МАТЕМАТИЧКЕ И ИНФОРМАЦИОНЕ ТЕХНОЛОГИЈЕ
МАТЕМАТИЧЕСКИЕ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ
MATHEMATICAL AND INFORMATIONAL TECHNOLOGIES

МИТ 2016

ВОДИЧ КОНФЕРЕНЦИЈЕ
СПРАВОЧНИК КОНФЕРЕНЦИИ
CONFERENCE INFORMATION

28.08. - 31.08.2016. Vrnjacka Banja, Serbia

01.09. - 05.09.2016. Budva, Montenegro

$$\frac{G(\mathbf{x})-H(\mathbf{x})}{A(\mathbf{x})-G(\mathbf{x})}\to 1, \quad \text{and} \quad \frac{A(\mathbf{x})-G(\mathbf{x})}{Q(\mathbf{x})-A(\mathbf{x})}\to 1,$$

if $\mathbf{x}\to\mathbf{a}, \quad \mathbf{x}=(x_1,\ldots,x_N), \qquad \mathbf{a}=(a,\ldots,a)$.

For any power mean value M_p we have

$$M_p(\mathbf{x})-A(\mathbf{x})=(p-1)\big(Q(\mathbf{x})-A(\mathbf{x})\ +o\ Q(\mathbf{x})-\mathbf{x})\ =\frac{p-1}{2a}\sigma^2(\mathbf{x})+o(\sigma^2(\mathbf{x})), \quad \mathbf{x}\to\mathbf{a},$$

where $\sigma^2(\mathbf{x})$ is a variance of $\mathbf{x}$. For the moments about the arithmetic mean

$\mu_n(\mathbf{x})=\left(\sum_{i=1}^{N}(x_i-A(\mathbf{x}))^p\right)/N$, for example, obtains

$\mu_3(\mathbf{x})=\big(Q(\mathbf{x})-A(\mathbf{x})\big)^2\cdot O$.

References

1. M. Bjelica, Asymptotic linearity of mean values, Matematički Vesnik 51 (1999), 15-19.

PARALLEL TEXT DOCUMENT CLUSTERING BASED ON GENETIC ALGORITHM

M. Mansurova[1], V. Barakhnin[2,3], S. Aubakirov[1], E. Khibatkhanuly[1], and A. Musina[1]

[1] Al-Farabi Kazakh National University, Almaty, Kazakhstan
[2] Institute of Computational Technologies SB RAS, Novosibirsk, Russia
[3] Novosibirsk State University, Novosibirsk, Russia

This work describes parallel implementation of algorithm FRIS-Tax for clustering of a corpus of documents. The algorithm is based on evaluated of the similarity between objects in a competitive situation, which leads to the notion of the competitive similarity function [1, 2, 3]. Similarity measure m on the set of documents D is assigned as follows (1):

$$m: D\times D\to[0,1], \tag{1}$$

and the function *m* in the case of full similarity takes a value 1 and in the case of full differences – 0. Similarity measure can be calculated using the formula (2) :

$$m(d_1,d_2)=\sum a_i m_i(d_1,d_2) \tag{2}$$

where *i* is number of the attribute of the bibliographic description, a_i is the weight coefficients, $m_i(a_1, a_2)$ is the similarity measure by the i^{th} element.

To determine the similarity measure, the attributes of bibliographic description of documents were chosen. As the attributes the year; the unified code; the keywords; the authors; the series and abstracts have been selected. To choose weight coefficient, which are used in the formula of similarity measure (2), a genetic algorithm has been developed. The genetic algorithm consists of the following stages ([4]): creation of initial population; selection; choice parents; crossing and mutation.

To create the initial population and its further evolution, the ordered chain of genes or genotype has the length equal to 13 representing a set of parameters made up on the basis of attributes of the bibliographic description of documents. In the genetic algorithm, a set

of individuals each with its own genotype presents some solution of clustering task. Let us suppose that we have generated an individual, i.e. a set of weight coefficients for defining the similarity measure is set, specified. Then, clustering FRIS-Tax is executed where the measure of closeness is computed with the given set of weight coefficients. In the algorithm, a fitness-function is set which allows to determine how well the task of clustering is executed. The quality of the obtained clusters in this work is evaluated with the help of an external criterion of clustering quality "Purity" [7]. To decide which of the individuals failed to be chosen and is dying and which one is surviving and will take part in reproduction, a lower boundary (Threshold) for fitness-function is set up.

