Social Network Problem in Enron Corpus

Jiří Dvorský, Petr Gajdoš, Eliška Ochodková, Jan Martinovič, and Václav Snášel

Department of Computer Science,
VŠB - Technical University of Ostrava,
tř. 17. listopadu 15, 708 33 Ostrava-Poruba
Czech Republic
Jiri.Dvorsky@vsb.cz, Eliska.Ochodkova@vsb.cz

Abstract. Traditional communication barriers are disappearing due to expansion of electronic communication devices. Forming of communities does not depend only on handshaking or sending a letter to other person. Modern communication devices give rise to originating of new types of communities without necessity of their geographical proximity. Fast communication brings disadvantages connected with determining of communities. The question is, if there are any methods, how to identify particular communities or how to identify topics of their communication. Members of community can be represented by vertices and communication channels by edges. The whole problem can be solved using graph theory and information retrieval methods. In our paper we describe method, how to identify these communities, based on searching of 2-connected components in social nets. Communication topics can be specified using clustering methods. To demonstrate our approach we use the Enron corpus.

Keywords: Social networks, clustering, graph theory, Enron corpus.

1 Introduction

Networks arising from real life are concerned with relations between real objects and are important part of modern life. Important examples include links between web pages, citations of references in scientific papers, social networks of acquaintance or other connections between individuals, electric power grids, etc. Word "network" is usually used for what mathematicians and a few computer scientists call graphs [11]. A network (graph) is a set of items called vertices with connections between them, called edges. The study of a graph theory, is one of the fundamental pillars of discrete mathematics.

A social network is a set of people or groups of people with some pattern of contacts or interactions between them. Social nets have been studied extensively since the beginning of 20th century, when sociologists realized the importance of the understanding how the human society is functioned. The traditional way to analyze a graph is to look at its picture, but this is unusable for large networks.

New approach to examine properties of graphs has been driven largely by the availability of computers and communication networks, that allow us to analyze data on a scale far larger than before now, see [12, 8].

Interesting source of reliable data about personal connections between people are communication records of certain kinds. For example, one could construct a network in which each the vertices represent email addresses and directed edges represent a message passing from one address to another.

Complex networks such as the World Wide Web or social networks often do not have an engineered architecture but instead are self-organized by the actions of a large number of individuals. From these local interactions nontrivial global phenomena can emerge as, for example, small-world properties or a scale-free distribution of the degree, see [11]. In small-world networks short paths between almost any two sites exist even though nodes are highly clustered. Scale-free networks are characterized by a power-law distribution of a nodes degree, defined as the number of its next neighbors, meaning that structure and dynamics of the network are strongly affected by nodes with a great number of connections. There is reported in [6] that networks composed of persons connected by exchanged emails show both the characteristics of small-world networks and scale-free networks.

Our approach combines graph theory with cluster analysis and techniques known in information retrieval community. This synthetic method can be understood as data mining method despite the fact that there is no database. Mining procedure use topological information about graphs and combines it with contens of emails' textual bodies which are evaluated using cosine similarity measurement coming from vector model.

The paper is organized as follows: Section 2 contains necessary definitions, description of clustering method etc. Section 3 introduces our approach to social network problem. Our experimental results are described in section 4. Conclusion and future works are given in section 5.

2 Background

2.1 Connected components

This paragraph introduces basic definitions from graph theory, for more details see [21].

Definition 1. *A vertex-cut in a graph G is a set U of vertices of G such that $G \setminus U$ is not connected.*

Definition 2. *The vertex-connectivity or simply connectivity $\kappa(G)$ of a graph G is the minimum cardinality of a vertex-cut of G (Hence $\kappa(G)$ is the minimum number of vertices whose removal results in a disconnected or trivial graph).*

Definition 3. *A graph G is said to be k-connected, $k \geq 1$ if $\kappa(G) \geq k$.*

Definition 4. (2-connected component) *Let G be a graph. A k-component of G is a maximal k-connected subgraph H of G.*

The higher the degrees of the vertices of a graph, the larger connectivity of graphs. A k-component of the graph is a maximal graph with a cut-set of the size k and with k or more independent paths between any pair of its nodes.

Two paths from i to j in the graph G are node independent if they have only nodes i and j in common. If there is at least one path linking every pair of actors in the graph then it is connected. If there are k node-independent paths connecting every pair, the graph is k-connected and called a k-component.

