ISSN 0146-4116, Automatic Control and Computer Sciences, 2007, Vol. 41, No. 6, pp. 320–331. © Allerton Press, Inc., 2007.
Original Russian Text © R.M. Alguliev, R.M. Alyguliev, F.F. Yusifov, 2007, published in Avtomatika i Vychislitel'naya Tekhnika, 2007, No. 6, pp. 32–17.

Automatic Identification of the Interests of Web Users

R. M. Alguliev, R. M. Alyguliev, and F. F. Yusifov

Institute of Information Technologies, National Academy of Sciences of Azerbaijan,
ul. F. Agaev 9, Baku, Az-1141 Azerbaijan
e-mail: rasim@science.az

Received June 14, 2007

Abstract—In the present article an approach to automatic determination of a user's sphere of interests is proposed. The approach is based on a method involving clustering of documents which the user is interested in. The process of clustering of documents is reduced to a problem of discrete optimization for which quadratic- and linear-type models are proposed. Identification of interests makes it possible to determine the context of a request without any effort on the user's part. Different methods are proposed for determining the context of a request. An ant algorithm for solving a quadratic-type discrete optimization problem is also proposed in the present study.

DOI: 10.3103/S0146411607060041

Key words: Web usage mining, user interest, clustering, personified search, context of request, ant algorithm

1. INTRODUCTION

With the appearance of the World Wide Web the Internet turned into one of the basic sources of information. In sending a request to search engines, the user wishes to find information which he (or she) requires. An information search consists in a set of operations that are needed to find information that satisfies a user's demand. The search for required information on the Internet is becoming increasingly more difficult with the increase in Web resources and the increasing dispersal of these resources. Moreover, the efficiency of a search is estimated not only by the search time, but also by the degree of relevance, that is, the extent to which the selected documents correspond to a user's request.

Millions of search operations are executed on Google-, Yahoo-, etc. type search engines each day. Despite their popularity, search engines nevertheless possess a number of drawbacks. First, search engines present a vast quantity of documents in response to a search request and, second, the documents that are presented in the leading part of the list usually do not correspond to the user's needs. In other words, the documents that are presented are ranked without regard to the user's needs. In fact, the user is generally interested in only a few of the results obtained in response to his request. A request, in fact, constitutes a description of information which the user wishes to obtain access to. To obtain relevant information the user sometimes utilizes more complex forms of representation of requests. But this naturally complicates the procedure used to process the requests and, consequently, increases the search time.

Traditional search engines generally treat a user request as a search for information in isolation from all the user's other requests and do not utilize previously obtained results. The information demand of two users who have transmitted the same request may be different. But the documents and the rankings that are presented by the search engines are always the same regardless of which user transmitted a particular request. Thus, to increase search efficiency it is necessary to develop methods for automatic determination of the user's objective (problem) and to discover semantic relationships between users, the problems of users, and Web pages. For this purpose, a system based on the PLSA (Probabilistic Latent Semantic Analysis) method was developed in [1] for the identification and analysis of navigation patterns.

The results of a search may be improved significantly once the context of a request is known [2]. A number of different tools for the determination of the context of a request have been developed. For example, in the Inquirus project [3], which is under development at the NEC Research Institute, context information is specified explicitly in the form of a specification of the category of the data which the user is requesting. Here context information is used for selecting particular search engines that are then applied to transmit the request, to modify the request, and to determine principles for use in ranking the documents produced. It is well known that the Inquirfus project requires that the user indicate explicitly the context in order to achieve increased precision. The Watson System [4] automatically determines the context of a request. Relying on

the content of documents which the user has previously edited by means of Microsoft Word or scanned by Internet Explorer, the system simulates the context of the request. In order to specify the context of a request as a means of increasing the relevance of search results, an ontological approach to the description of the application domain the user is interested in was proposed in [5]. The user himself formulates this ontology and then specifies the context of the request explicitly by means of the ontology. Researchers are now working on a more complex system that will constantly collect data about the user and may even help predict how the user's interests will change in the future. A new approach to the classification of navigation patterns and prediction of future user requests was proposed in [6] using user profiles. The user's navigation profile is extracted by means of a combined analysis of the Web data log and the content of Web pages.

The objective of a search personification on the Web is to take into account the user's information needs in the search process [7–11]. In personalization of the search process, data on previous requests and on the user's sphere of interests are utilized. Several different approaches have been proposed as ways of automatically creating a user's personal settings [10, 11]. For example, in [10] the user parameters are modified through the accumulation of personal settings reflected in the chronology of a scan. And in [11], the user profile is represented by a tree of hierarchical categories and corresponding key words are associated with each category. The user profile is created automatically, through a study of the chronology of the user's search.

Many search mechanisms on the Web are focused on the analysis of Web structure (hyperlink). For example, a Google search engine computes a universal PageRank vector to determine the relative degree of significance of different Web pages. Personified variants of the PageRank algorithm have been proposed for the purpose of personalizing a Web search. For example, the Topic Sensitive PageRank (TSPR) algorithm was proposed in [12]. In place of a single global PageRank indicator, the TSPR algorithm computes for each Web page a PageRank vector, where each element of the vector corresponds to the PageRank indicator of the Web page with respect to a particular topic. Before computing the PageRank indicator of Web pages, in the TSPR algorithm the Web pages are first grouped according to topic and the PageRank indicator of each of the pages is then computed for each topical section. The TSPR algorithm does not utilize any user context information whatsoever in computing the PageRank indicator, and it is therefore difficult to evaluate whether the results obtained satisfy the user's information needs. A personified PageRank algorithm, called UPR (Usage Based PageRank), was proposed in [13] as a way of taking into account the user's information needs. Based on the user's previous navigational behavior, the UPR algorithm constructs a personified oriented navigational graph to a particular Web page, and on the basis of this graph produces a ranking of Web pages.

A search engine that utilizes the HITS algorithms (Hypertext Induced Topic Search) and a fuzzy concept network was proposed in [14] as a means of finding relevant documents corresponding to the user's needs. Through application of the HITS algorithm, the search engine, in response to a user request, first computes the ranks of documents as "authority" and "hub," and then personifies the search results relative to the user's interests. The search engine finds those relevant documents which the user is interested in and then reorders them relative to the user's interests.

In the present study an approach involving automatic determination of the user's sphere of interests is proposed on the basis of scanning chronology. The proposed approach makes it possible to determine the context of a request without any effort on the user's part.