The time of FRIS-Tax operation increases exponentially with the increase in the amount of articles. In this relation, to speed up the work at two stages of the algorithm, technologies of parallel computations were used. First, when choosing individuals in a genetic algorithm. The parallel genetic algorithm is implemented on high performance platform MPJ Express. Secondly, during direct implementation of the clustering algorithm. The loading test revealed two slowest stages in FRIS-Tax algorithm. They appeared to be finding of the first pillar and finding of the next pillar. To speed up these stages, the technology Streams JAVA 8 was used. For monitoring of the algorithm implementation, we developed a web interface which allows observing the current values of genetic parameters and achieved values of the fitness function. The work presents quantitative values of the process execution time demonstrating the advantage of parallel implementation of the algorithm.

References

1. Borisova I.A., Zagoruiko N.G. Functions rival similarity in the problem of taxonomy, Proc. Conf. with international participation "Knowledge - Ontology - Theory" (Umbrella-07). Novosibirsk, 2007. T. 2. P. 67-76.
2. Barakhnin V.B., Nekhaeva V.A., Fedotov A.M. On the statement of the similarity measure for the clustering of text documents, Vestn. Novosib. state. Univ. Series: Information technology. 2008. T. 6, no. 1. S. 3-9.
3. Zagoruiko N.G., Barakhnin V.B., Borisova I.A., Tkachev D.A. Clustering of text documents from an electronic database of publications algorithm FRiS-Tax, Computational technologies. - T. 18, number 6, 2013. C. 62-74.
4. Gladkov L.A. Kureichik V.V., V.M. Kureichik Genetic algorithms, Ed. V.M. Kureichik. - 2nd ed., Rev. and add. - M .: FIZMATLIT, 2006. - 320 p.

METHODS AND TOOLS OF PARALLEL PROGRAMMING

V. N. Kasyanov[1,2] and E. V. Kasyanova,[1,2]

[1] *Institute of Informatics Systems of SB RAS, Novosibirsk, Russia*
[2] *Novosibirsk State University, Novosibirsk, Russia*

Using traditional methods, it is very difficult to develop high-quality, portable software for parallel computers. In particular, parallel software for supporting of enterprise information systems cannot be developed on low-cost, sequential computers and then moved to high-performance parallel computers without extensive rewriting and debugging. Functional programming [1] is a programming paradigm, which is entirely different from the conventional model: a functional program can be recursively defined as a composition of functions where each function can itself be another composition of functions or a primitive operator (such as arithmetic operators, etc.). The programmer need not be concerned with explicit specification of parallel processes since independent functions are activated by the predecessor functions and the data dependencies of the program. This also means that control can be distributed. Further, no central memory system is inherent to the model since data is

USING NON-NEGATIVE MATRIX FACTORIZATION FOR TEXT SEGMENTATION

A. Nugumanova[1], M. Mansurova[2], Ye. Alimzhanov[2], and Ye. Baiburin[1]

1 D. Serikbayev East Kazakhstan State Technical University, Ust-Kamenogorsk, Kazakhstan
2Al-Farabi Kazakh National University, Almaty, Kazakhstan

Segmentation of text documents is one of important and interesting research problems in the sphere of natural language processing (NLP). It comes in many applications of information retrieval the functions of which provide operative and purposeful access to the full text content of digital repositories. As full text documents available in digital repositories are, as a rule, large in volume, the main function of such applications is to provide the user with the possibility of high-accuracy access to the chosen selected fragments of the document containing information relevant to the topics of his request.

The main problems to be solved in the course of the topic segmentation of the document are: 1) to divide the document into initial fragments (for example, paragraph or sentences); 2) to define topics of separated fragments or evaluate topic closely related fragments; 3) to combine closely related fragments into larger, topically homogeneous blocks. Topic blocks obtained due to segmentation are arranged according to the degree of their relevance to the user's request, and the most relevant blocks are chosen for the concluding presentation of the document (for example, in the form of the so-called snippets).