2.2 Cluster analysis

Finding of groups of objects with the same or similar features within given set of objects is the goal of cluster analysis [7, 23]. These groups are called *clusters*. In other words, the groups of similar objects forms the cluster.

Hierarchical clustering methods are important tool of cluster analysis. Hierarchical methods create hierarchy of clusters, which are grouped in levels. Agglomerative clustering method is used in this paper. Basic principle of the clustering method can be seen in Figure 1 (see [20]).

1. **Create matrix of objects' distances**
   Matrix of objects' distances will be equal to document-term weight matrix D.
2. **Define each object as cluster**
   At the beginning each object is considered as one cluster i.e. there are as many clusters as objects. Sequentially, clusters are joined together and number of clusters drops down, when finally there is one cluster.
3. The most similar clusters p and q are found in similarity matrix and their similarity is determined as $s_{p,q}$.
4. Clusters p and q are joined together and the number of cluster is reduced. New formed cluster is determined as t. The cluster t replaces cluster q in similarity matrix. Distances of new cluster t to other clusters r are recomputed according to following formula [22]:
   $$s_{tr} = \frac{N_p s_{pr} + N_q s_{qr}}{N_p + N_q}$$
   Now cluster p is removed from similarity matrix i.e. size of matrix is reduced.
5. **if there are more than one cluster, go to step 3**

Fig. 1. Agglomerative clustering algorithm

Matrix of objects' distances will be equal to document(email)-term weight matrix D. The matrix is computed using a method taken from *vector model.* In the vector model there are documents (and potential users queries) represented by vectors.

We have used m different terms $t_1, \ldots, t_m$ for indexing n documents. Then each document d_k is represented by a vector:

$$d_k = (w_{k1}, w_{k2}, \ldots, w_{km})$$

where w_{kl} is the weight of the term t_l in the document d_k. Greater weight corresponds to more important term in the document.
An index file of the vector model is represented by following matrix:

$$D = \begin{pmatrix} w_{11} & w_{12} & \ldots & w_{1m} \\ w_{21} & w_{22} & \ldots & w_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \ldots & w_{nm} \end{pmatrix}$$

where k-th row matches k-th document, and l-th column matches l-th term.

The similarity of two documents is given by following formula (*cosine measure* [2]):

$$sim(d_k, d_l) = \frac{\sum_{i=1}^{m} w_{ki} w_{li}}{\sqrt{\sum_{i=1}^{m} w_{li}^2} \sqrt{\sum_{i=1}^{m} w_{ki}^2}}$$

3 Our approach

3.1 Triangle-based approach

In many networks it is found that if the vertex A is connected to the vertex B and the vertex B to the vertex C, then there is a heightened probability that vertex A will also be connected to the vertex C. In the language of social networks, the friend of your friend is likely also to be your friend. In terms of network topology, transitivity means the presence of a heightened number of triangles in the networksets. It can be quantified by defining a clustering coefficient C thus:

$$C = \frac{3 * \; number\ of\ triangles\ in\ the\ network}{number\ of\ connected\ triples\ of\ vertices} \tag{1}$$

The clustering coefficient measures the density of triangles in a network, see [11].

3.2 Description of our approach

Widely used approach to analyze communities structures is based on searching for triangles in social networks, [15]. According to the graph theory, we can say that each triangle is a smallest 2-connected component because it satisfies the conditions of the definition 4. We have used this knowledge in our approach. Also all 2-connected components in the graph of the Enron Corpus have been computed. That is the way how to obtain all possible social groups and analyze the channels between all members of each group. Whole process is described in the following paragraph.

Firstly, an undirected graph based on the Enron Corpus dataset was created. A vertex represents one subject of communication and an edge represents the fact, that there was a communication between two subjects (vertices). It is clear there are usually many data flows represented by emails between two subjects. These data flows give rise to many edges in our graph. Therefore the number of edges was reduced by removing multiple channels. We do not need them in the computation of 2-connected components, but we use them in future. After that, we have computed all 2-connected components by our improved algorithm described in the following chapter. Then we used a hierarchical clustering to analyze all emails which have been sent between the members of each group. All available emails make an input of this step.

3.3 2-connected components.