2. AUTOMATIC DETERMINATION OF A USER'S SPHERE OF INTERESTS

To determine the user's sphere of interests, i.e., the user's specific information needs, it seems best to incorporate information about the user's navigational strategy on Web pages and the results of a content analysis of these pages. In this section the user's sphere of interests is determined on the basis of a content analysis of documents which the user has previously edited or scanned. The proposed approach is based on the notion of clustering of documents.

2.1. Statement and Mathematical Models of Clustering Problem

The process of Web usage mining consists of three stages:

- collection and processing (or transformation) of data;
- identification of images;
- image analysis [15, 16].

On the first stage, a "raw" Web log of data that have accumulated on Web servers is transformed into operational data; these data may be processed by means of different methods of data mining. On the second

stage, images (or profiles) are identified by means of data mining methods. Finally, on the third stage, an analysis or interpretation of the identified images is undertaken in different applications, for example, in a personification of a Web search.

Clustering of documents, a technique that makes it possible to discover hidden relations between the elements of documents, is the most appropriate method for use in determining a user's sphere of interests. Clustering is employed in the most diverse domains, for example, in image and signal recognition, in marketing studies, in medicine, biology, physics, and elsewhere. Processing of text documents, for example, determining the semantic proximity of documents, is among the most important fields of application of the clustering method.

Suppose $\mathbf{D} = \{d_1, d_2, \ldots, d_n\}$ is a set of documents which the user has previously edited or scanned. The problem of clustering of the set $\mathbf{D} = \{d_1, d_2, \ldots, d_n\}$ consists in finding a partition of this set into subsets $\mathbf{C} = \{C_1, C_2, \ldots, C_k\}$, called clusters, such that $C_1 \cup C_2 \cup \ldots \cup C_k = \{d_1, d_2, \ldots, d_n\}$, $2 < k < n$, $C_p \cap C_q = \varnothing$, $p \neq q$ and such that the elements of each cluster C_p, $p = 1, \ldots, k$, are topically close to each other.

Let us now pass on to a mathematical statement of the problem of clustering, which requires representing the documents in an appropriate form. The most widely used method used to represent documents is the Vector Space Model. The vector space model is a classical representative of the class of algebraic models. Suppose that $\mathbf{T} = \{T_1, T_2, \ldots, T_m\}$ is a set of terms that occur in the set $\mathbf{D} = \{d_1, d_2, \ldots, d_n\}$ and a set of user requests. Within the framework of the model, some nonnegative weight w_{it} is compared with each term T_t in document d_i. Thus, each document is represented in the form of an m-dimensional vector $d_i = (w_{i1}, w_{i2}, \ldots, w_{im})$, $i = 1, 2, \ldots, n$. The weight of a term may be computed by any one of a number of different methods. The most common method is that of TF*IDF (Term Frequency * Inverse Document Frequency). The TF*IDF scheme determines the discriminant force of a term, i.e., how frequently a given term is used in other documents in the particular set. The TF*IDF scheme works better if statistics on the usage of terms broken down by set are available. The scheme assigns a weight to a term T_t in document d_i that is proportional to the number of occurrences of the term in the document and inversely proportional to the number of documents in the set $\mathbf{D} = \{d_1, d_2, \ldots, d_n\}$ in which the term is encountered at least once:

$$w_{it} = tf_{it}\log\left(\frac{n}{n_t}\right), \tag{1}$$

where tf_{it} is the number of occurrences of term T_t in document d_i; n, total number of documents, and n_t, number of documents in which term T_t is encountered.

In order to perform a clustering of the set $\mathbf{D} = \{d_1, d_2, \ldots, d_n\}$ of documents it is necessary to define a measure of proximity between the elements of the documents. The cosine metric is among the most commonly employed measures for text documents. The cosine metric between pairs of documents (d_i, d_j) is defined thus:

$$\cos(d_i, d_j) = \frac{\sum_{t=1}^{m} w_{it}w_{jt}}{\sqrt{\sum_{t=1}^{m} w_{it}^2}\sqrt{\sum_{t=1}^{m} w_{jt}^2}}, \quad i, j = 1, \ldots, n. \tag{2}$$

The quality of a clustering depends directly on the choice of criterion function. The choice of a particular type of criterion function determines the quality of the clustering. The criterion function must be determined in such a way that it is in conformity with the basic principle of the clustering problem, i.e., all the documents within the same cluster are similar to each other, in contrast to documents in different clusters. A minimal degree of similarity of documents assigned to different clusters and maximal proximity of documents within the same cluster will be assumed to be the optimality criterion.

Remark. One of the basic problems that arises in clustering of text documents is the high dimension of the attribute space. Thus, the problem of reducing the dimension assumes special importance. Moreover, a dictionary of stop words is usually used; such a dictionary includes often used words and, in certain cases, isolation of the roots of a word (stemming) is additionally employed to achieve a further reduction.

2.2. Quadratic Model

Suppose that the binary variable $x_{ip} \in \{-1, 1\}$ such that

$$x_{ip} = \begin{cases} -1, & d_i \notin C_p; \\ 1, & d_i \in C_p; \end{cases} \quad p = 1, 2, \ldots, k \tag{3}$$

corresponds to document $d_i \in \mathbf{D}$ ($i = 1, \ldots, n$).

We let $c_{ij}^{+} = \cos(d_i, d_j)$ if d_i and d_j belong to the same cluster, otherwise, if d_i and d_j belong to different clusters, we set $c_{ij}^{-} = \cos(d_i, d_j)$. By (2), $c_{ij}^{+} = c_{ji}^{+}$ and $c_{ij}^{-} = c_{ji}^{-}$.

