayor Directorate of Moscow City Government for the work with textual database, "Sustainable development of cities of Russia." Now TextRecognizer is being tested in the Department of Environment Protection of Mexico City Government for the working with the text archive of ecological data. In our future work, we plan to implement more functions reported as desirable by the current users, in order to turn our system into a convenient workplace for text classification.

REFERENCES

Alexandrov, M., Makagonov, P., and Sboychakov, K.(1999): Searching similar texts: some approaches to solution. In V. Vorsevich et al (Eds.): *Acta Academia.* Intern. Inform. Acad. with UN, Chisinau,Moldova., 215-223.

Feldman, R., Dagan, I. (1995). Knowledge Discovery in Textual Databases. *Proc. of Intern. Symposium "KDD-95".* Montreal, pp.112-117

Gelbukh, A., G. Sidorov, and A. Guzmán-Arenas. A Method of Describing Document Contents through Topic Selection. *Proc. SPIRE'99, International Symposium on String Processing and Information Retrieval*, Cancun, Mexico, September 22 – 24. IEEE Computer Society Press, 1999, pp. 73-80.

Grishin, V.G. (1982). Pictorial analysis of data structures and object states. *IFAC/IFIP/IFORS International Conference on analysis, design, and evaluation of man-machine systems.* Pergamon Press, pp. 319–327.

Guzman-Arenas, A. (1998). Finding the main themes in a Spanish document. *Intern. J. Expert Systems with Applications*, v. 14, N 1/2, 139-148.

Makagonov, P., Alexandrov, M. and Sboychakov, K. (1999): Searching in full text Data Bases by using text patterns. In Pedro Galicia (Ed): *Proceedings of International Computer Symposium CIC'99 (Mexico, 1999).* National Polytechnic Institute, Mexico, 17-29

Mikhail Alexandrov is a professor and researcher of the Natural Language Processing laboratory of the Computing Research Center (CIC) of the National Polytechnic Institute (IPN), Av. Juan Dios Bátiz s/n esq. Mendizabal, col. Zacatenco, CP 07738, DF, Mexico. His main interests include mathematical modeling and mathematical methods of classification and clusterization. He can be reached at dyner@cic.ipn.mx.

Alexander Gelbukh is a professor and researcher and the head of the same laboratory. He is author of about 90 publications in the field of computational linguistics, including computational morphology, syntax, semantics. He can be reached at gelbukh@cic.ipn.mx or gelbukh@earthling.net, see also http://www.cic.ipn.mx/~gelbukh.

Pavel Makagonov is the Deputy Chief of the Moscow Mayor's Directorate of the Moscow City Government, Novi Arbat 36, 13-th floor, Moscow 121205, Russia. His main interests include the problems of megapolis management, statistical methods of decision making, mathematical and visual methods of classification and clusterization. He can be reached at makagon@maria3.munic.msk.su.

The work was done under partial support of CONACyT, REDII, and CGEPI-IPN, Mexico.