Algorithm [16] for finding the 2-connected components of a graph breaks a graph into its 2-connected components by performing a depth-first search along the edges of the graph. Each new point reached is placed on a stack, and for each point a record is kept of the lowest point on the stack to which it is connected by a path of unstacked points. When a new point cannot be reached from the top of the stack, the top point is deleted, and the searching is continued from the next point on the stack. If the top point does not connect to a point lower than the second point on the stack, then this second point is an articulation point of the graph. All edges examined during the search are placed on another stack, so that when an articulation point is found the edges of the corresponding 2-connected component may be retrieved and placed in an output array. When the stack is exhausted, a complete search of a connected component has been performed. If the graph is connected, the process is complete. Otherwise, an unreached vertex is selected as a new starting point and the process repeated until all of the graph has been exhausted. Isolated points are not listed as 2-connected components, since they have no adjacent edges. They are merely skipped.

4 Experimental results

A large set of email messages, the *Enron corpus*[1], was made public during the legal investigation concerning the Enron corporation. The raw Enron corpus contains 619 446 messages belonging to 158 users (see [1]). Each message present in the folders contains the senders and the receiver email address, date and time, subject, body, text and some other email specific technical details.

Our method consists of the following three steps:

1. The dataset has a lot of integrity issues. It has many duplicate and corrupt messages. We cleaned the corpus before this analysis by removing certain folders from each user, such as "discussion_threads", "all_documents",

[1] Corpus is available at `http://www-2.cs.cmu.edu/~enron/`

"_sent_email". These folders were presented for most users, and did not appear to be used directly by the users, but rather were computer generated. Our goal in this paper is to analyze graph properties of Enron corpus and perform cluster analysis on selected emails, so these folders would have likely been misleading. In our cleaned Enron corpus, there are a total of 300 458 messages belonging to 150 users. The graph G has 77784 vertices and 332777 edges.

2. All 2-connected components were found in the second step. 44685 components were found in graph G. Maximal components has 31420 vertices and contains most of the graph G, 44685 components have only two vertices. These components are unusable for our purposes, because they contains all or nothing from social networks point of view. Detailed number of 2-connected components is given in table 1. Finding 2-connected component was implemented in C# language and takes only about one minute on Pentium 4 2.4 GHz.

Table 1. List of components in graph G

# of vertices	# of component
2	44619
3	32
4	17
5	3
6	3
7	1
9	1
10	1
11	2
17	1
20	1
36	1
61	1
343	1
31420	1

3. The third step in our approach was cluster analysis. Hierarchical agglomerative clustering method, with average selection strategy, was used. Clustering was performed on all components with at least 3 vertices and at most 343 vertices.

4.1 Results

Computed clusters were depicted as *dendograms*. The dendogram shows on axis x maximal level of similarity, where joining of two clusters was performed. Joining is done on interval $\langle 0; 1\rangle$. The interval corresponds to cosine measure of similarity of emails in document-document similarity matrix. Zero value means that two clusters are totally different and value 1 means that clusters are the same.

Several typical components and corresponding cluster structures were identified among many particular components (basic properties of selected components are given in table 2). Identification is based on comparison of emails' textual bodies and corresponding cluster structure:

distribution lists and their members – these components have relatively a lot of vertices i.e. users with email addresses (see figure 2). There are also two vertices with high degree - distribution lists. Cluster analysis also shows presence of several groups of similar emails. Emails in particular group are very close to each other (clustering level is close to 1), and at the same time each groups is dissimilar to other group of emails (clustering level from 0 to 0.3).

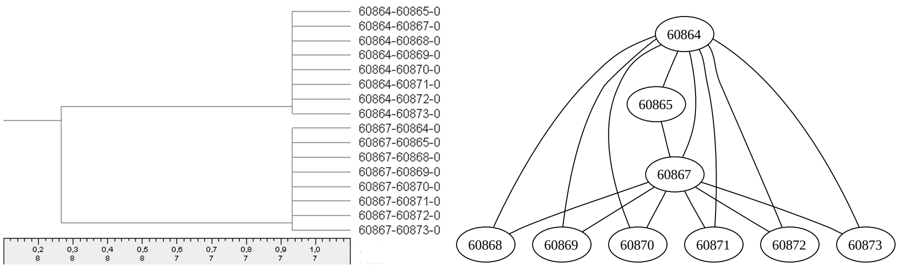
distribution lists and answers – these components also represent distribution lists. Cluster analysis shows several clusters of similar emails, represented as cluster at the level approximately 1. But there are different emails with lower similarity coefficient – answers to emails from the distribution lists. The answer can be detected as a small cluster connected to cluster formed at similarity level close to 1 (see figure 3).