In view of the above notation and the assumptions, we will formulate the mathematical model of the clustering problem by means of a quadratic-type problem:

$$f(x) = \sum_{p=1}^{k} \sum_{q=1}^{k} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (1 - x_{ip}x_{jq})c_{ij} - 2(k-1) \sum_{p=1}^{k} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (1 - x_{ip}x_{jp})c_{ij}^{-} \longrightarrow \max \tag{4}$$

under the constraints

$$\sum_{p=1}^{k} x_{ip} = 2 - k, \quad \forall i = 1, 2, \ldots, n, \tag{5}$$

$$x_{ip}^2 = 1 \quad \forall i = 1, 2, \ldots, n, \quad \forall p = 1, 2, \ldots, k, \tag{6}$$

where $c_{ij} = \begin{cases} c_{ij}^{+} \text{ if } d_i \text{ and } d_j \text{ belong to the same cluster;} \\ c_{ij}^{-} \text{ if } d_i \text{ and } d_j \text{ belong to different clusters.} \end{cases}$

Problem (4)–(6) is a multi-extremum nonlinear-type problem in which the function (4) and the constraints (6) are quadratic "homogeneous" functions, i.e., they do not contain linear terms that are functions of the variables x_{ip}, $i = 1, 2, \ldots, n$; $p = 1, 2, \ldots, k$, while the constraints (5) are linear. Constraints (5) guarantee that there are no common points among the clusters. This follows directly from the definition of the variables x_{ip}. By Definition (3), variable x_{ip} assumes a value equal to 1 once and assumes a value equal to –1 a total of $(k - 1)$ times, and, consequently,

$$\sum_{p=1}^{k} x_{ip} = 1 + (k-1)(-1) = 1 - k + 1 = 2 - k.$$

Despite the fact that the first term in (4) contains both weights c_{ij}^{+} and c_{ij}^{-} while the second term contains only the weight c_{ij}^{-}, ultimately the two weights c_{ij}^{+} and c_{ij}^{-} occur in the objective function (4) with the same coefficient but with opposite sign. In other words, the objective function (4) balances two criteria, intra-cluster similarity and inter-cluster difference. This follows directly from the following assertion.

Proposition. For any pair (i, j) the weight coefficient c_{ij}, i.e., the weights c_{ij}^{+} and c_{ij}^{-}, in the first term of the objective function (4) is equal to $4(k - 1)$, while the weight coefficient c_{ij}^{-} in the second term is equal to $8(k - 1)$.

Proof. Let us first prove the first part of the Proposition.

Let $d_i \in C_{p_0}$ and $d_j \in C_{q_0}$ $(p_0 \neq q_0)$, i.e., $x_{ip_0} = 1$ and $x_{jq_0} = 1$.

Then, by definition, for all $p \neq p_0 = 1, 2, \ldots, k$, it turns out that $x_{ip} = -1$ while for all $q \neq q_0 = 1, 2, \ldots, k$, $x_{jq} = -1$. Hence, it follows that with $q \neq q_0 = 1, 2, \ldots, k$ we have $(1 - x_{ip_0}x_{jq}) = 2$, while with $q = q_0$ we have $(1 - x_{ip_0}x_{jq}) = 0$. Analogously, we find that with $p \neq p_0 = 1, 2, \ldots, k$, it turns out that $(1 - x_{ip}x_{jq_0}) = 2$, and with $p = p_0$ we have $(1 - x_{ip}x_{jq_0}) = 0$.

Expanding the sums with respect to p and q, we have

$$\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} c_{ij}\sum_{p=1}^{k}\sum_{q=1}^{k}(1-x_{ip}x_{jq}) = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} c_{ij}\Bigg\{\sum_{q=1}^{k}(1-x_{i1}x_{jq}) + \ldots + \sum_{q=1}^{k}(1-x_{i,\,p_0-1}x_{jq}) + \sum_{q=1}^{k}(1-x_{ip_0}x_{jq}) + \sum_{q=1}^{k}(1-x_{i,\,p_0+1}x_{jq}) + \ldots + \sum_{q=1}^{k}(1-x_{ik}x_{jq})\Bigg\}$$

$$= \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} c_{ij}\{[(1-x_{i1}x_{j1}) + \ldots + (1-x_{i1}x_{j,\,q_0-1}) + (1-x_{i1}x_{jq_0}) + (1-x_{i1}x_{j,\,q_0+1}) + \ldots + (1-x_{i1}x_{jk})] + \ldots + [(1-x_{i,\,p_0-1}x_{j1}) + \ldots + (1-x_{i,\,p_0-1}x_{j,\,q_0-1}) + (1-x_{i,\,p_0-1}x_{jq_0}) + (1-x_{i,\,p_0-1}x_{j,\,q_0+1}) + \ldots + (1-x_{i,\,p_0-1}x_{jk})] + [(1-x_{ip_0}x_{j1}) + \ldots + (1-x_{ip_0}x_{j,\,q_0-1}) + (1-x_{ip_0}x_{jq_0}) + (1-x_{ip_0}x_{j,\,q_0+1}) + \ldots + (1-x_{ip_0}x_{jk})] + [(1-x_{i,\,p_0+1}x_{j1}) + \ldots + (1-x_{i,\,p_0+1}x_{j,\,q_0-1}) + (1-x_{i,\,p_0+1}x_{jq_0}) + (1-x_{i,\,p_0+1}x_{j,\,q_0+1}) + \ldots + (1-x_{i,\,p_0+1}x_{jk})] + \ldots + [(1-x_{i,k}x_{j1}) + \ldots + (1-x_{i,k}x_{j,\,q_0-1}) + (1-x_{i,k}x_{jq_0}) + (1-x_{i,k}x_{j,\,q_0+1}) + \ldots + (1-x_{i,k}x_{jk})]\} \tag{7}$$

$$= \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} c_{ij}\{[\underbrace{2+\ldots+2}_{(p_0-1)\text{ times}}] + 2(k-1) + [\underbrace{2+\ldots+2}_{(k-p_0)\text{ times}}]\}$$

$$= \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} c_{ij}\{2(p_0-1+k-1+k-p_0)\} = 4(k-1)\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} c_{ij}.$$

The first part of the Proposition is proved. The second part of the Proposition is proved in analogous fashion.

Without loss of generality, we may suppose that $p_0 < q_0$. In the second term in (4), we expand the sum with respect to p:

$$\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} c_{ij}^{-}\sum_{p=1}^{k}(1-x_{ip}x_{jp}) = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} c_{ij}^{-}\{(1-x_{i1}x_{j1}) + \ldots + (1-x_{ip_0}x_{jp_0}) + \ldots + (1-x_{iq_0}x_{jq_0}) + \ldots + (1-x_{ik}x_{jk})\}. \tag{8}$$

Since $x_{ip_0} = 1$ and $x_{jq_0} = 1$, we find that only two of the terms within the braces are nonzero, i.e., $(1 - x_{ip_0}x_{jp_0}) = 2$ and $(1 - x_{iq_0}x_{jq_0}) = 2$, while the other terms are equal to zero. Thus,

$$\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} c_{ij}^{-}\sum_{p=1}^{k}(1-x_{ip}x_{jp}) = 4\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} c_{ij}^{-}, \tag{9}$$

whence follows the proof of the second part of the Proposition. Thus, the Proposition is completely proved.