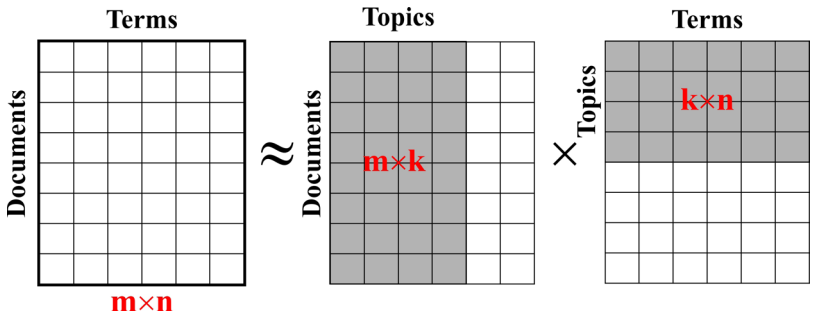

Fig. 1 Non-negative matrix factorization.

The aim of this work is to study the possibility of topic segmentation of document on the basis of non-negative matrix factorization. Non-negative matrix factorization is a method for isolation of significant factors in the data set being analysed [1]. The rows of such matrices correspond to objects (observations) and the columns – to the observed features. The algorithm of non-negative matrix factorization reduces the number of features; as a result, the basis matrix includes observations written down with the help of new larger factors and the features matrix presents weights of initial features in the composition of selected factors. The algorithm iteratively changes the elements of the indicated matrices so that their product would approximate the initial matrix as well as possible but the values of matrix elements would retain their non-negativity.

This work deals with a subject-oriented collection of documents as a data set. To use non-negative matrix factorization, the collection is presented in the form of a special matrix, the line of which corresponds to documents and columns-to term. The matrix elements present frequencies of terms usage in documents. Non-negative matrix factorization allows to execute representation of each document from the space of terms to the space of topics (Fig. 1).

The obtained as a result of factorization basic matrix contains economical representations of documents in the form of topic combinations. The matrix of features represents distribution of terms according to the topics [2]. The main idea of this work is to sort out the weights of each isolated topic according to their decrease and, in this way, determine the most important terms of this topic. It is assumed that these terms can be used as more effective supporting features during topic classification of the document fragments. This work is devoted to the study of this problem.

References

1. Lee D. D., Seung H. S. Learning the parts of objects by non-negative matrix factorization, Nature. 1999. T. 401. №. 6755. – C. 788-791.
2. Wang D., Li T., Ding C. Weighted feature subset non-negative matrix factorization and its applications to document understanding //Data Mining (ICDM), 2010 IEEE 10th International Conference on. IEEE, 2010. 541-550.

INFORMATION-THEORETIC APPROACH TO CLASSIFICATION OF SCIENTIFIC DOCUMENTS

A. Guskov[1,2,3], B. Ryabko[2,1,3], and I. Selivanova[1,3]

[1] The State Public Scientific Technological Library of SB RAS, Novosibirsk, Russia
[2] Institute of Computational Technologies of SB RAS, Novosibirsk, Russia
[3] Novosibirsk State University, Novosibirsk, Russia

Nowadays the problem of classification of scientific documents is of a great importance, because a flow of scientific documents is growing in fact exponentially. The development of methods of automatic classification of scientific documents attracts attention of many researchers over the world, see [1-6]. One of the most difficult tasks is the process of automation the thematic classification of documents, the result of which is assigning a document to one or more classes (e.g. mathematics, physics, chemistry, etc.) In spite of many efforts, an efficient automatic method for the thematic classification of scientific documents does not exist yet.

In this report we propose to use data compression methods in order to automatically determine a thematic affiliation of scientific texts. The main idea of the suggested method is quite natural: scientific texts (articles, books, etc.) use similar terminology if they belong to the same area. On the other hand, the data compressor uses frequencies of occurrence of words in the text and "compresses" the data the better, the more repeated words. Based on this observation, we suggest the following classification scheme: for any scientific area we form a set of papers, which represents the area. Then a new text is compressed together with each set of texts representing the thematic areas and refers to that area for which it is compressed to a minimum size.

For an assessment of the possible practical applications of this method, an experiment was conducted. We used data provided on the website arxiv.org to select subject domains and the formation of the texts describing them. This arxiv contains more than a million articles pertaining to various areas of science. When placing the article on the site, an author refers to the work of one of the scientific sections. The first section, pointed by author, we will call "the main category", other - "secondary".

For our experiment, we have chosen thirty research fields, presented in the arxiv (For example, information theory, logic in computer science, artificial Intelligence, cryptography