users' communication – the third type of components contains a small number of group emails. But there are a lot of individual emails between two users. There are several clusters at level 1 and many small clusters at higher levels. Emails in these clusters represents communication between users and topic of emails are under development in time. Emails typically contains original email, answer and/or question to the sender of original emails etc. In this way emails come different, and form clusters at higher level (see figure 4).

collaborative groups – these components represent small group of users with communication within group, but these groups are not separated. Some users belong to more than one group. Example of component can be seen in figure 5.

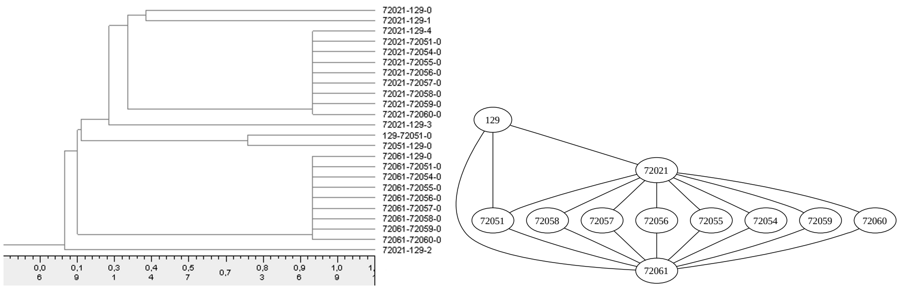
Table 2. Properties of components

	Component Id number			
	44362	40106	38413	11480
Number of vertices	9	11	17	20
Number of edges	15	19	30	72
Number of emails	16	24	70	91



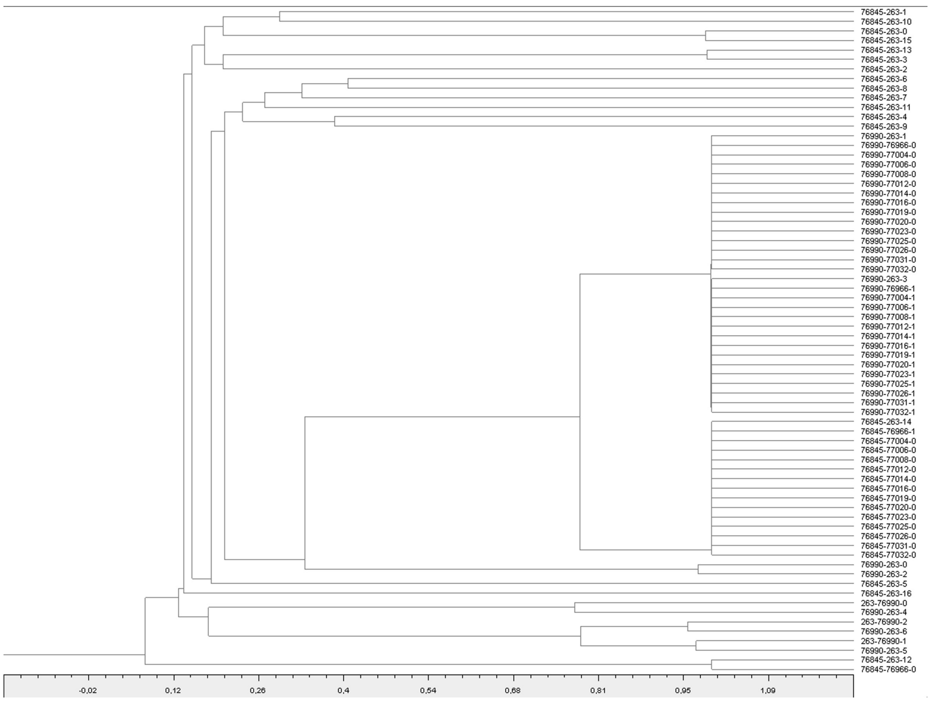
(a) Cluster of component 44362 (b) Graph of component 44362