Substituting formulas (7) and (9) into (4), we find

$$f(x) = 4(k-1)\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} c_{ij} - 8(k-1)\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} c_{ij}^{-} = 4(k-1)\Bigg\{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} c_{ij}^{+} - \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} c_{ij}^{-}\Bigg\}. \tag{10}$$

By introducing new variables and applying functionally redundant constraints [17], the quadratic model (4)–(6) may be reduced to a linear-type model.

2.3. Linear Model

Let $y_{ipjq} = x_{ip}x_{jq}$ be new variables for all i, j, p, and q such that $1 \le i < j \le n$ and $1 < p, q \le k$. The number of new variables y is equal to $\frac{n(n-1)}{2}k^2$.

The new model expressed in terms of the new variables y for the clustering problem will be formulated by means of a quadratic-type problem,

$$f(y) = \sum_{p=1}^{k}\sum_{q=1}^{k}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(1-y_{ipjq})c_{ij} - 2(k-1)\sum_{p=1}^{k}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(1-y_{ipjp})\bar{c}_{ij} \longrightarrow \max, \tag{11}$$

under the constraints

$$\sum_{p=1}^{k}\sum_{q=1}^{k} y_{ipjq} = (k-2)^2, \quad \forall (i, j): 1 \le i < j \le n, \tag{12}$$

$$y_{ipjq}^2 = 1, \quad \forall i, j, p, q: 1 \le i < j \le n \quad \text{and} \quad 1 < p, \quad q \le k, \tag{13}$$

$$y_{ipjq} - y_{iplr}y_{jqlr} = 0, \quad \forall i, j, p, q: 1 \le i < j \le n \quad \text{and} \quad 1 < p, \quad q \le k, \tag{14}$$

$$y_{iplr} - y_{ipjq}y_{jqlr} = 0, \quad \forall i, j, p, q: 1 \le i < j \le n \quad \text{and} \quad 1 < p, \quad q \le k, \tag{15}$$

$$y_{jqlr} - y_{ipjq}y_{iplr} = 0, \quad \forall i, j, p, q: 1 \le i < j \le n \quad \text{and} \quad 1 < p, \quad q \le k. \tag{16}$$

Constraints (12)–(16) have the following sense. Constraints (12) are analogs of constraints (5) expressed in terms of the new variables y. Constraints (13) correspond to the binarity of the variables y and follow from the rules under which they were introduced, since by (6), the following equality is always valid for any i, j, p, and q such that $1 \le i < j \le n$, $1 < p$, and $q \le k$:

$$y_{ipjq}^2 = (x_{ip}x_{jq})^2 = x_{ip}^2x_{jq}^2 = 1. \tag{17}$$

Constraints (14)–(16) specify functionally redundant constraints that follow from the definition of the new variables y and Constraints (6),

$$y_{ipjq} = x_{ip}x_{jq}(x_{lr})^2 = (x_{ip}x_{lr})(x_{jq}x_{lr}) = y_{iplr}y_{jqlr}, \tag{18}$$

$$y_{iplr} = x_{ip}x_{lr}(x_{jq})^2 = (x_{ip}x_{jq})(x_{jq}x_{lr}) = y_{ipjq}y_{jqlr}, \tag{19}$$

$$y_{jqlr} = x_{jq}x_{lr}(x_{ip})^2 = (x_{ip}x_{jq})(x_{ip}x_{lr}) = y_{ipjq}y_{iplr}, \tag{20}$$

which are valid for arbitrary indices i, j, l, p, q, and r, $1 \le i < j < l \le n$, $1 < p, q, r \le k$.

Let us show that a linear model may be obtained from the model (11)–(16). The following relations are obtained from Constraints (14)–(16):

$$2y_{iplr}y_{jqlr} + 2y_{ipjq}y_{jqlr} + 2y_{ipjq}y_{iplr} - 2(y_{ipjq} + y_{iplr} + y_{jqlr}) = 0, \tag{21}$$

$$2y_{iplr}y_{jqlr} - 2y_{ipjq}y_{jqlr} - 2y_{ipjq}y_{iplr} - 2(y_{ipjq} - y_{iplr} - y_{jqlr}) = 0, \tag{22}$$

$$-2y_{iplr}y_{jqlr} + 2y_{ipjq}y_{jqlr} - 2y_{ipjq}y_{iplr} - 2(-y_{ipjq} + y_{iplr} - y_{jqlr}) = 0, \tag{23}$$

$$-2y_{iplr}y_{jqlr} - 2y_{ipjq}y_{jqlr} + 2y_{ipjq}y_{iplr} - 2(-y_{ipjq} - y_{iplr} + y_{jqlr}) = 0. \tag{24}$$

Relations (21)–(24) are obtained by multiplying the four combinations constructed from Constraints (14)–(16) by 2. Relation (21) is obtained as a result of adding all the constraints in (14)–(16). Relation (22) is obtained as a result of subtracting Constraints (15) and (16) from Constraint (14), while (23) is obtained as a result of subtracting Constraints (14) and (16) from Constraint (15). Finally, relation (24) is obtained as a result of subtracting Constraints (14) and (15) from Constraint (16).

In light of (12), relation (21) may be rewritten thus:

$$y_{ipjq}^2 + y_{iplr}^2 + y_{jqlr}^2 + 2y_{iplr}y_{jqlr} + 2y_{ipjq}y_{jqlr} + 2y_{ipjq}y_{iplr} - 2(y_{ipjq} + y_{iplr} + y_{jqlr}) - 3 = 0 \tag{25}$$

or

$$(y_{ipjq} + y_{iplr} + y_{jqlr})^2 - 2(y_{ipjq} + y_{iplr} + y_{jqlr}) + 1 - 4 = 0, \tag{26}$$

$$(y_{ipjq} + y_{iplr} + y_{jqlr} - 1)^2 = 4. \tag{27}$$

From the latter equality it follows that

$$y_{ipjq} + y_{iplr} + y_{jqlr} = 3 \tag{28}$$

or

$$y_{ipjq} + y_{iplr} + y_{jqlr} = -1. \tag{29}$$

Thus, the following linear inequality may be written for the sum of the variables y_{ipjq}, y_{iplr}, and y_{jqlr}:

$$y_{ipjq} + y_{iplr} + y_{jqlr} \geq \min\{-1, 3\} = -1. \tag{30}$$

Analogous arguments hold for relations (22)–(24). As a result, we have the equalities

$$(y_{ipjq} - y_{iplr} - y_{jqlr} - 1)^2 = 4, \tag{31}$$

$$(-y_{ipjq} + y_{iplr} - y_{jqlr} - 1)^2 = 4, \tag{32}$$

$$(-y_{ipjq} - y_{iplr} + y_{jqlr} - 1)^2 = 4, \tag{33}$$

from which follow, as previously, linear constraints of a similar type:

$$y_{ipjq} - y_{iplr} - y_{jqlr} \geq \min\{-1, 3\} = -1, \tag{34}$$

$$-y_{ipjq} + y_{iplr} - y_{jqlr} \geq \min\{-1, 3\} = -1, \tag{35}$$

$$-y_{ipjq} - y_{iplr} + y_{jqlr} \geq \min\{-1, 3\} = -1. \tag{36}$$

Thus, a new linear model is obtained by means of simple computations from the quadratic model (11)–(16):

$$f(y) = \sum_{p=1}^{k}\sum_{q=1}^{k}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(1 - y_{ipjq})c_{ij} - 2(k-1)\sum_{p=1}^{k}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(1 - y_{ipjp})c_{ij}^{-} \longrightarrow \max \tag{37}$$

under the constraints

$$y_{ipjq} + y_{iplr} + y_{jqlr} \geq -1, \quad \forall i, j, p, q\colon 1 \leq i < j \leq n \quad \text{and} \quad 1 < p, \quad q \leq k, \tag{38}$$

$$y_{ipjq} - y_{iplr} - y_{jqlr} \geq -1, \quad \forall i, j, p, q\colon 1 \leq i < j \leq n \quad \text{and} \quad 1 < p, \quad q \leq k, \tag{39}$$

$$-y_{ipjq} + y_{iplr} - y_{jqlr} \geq -1, \quad \forall i, j, p, q\colon 1 \leq i < j \leq n \quad \text{and} \quad 1 < p, \quad q \leq k, \tag{40}$$

$$-y_{ipjq} - y_{iplr} + y_{jqlr} \geq -1, \quad \forall i, j, p, q\colon 1 \leq i < j \leq n \quad \text{and} \quad 1 < p, \quad q \leq k, \tag{41}$$

$$\sum_{p=1}^{k}\sum_{q=1}^{k} y_{ipjq} = (k-2)^2, \quad \forall (i, j)\colon 1 \leq i < j \leq n. \tag{42}$$

In the linear model (37)–(42) the new variable $y_{ipjq} = x_{ip}x_{jq}$ denotes nothing other than the binary variable

$$y_{ipjq} = \begin{cases} -1 & \text{if} \quad d_i \in C_p, \quad d_j \notin C_q \quad \text{or} \quad d_i \notin C_p, \quad d_j \in C_q \\ 1 & \text{if} \quad d_i \in C_p, \quad d_j \in C_q \quad \text{or} \quad d_i \notin C_p, \quad d_j \notin C_q \end{cases} \tag{43}$$

which follows from the informal meaning of the variables x_{ip} and x_{jq}.

It was proved in [17] that functionally redundant constraints improve the upper bound of the solution of the problem.

To avoid repeated clustering, i.e., to assure the adaptability of the proposed approach, each new element d_{n+1} is classified by means of the k-Nearest Neighbor (kNN) method into one of the classes (topical sections). The kNN method assigns a relevance count to each topical section C_p, using the following formula:

$$\text{score}(d_{n+1}, C_p) = \sum_{d \in kNN(d_{n+1})} \text{sim}_{\cos}(d_{n+1}, d),$$

where $k\text{NN}(d_{n+1})$ denotes the set of k-nearest neighbors of document d_{n+1} in cluster C_p. Document d_{n+1} belongs to the topical section C_p for which score (d_{n+1}, C_p) is of maximal value.

If there are no documents in the set of documents that are similar to any of the new documents, the new documents are placed in a new topical section.

3. DETERMINATION OF THE CONTEXT OF A REQUEST

In the present section methods of determining the context of a request will be given.

Method 1. The probability of membership of a request to a topical section is calculated. The probability $P(C_p|Q)$ that request Q belongs to topical section C_p may be determined from the formula

$$P(C_p|Q) = \sum_{t=1}^{m} P(C_p|Q, \text{“}T = T_t\text{”}) P(\text{“}T = T_t\text{”}|Q), \tag{44}$$

where “$T = T_t$” denotes that the term T that has been selected randomly from request Q is equal to T_t.

In the latter formula we obtain, using Bayes' formula, assuming that C_p and Q are independent,

$$P(C_p|Q) = P(C_p) \sum_{t=1}^{m} \frac{P(\text{“}T = T_t\text{”}|C_p) P(\text{“}T = T_t\text{”}|Q)}{P(\text{“}T = T_t\text{”})}. \tag{45}$$

The probabilities present on the right-hand side of formula (45) are defined as follows:

- $P(\text{“}T = T_t\text{”}|Q) = \dfrac{tf(Q, T_t)}{tf(Q)}$ – relative frequency of occurrence of term T_t in request Q;
- $P(\text{“}T = T_t\text{”}|C_p) = \dfrac{tf(C_p, T_t)}{tf(C_p)}$ – relative frequency of occurrence of term T_t in cluster C_p;
- $P(\text{“}T = T_t\text{”}|C_p) = \dfrac{tf(\mathbf{D}, T_t)}{tf(\mathbf{D})}$ – relative frequency of term T_t in collection of documents $\mathbf{D}$;
- $P(C_p) = \dfrac{|C_p|}{n}$ – relative frequency of occurrence of documents in cluster C_p,

where $tf(Q)$ is the total number of terms in request Q; $tf(Q, T_t)$, number of occurrences of term T_t in request Q; $tf(C_p)$, total number of terms in cluster C_p; $tf(C_p, T_t)$, number of occurrences of term T_t in cluster C_p; $tf(\mathbf{D})$, total number of terms in collection of documents $\mathbf{D}$; $tf(\mathbf{D}, T_t)$, number of occurrences of term T_t in collection of documents $\mathbf{D}$; and $|C_p|$, number of documents belonging to cluster C_p, $p = 1, 2, \ldots, k$.