Fig. 2. Distribution lists and their members

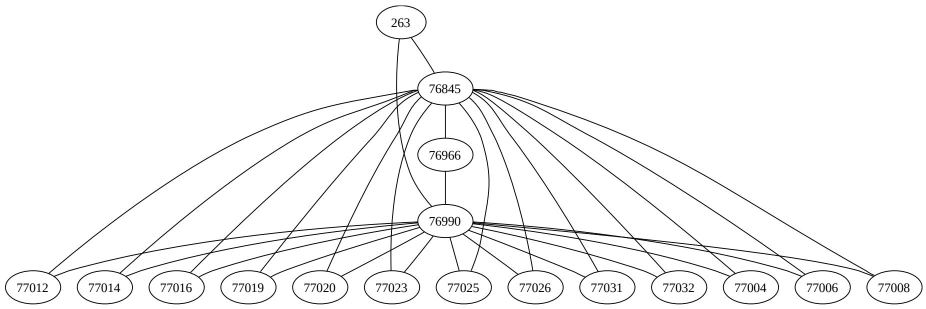


(a) Cluster of component 40106 (b) Graph of component 40106

Fig. 3. Distribution lists and answers

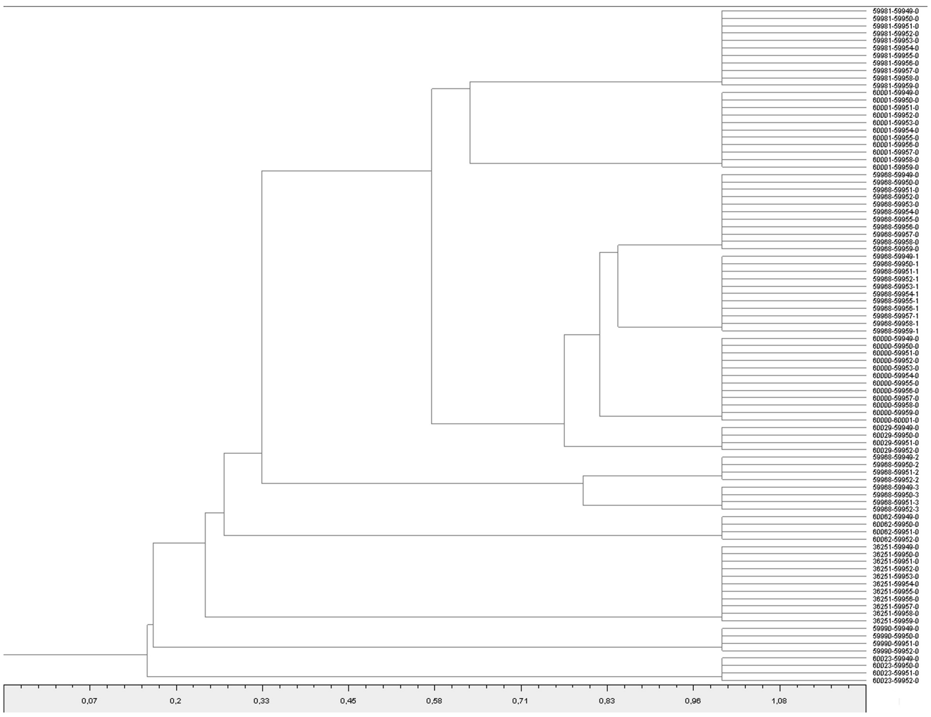


(a) Cluster of component 38413

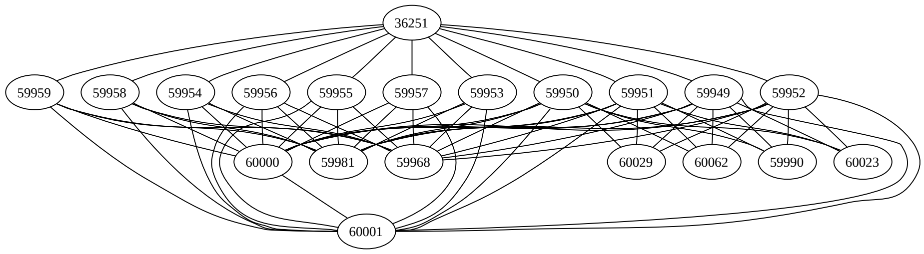


(b) Graph of component 38413

Fig. 4. Users' communication



(a) Cluster of component 11480



(b) Graph of component 11480

Fig. 5. Collaborative groups

5 Conclusion and future works

In this article we have reviewed some recent work on the structure and function of networked systems. Work in this area has been motivated to a high degree by empirical studies of real-world networks such as the Internet, the World Wide Web, emails corpora, social networks, collaboration networks, citation networks, and a variety of biological networks.