Method 2. Before determining the context of a request, we will first calculate the mean weight of the terms in each topical section. The mean weight W_{pt}^{avg} of term T_t in cluster C_p is computed thus:

$$W_{pt}^{avg} = \frac{1}{|C_p|} \sum_{d_i \in C_p} \frac{w_{it}}{\text{len}(d_i)} = \frac{1}{2|C_p|} \sum_{i=1}^{n} \frac{w_{it}(1 + x_{ip})}{\text{len}(d_i)}, \tag{46}$$

where $\text{len}(d_i) = \sum_{t=1}^{m} f_{it}$ is the length (number of terms) of document d_i.

Then the weight of request Q in topical section C_p will be defined as the sum of the mean weights of terms encountered in request Q:

$$W(Q, C_p) = \sum_{T_i \in Q} W_{pt}^{avg} = \frac{1}{2|C_p|} \sum_{T_t \in Q} \sum_{i=1}^{n} w_{it}(1 + x_{ip}). \quad (47)$$

Method 3. In this method the context of a request is computed as being equal to the mean measure of proximity between request Q and documents belonging to cluster C_p:

$$\mathrm{sim}_{\cos}(Q, C_p) = \frac{1}{|C_p|} \sum_{d_i \in C_p} \cos(Q, d_i) = \frac{1}{2|C_p|} \sum_{i=1}^{n} \cos(Q, d_i)(1 + x_{ip}), \quad p = 1, 2, \ldots, k. \quad (48)$$

Method 4. The degree of proximity of request Q to topical section C_p may be defined as the measure of proximity between Q and C_p, on the one hand, and the center O_p of cluster C_p, on the other hand:

$$\mathrm{sim}_{\cos}(Q, C_p) = \cos(Q, O_p), \quad p = 1, 2, \ldots, k. \quad (49)$$

Method 5. The χ^2 criterion is a classical criterion used to establish the existence of a relationship between a request Q and a topical section C_p. The criterion is usually used to identify informational attributes (terms) in a problem involving classification of text documents [18].

Let N_{11} be the number of cases in which $T \in Q$ and C_p are encountered together; N_{10}, number of cases in which $T \in Q$ is encountered without C_p; N_{01}, number of cases in which C_p is encountered without $T \in Q$; N_{00}, number of cases in which neither $T \in Q$ nor C_p is encountered. The degree of proximity between term $T \in Q$ and cluster C_p is then calculated in the form

$$\chi^2(T \in Q, C_p) = n \frac{(N_{11}N_{00} - N_{10}N_{01})^2}{(N_{11} + N_{10})(N_{01} + N_{00})(N_{11} + N_{01})(N_{10} + N_{00})}. \quad (50)$$

Then the context of a request may be defined in the following way:

$$\chi^2(Q, C_p) = \sum_{T \in Q} \chi^2(T \in Q, C_p), \quad p = 1, 2, \ldots, k. \quad (51)$$

If the quantities determined by formulas (45), (47), (48), (49), and (51) are large, this will mean that request Q belongs to the p-th topical section.

4. ESTIMATION OF QUALITY OF CLUSTERING AND OPTIMAL NUMBER OF CLUSTERS

One of the most difficult parts of a clustering problem is to determine the number of clusters. Note that at the start the number of clusters may not be known, since a priori information about the data and about the nature of the data may be lacking. Therefore, the number of clusters is usually selected on the basis of the conditions of the particular problem. After performing a partitioning if necessary, some of the clusters may be merged together or partitioned and the clustering process repeated now with a new number of clusters. The following algorithm may be proposed for the purpose of determining the optimal number of clusters. An index for estimating the quality of a clustering is first introduced. Let $\mathrm{Sep}(k) = \frac{1}{k(k-1)} \sum_{p=1}^{k-1} \sum_{q=p+1}^{k} \cos(O_p, O_q)$, where O_p is the center of cluster C_p. A minimal value of the quantity $\mathrm{Sep}(k)$ shows that the clusters are dispersed from each other. The quantity $\mathrm{Comp}(k) = \sum_{p=1}^{k} \frac{1}{|C_p|} \sum_{d_i \in C_p} \cos(d_i, O_p)$ is introduced as a means of determining the degree of compactness of the clusters. If $\mathrm{Comp}(k)$ has a high value, this means that the documents are compact around the corresponding centers. An index for estimating the quality of a clustering will be defined by the ratio of the two quantities $\mathrm{Sep}(k)$ and $\mathrm{Comp}(k)$, thus: $V_{\mathrm{index}}(k) = \frac{\mathrm{Sep}(k)}{\mathrm{Comp}(k)}$. If minimized, the latter index ensures both compactness and remoteness of the clusters. Obviously, with increasing number of clusters the compactness of each of the clusters grows, while the centers of the clusters become more remote from the center of the collection of documents. In other words, the estimation index $V_{\mathrm{index}}(k)$ is a nondecreasing function of the parameter k, $V_{\mathrm{index}}(k) \geq V_{\mathrm{index}}(k+1)$. The following technique may then be proposed for determining the number of clus-

ters. Suppose a degree of precision $\varepsilon > 0$ is given. If for some number k^*, the condition $\frac{V_{\text{index}}(k^*) - V_{\text{index}}(k^*+1)}{V_{\text{index}}(k^*)} < \varepsilon$ holds, it is adopted as the optimal number of clusters. A choice of a particular value of the parameter ε may lead to different values for the number of clusters. A larger value of ε may lead to the appearance of large clusters that combine other clusters, while a small value of ε may lead to the appearance of small clusters, i.e., artificial clusters. The choice of the value of the parameter ε depends on the nature of the data, i.e., on the degree of topical diversity of the documents.

5. ANT ALGORITHM FOR USE IN SOLVING THE CLUSTERING PROBLEM

In the general case, problem (4)–(6) is *NP*-hard [19]. As a result meta-heuristic methods are of interest as tools for the solution of the problem. In this section an algorithm based on the meta-heuristic of an ant colony will be proposed for the solution of problem (4)–(6). As a multi-agent approach, ant algorithms were first proposed in [20] for the purpose of solving the traveling salesman problem. Ant algorithms appeared as a result of an analysis of the behavior of an actual ant colony. The principal advantage of "ant colony" algorithms (simply, ant algorithms) lies in parallel search for the solutions of given problems. As a meta-heuristic method, ant algorithms have been used to solve a number of complex problems of combinatorial optimization. Such problem as vehicular transport routing, quadratic assignment, calendar scheduling, and the graph coloring problem are among examples that illustrate the use of ant algorithms for the solution of problems of combinatorial optimization [21].

In ant colony algorithms a group of artificial agents iteratively construct feasible solutions on the basis of their cooperative interaction. Transmission of information between agents is achieved through the exchange of food (direct path) or stigmergy (indirect path). Biologically speaking, stigmergy is realized through pheromones, a special substance that is deposited on the trail as the ant travels. The construction of solutions is performed on the basis of virtual trails of pheromones that are specialized for the problem of heuristic information. Through successive assignment of documents to appropriate clusters $\mathbf{C} = (C_1, C_2, \ldots, C_k)$, each ant constructs its own feasible solution. An assignment of document d_i to cluster C_p is denoted by the pair (i, p). For each ant the process is repeated until all the documents have been assigned. A stochastic path of assignment is employed by the ants on each step. For the s-th ant, the probability P_{ip}^s that document d_i is assigned to cluster C_p depends on the following three quantities:

- local information η_{ip}, computed on the basis of heuristic information and demonstrating the utility of the given assignment (i, p);
- level of pheromone τ_{ip}, showing how successful was the application of the given assignment (i, p);
- taboo list, denoting the set of documents which may not be repeatedly assigned.

The taboo list grows once the documents are assigned and is deleted at the start of each iteration. We denote by A_p^s the set of documents which the s-th ant must assign to one of the clusters C_p. Consequently, the set A_p^s is the complement to the taboo list. For the s-th ant, the probability P_{ip}^s that document d_i is assigned to cluster C_p is denoted by the following formula:

$$P_{ip}^s = \begin{cases} 1 \quad \text{if} \quad \theta < \theta_0 \quad \text{and} \quad i = \arg\max\limits_{u \in A_p^s} \{\alpha\tau_{up} + (1-\alpha)\eta_{up}\}; \\ 0 \quad \text{if} \quad \theta < \theta_0 \quad \text{and} \quad i \neq \arg\max\limits_{u \in A_p^s} \{\alpha\tau_{up} + (1-\alpha)\eta_{up}\}; \\ \dfrac{\alpha\tau_{ip} + (1-\alpha)\eta_{ip}}{\sum\limits_{d_u \in A_p^s} [\alpha\tau_{up} + (1-\alpha)\eta_{up}]} \quad \text{if} \quad \theta \geq \theta_0. \end{cases} \tag{52}$$

In the assignment of the documents the parameter α specifies the importance of the trail τ_{ip} of pheromones relative to the potential utility of the heuristic function η_{ip}, $0 \leq \alpha \leq 1$. The rule (52) defines the probabilities that a particular document is assigned to cluster C_p. It is interpreted thus. A number θ is selected equi-probably from the interval (0, 1). If it is less than the given parameter $0 < \theta_0 < 1$, $\theta < \theta_0$, the cluster C_p

is assigned document d_i such that $\arg\max_{u \in A_p^s}\{\alpha\tau_{up} + (1-\alpha)\eta_{up}\}$. But if $\theta \geq \theta_0$, then with probability $\frac{\alpha\tau_{ip} + (1-\alpha)\eta_{ip}}{\sum_{d_u \in A_p^s}[\alpha\tau_{up} + (1-\alpha)\eta_{up}]}$, cluster C_p is assigned a document selected from the set A_p^s.

Local information η_{ip}, which asserts the utility of the assignment (i, p), is computed thus:

$$\eta_{ip} = \sum_{q=1}^{k}\sum_{j=1}^{i-1}(1 - x_{ip}x_{jq})c_{ij} - (2k-1)\sum_{j=1}^{i-1}(1 - x_{ip}x_{jp})\bar{c}_{ij}. \tag{53}$$

In assigning document d_i to cluster C_p the ants vary the amount of pheromone in the pair (i, p). Such a change constitutes a local updating of the pheromone. The amount of pheromone per pair (i, p) on iteration ν is denoted $\tau_{ip}(\nu)$. In the course of searching for a solution per pair (i, p) updates the amount of pheromone, moreover, reporting its choice to the other ants. It is known that the use of only positive feedback may lead to stagnation, i.e., premature convergence. Therefore, evaporation of pheromone must be assured in order to investigate the entire solution space. By $\rho \in (0, 1)$ we denote the evaporation coefficient. The pheromone is updated on each iteration. In addition, in assignments of "good" solutions, the pheromone level grows, while in other solutions it decreases. At the start of each iteration the amount of pheromone per pair (i, p) is set equal to some positive number τ_0. Then following ν iterations, the amount of pheromone per pair (i, p) will be updated according to the rule

$$\tau_{ip}(\nu + 1) = (1 - \rho)\tau_{ip}(\nu) + \rho\tau_0, \tag{54}$$

Following each iteration only the ant which has found the best solution will update the pheromone in the pair (i, p), using the following global rule:

$$\tau_{ip}(\nu + 1) = (1 - \rho)\tau_{ip}(\nu) + \rho\Delta\tau_{ip}(\nu). \tag{55}$$

In formula (55) the quantity $\Delta\tau_{ip}$ is given as follows:

$$\Delta\tau_{ip} = \begin{cases} f_{\text{best}} & \text{if } (i, p) \text{ belongs to the current best solution;} \\ 0 & \text{in otherwise.} \end{cases} \tag{56}$$

where f_{best} is the value of the best current solution determined by formula (4).

In the course of optimization the number of ants remains constant. Numerous colonies lead to premature convergence of the suboptimal solution. But when there are few ants, the danger arises of a loss of cooperativeness due to limited interaction and rapid evaporation of pheromone. In our case, the number of ants is set equal to the quantity $\left[\frac{n}{k}\right]$, where $[a]$ denotes the integral part of the number a.

6. CONCLUSION

Web users are constantly forced to deal with the problem of finding needed information. The dynamic and rapid development of the World Wide Web has entailed a several-fold increase in the quantity of information available to users. As a consequence, the need for the development of different types of search engines that, in response to a request, would be able to find among a vast quantity of documents those which correspond to the needs of users has become acute.