Our work is inspired by many researchers [11, 12, 14], that have proposed models of networks that typically seek to explain either how networks come to have the observed structure, or what the expected effects of that structure will be.

Our approach combines cluster analysis and algorithms taken from graph theory, especially 2-connected components. 2-connected components allow us to find structure of social communities. Topics discussed in particular communities have been found by cluster analysis. The intersection of social communities and 2-connected components shows behavior of users in network. For example in figure 2 we can see behavior of user 60865 - the topmost user sent email to all member in the group. The user 60865 resent again the same email to all member, as can be seen in corresponding cluster.

Outgoing edges from one 2-connected component to another will be direction of out future works. These edges represents communication between particular communities and may signal leakage of information.

References

1. Klimt B., Yang Y.: Introducing the Enron Corpus, Proceedings of First Conference on Email and Anti-Spam (CEAS) (2004)
2. Baeza-Yates R., Ribeiro-Neto B.: Modern Information Retrieval. Addison Wesley, New York (1999)
3. Berry, M. W (Ed.): Survey of Text Mining: Clustering Classification, and Retrieval. Springer Verlag (2003)
4. Chatrand G., Lesniak L.: Graphs & Digraphs. Chapmann & Hall/CRC (1996)
5. Dvorský J., Martinovič J., Snášel V.: Query Expansion and Evolution of Topic in Information Retrieval Systems, DATESO (2004)
6. Ebel H., Mielsch L.I., Bornholdt S.: Scale-free topology of e-mail networks, Phys. Rev. E, 66 (2002)
7. Faloutsos Ch., Oard D.: A Survey of Information Retrieval and Filtering Methods, Univ. of Maryland Institute for Advanced Computer Studies Report, College Park (1995)
8. Guimerà R., Danon L., Díaz-Guilera A., F. Giralt F., Arenas A.: Self-similar community structure in a network of human interactions. Physical Review, vol. 68 (2003)
9. Hung-Yu K., Shian-Hua L., Jan-Ming H., Ming-Syan Ch.: Mining Web Informative Structures and Contents Based on Entropy Analysis. IEEE Transactions on Knowledge and Data Engineering vol. 16, No. 1. (2004)
10. Kummamuru K., Lotlikar R., Roy S., Singal K., Krishnapuram R.: A Hierarchical Monothetic Document Clustering Algorithm for Summarization and Browsing Search Results, WWW2004, New York, USA.

11. Newman M. E. J.: The Structure and Function of Complex Networks. SIAM Review, vol. 45 (2003) 167–256
12. Newman M. E. J., Balthrop J., Forrest S., Williamson M. M.: Technological networks and the spread of computer viruses. Science, vol. 304 (2004) 527–529
13. Pool I., Kochen M.: Contacts and influence. Social Networks, 1 (1978) 148
14. Newman M. E. J.: Fast algorithm for detecting community structure in networks. Phys. Rev. E 69, 066133 (2004)
15. Ravasz E., Barabási A.-L.: Hierarchical organization in complex networks. Phys. Rev. E, 67 (2003)
16. Tarjan R.: Testing Graph Connectivity. Sixth Annual ACM Symposium on Theory of Computing, Assoc. Comput. Mach. New York (1974) 185–193
17. Rijsbergen C. J. van: Information Retrieval (second ed.). London, Butterworths (1979)
18. Rijsbergen Keith Van: The Geometry of Information Retrieval, Cambridge University Press (2004)
19. Vempala S.: The Random Projection Method, Dimacs Series in Discrete Mathematics and Theoretical Computer Science (2004)
20. Dvorský J., Martinovič J., Pokorný J., Snášel V.: A Search topics in Collection of Documents.(in Czech), Znalosti (2004)
21. Tutte W. T.: Graph Theory, Encyclopedia of mathematics and its applications, volume 21, Addison Wesley (1984)
22. Martinovič J., Snášel V.: Topic Evolution Usage in Vector Query Improvement (in Czech), ZNALOSTI (2005)
23. Jain A. K., Dubes R. C.: Algorithms for Clustering Data, Prentice Hall Advanced Reference Series: Computer Science, Prentice Hall (1988)