The problem of increasing the efficiency of searches on the Web has been investigated by many specialists and designers of information retrieval systems. The principal focus has been on the creation of user profiles and the determination of the context of a request. Search efficiency may be significantly increased through the use of the request context and information about the users's sphere of interests, i.e., by knowing the user's intentions, it becomes possible to obtain relevant documents. Basically, the profile is created by the users themselves. It is known that most users report different types of additional information about themselves to the computer out of habit.

Consequently, it is necessary to create a system that will automatically track the user's sphere of interests in order to establish the context of a request. The log is the basic source of information about personal interests. It provides a record of Web sites which the user has visited and a record of the programs which the user has recently launched. By recalling which documents the user has opened, which documents he has scanned

and which he has printed out, a search engine may analyze the user's activity and use the results obtained by conducting a search in a particular direction.

In the present study it has been shown how a user's sphere of interests can be determined automatically A user's sphere of interests is determined through clustering of those documents which he has previously scanned or edited. The process of clustering of documents is reduced to problems of discrete optimization with binary variables, where quadratic- and linear-type models are proposed. A linear model is obtained from a quadratic model by the introduction of new variables and functionally redundant constraints. Since a problem of discrete optimization of high dimension requires considerable computational resources, in the present study a meta-heuristic algorithm for the solution of the problem was proposed.

The clustering models and ant algorithm that have been developed here may be useful in different areas of research, in particular, in image and signal recognition, in business studies, in medicine, and elsewhere.

REFERENCES

1. Jin, X., Zhou, Y., and Mobasher, B., Web Usage Mining Based on Probabilistic Latent Semantic Analysis, in *Proc. 10th ACM SIGKDD Inter. Conf. Knowledge Discovery and Data Mining (KDD'04)*, August 22–24, 2004, Seattle, pp. 197–205.
2. Sugiyama, K., Hatano, K., and Yoshikawa, M., Adaptive Web Search Based on User Profile Constructed without any Effort from Users, in *Proc. 13th Intern. Conf. World Wide Web (WWW13)*, New York, May 17–22, 2004, pp. 675–684.
3. Glover, E.J., Lawrence, S., Gordon, M.D., Birmingham, W.P., and Giles, C.L., Web Search–Your Way, *Comm. ACM*, 2001, vol. 44, no. 12, pp. 97–102.
4. Budzik, J. and Hammond, K.J., User Interactions with Everyday Applications as Context for Just-in-Time Information Access, *Proc. 5h Intern. Conf. Intelligent User Interfaces*, New Orleans, January 9–12, 2000, pp. 44–51.
5. Rogushina, Yu.V., Use of the Context of a Request for Increasing the Relevance of an Information Search on the Internet, *Upravlyayushch. Sistemy Mashiny*, 2004, no. 2, pp. 74–84.
6. Liu, H. and Keselj, V., Combined Mining of Web Server Logs and Web Contents for Classifying User Navigation Patterns and Predicting Users' Future Requests, *Data and Knowledge Engineering*, 2007, vol. 61, no. 2, pp. 304–330.
7. Pitkow, J., Schultze, H., Cass, T., Cooley, R., Turnbull, D., Edmonds, A., Adar, E., and Breuel, T., Personalized Search, *Comm. ACM*, 2002, vol. 45, no. 9, pp. 50–55.
8. Sun, J.-T., Zeng, H.-J., and Liu, H., CubeSVD: A Novel Approach to Personalized Web Search, in *Proc. 14th Intern. Conf. World Wide Web (WWW14)*, Chiba, May 10–14, 2005, pp. 382–390.
9. Scheme, K.-D. and Thalheim, B., Personalization of Web Information Systems – A Term Rewriting Approach, *Data and Knowledge Engineering*, 2007, vol. 62, no. 1, pp. 101–117.
10. Qiu, F. and Cho, J., Automatic Identification of User Interest for Personalized Search, in Proc. 15th Intern. Conf. World Wide Web (WWW15), Edinburgh, Scotland, May 23–26, 2006, pp. 727–736.
11. Liu, F., Yu. C., and Meng, W., Personalized Web Search for Improving Retrieval Effectiveness, *IEEE Trans. Knowledge and Data Engineering*, 2004, vol. 16, no. 1, pp. 28–40.
12. Haveliwala, T.H., Topic-sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search, *IEEE Trans. Knowledge and Data Engineering*, 2003, vol. 15, no. 4, pp. 784–796.
13. Erinaki, M. and Vazirgiannis, M., Usage-based PageRank for Web Personalization, Proc. 5th IEEE Intern. Conf. Data Mining (ICDM-05), Louisiana, USA, November 27–30, 2005, pp. 130–137.
14. Kim, K.-J. and Cho, S.-B., Personalized Mining of Web Documents Using Link Structures and Fuzzy Concept Networks, *Applied Soft Computing*, 2007, vol. 7, no. 1, pp. 398–410.
15. Chen, H. and Chau,M.,Web Mining: Machine Learning for Web Applications, *Ann. Rev. Inf. Sci. Technol.*, 2004, vol. 38, pp. 289–329.
16. Wang, X., Abraham, A., and Smith, K.A., Intelligent Web Traffic Mining and Analysis, *J. Network Comput. Appl.*, 2005, vol. 28, no. 2, pp. 147–165.
17. Stetsyuk, P.I., New Quadratic-Type Models for the Problem of the Maximal Weighted Section of a Graph, *Kibernetika Sistem. Analiz*, 2006, no. 1, pp. 63–75.
18. Tolcheev, V.O., Methods of Identifying Information-Bearing Attributes in the Problem of Classification of Text Documents, *Informats. Tekhnol,* 2005, no. 8, pp. 14-21.
19. Gehry, M and Johnson, D., Computational Machines and Hard-to-Solve Problems
20. Dorigo, M., Maniezzo, V., and Colorni, A., The Ant System: Optimization by a Colony of Cooperating Agents, *IEEE Trans. Systems, Manufactures and Cybernetics. Part B*, 1996, vol. 26, no. 1, pp. 29–41.
21. Dorigo, M., Di Caro, G., and Gambardella, L.M., Ant Algorithms for Discrete Optimization, *Artificial Life* (MIT Press), 1999, vol. 5, no. 2, pp. 137–172.

 Spell: OK