2005

Document Clustering in Antimicrobial Peptides Research

Yi Peng
University of Nebraska at Omaha, ypeng@mail.unomaha.edu

Nian Yan
University of Nebraska at Omaha, nyan@mail.unomaha.edu

Gang Kou
University of Nebraska at Omaha, gkou@mail.unomaha.edu

Zhengxin Chen
University of Nebraska at Omaha, zchen@mail.unomaha.edu

Yong Shi
University of Nebraska at Omaha, yshi@mail.unomaha.edu

Document Clustering in Antimicrobial Peptides Research

Yi Peng
ypeng@mail.unomaha.edu

Nian Yan
nyan@mail.unomaha.edu

Gang Kou*
gkou@mail.unomaha.edu

Zhengxin Chen
zchen@mail.unomaha.edu

Peter Kiewit Institute of Information Science, Technology & Engineering, University of Nebraska, Omaha, NE 68182, , Phone number: ++1(402)5543429 or ++1(402)5543625

Yong Shi
yshi@mail.unomaha.edu
Peter Kiewit Institute of Information Science, Technology & Engineering, University of Nebraska, Omaha, NE 68182, Phone number: ++1(402)5543652

yshi@gscas.ac.cn
Chinese Academy of Sciences Research Center on Data Technology and Knowledge Economy, Beijing 100039, China, Phone number: ++8613651346898

ABSTRACT

Antimicrobial peptides are small peptides encoded by genes. The research area of antimicrobial peptides has attracted intense attention in recent years because "their potential use in the cure of infectious diseases caused by pathogens that have become counteractive to traditional antibiotics" (Boman 1994). There exist huge amount of antimicrobial peptides research articles and this number is continuously increasing. Although some biomedical databases, such as PubMed, have been well established, they provide only query-based information retrieval and end-users need to manually find out relevant information from thousands of retrieved articles.

The objective of this paper is to apply one of the text mining techniques, *document clustering*, which groups similar documents into clusters, to text documents collected from PubMed using keyword "antimicrobial peptides". The results of our work can help researchers to discover meaningful groups of antimicrobial peptides articles in an efficient manner.

Keywords: text mining, antimicrobial peptides, document clustering, PubMed

INTRODUCTION

The field of data mining has achieved a significant growth and attracted strong interest from researchers and practitioners during the last decade. Although various data mining algorithms and tools have been developed for structured databases, few have been investigated for text documents. Plenty of useful information is stored in text formats, such as call data, mail-order addresses, sales histories, web transactions, free-form text notes (SAS 2005), library databases, and research articles. Contrasted with structured databases, text documents are stored in unstructured or semi-structured formats that are difficult to decipher. However, the basic idea of mining text documents is akin to data mining for structured data. A general definition of text mining: "text mining is about looking for regularities, patterns or trends in natural language text, and usually is about analyzing text for particular purposes (Nahm 2001)" reveals this similarity.

Similar to data mining, text mining is a cross-disciplinary field including, but not limited to: information retrieval, information extraction, natural language processing, computational linguistics, machine learning, data mining, and

information visualization (Nahm 2001). Text mining has various applications. According to Porter (2002), text mining applications can be classified into five categories: retrieving document, identify infrastructure, identify technical themes/relationships, discovery from literature, and technology forecasting. These applications require a spectrum of text mining techniques. For document retrieving, information retrieval techniques are enough. For discovery from literature and technology forecasting, more advanced text mining algorithms and techniques are needed. Among these text mining techniques, clustering techniques have been used in many areas to discover groups of documents and to identify potential abstract structures, especially when the features of databases are unknown (Kogan, Nicholas, and Teboulle, 2003).

In life sciences, clustering has been used in the analysis of clinical information, phylogeny, genomics, and proteomics, to name just a few (Zhao and Karypis, 2002). This paper is trying to apply document clustering to antimicrobial peptides-related research articles. Antimicrobial peptides are small peptides encoded by genes (Boman 1994). The research area of antimicrobial peptides has attracted intense attention in recent years because "their potential use in the cure of infectious diseases caused by pathogens that have become counteractive to traditional antibiotics" (Boman 1994). There exist huge amount of antimicrobial peptides research articles and this number is continuously increasing. Although some biomedical databases, such as PubMed, have been well established, they provide only query-based information retrieval and end-users need to manually find out relevant information from hundreds or even thousands of retrieved articles. This task is time-consuming and inefficient. To the best of our knowledge, no clustering technique has been used to group antimicrobial peptides articles. The primary goal of this paper is to help researchers understand current antimicrobial peptides development status efficiently and identify potential research directions of antimicrobial peptides using partitioning clustering approach.

The paper is organized in three parts. The first part provides information about PubMed, a national digital archive of biomedical and life sciences journal literature, from where antimicrobial peptides research articles was collected. The second part describes document clustering concepts and techniques used in this research. The third part represents the clustering procedure and results. The article concludes with future research directions.

PUBMED

PubMed is the U.S. National Library of Medicine's premiere search system for health information and biomedical related research publications. It is free at http://pubmed.gov. PubMed includes over 15 million citations for biomedical articles and provides access to MEDLINE, OLDMEDLINE, in-process citations, and publisher-supplied citations (PubMed 2005). As a specialized system, PubMed has the search capabilities for biomedical terminologies and has been widely used in biomedical field.

In this article, PubMed was used to collect up to date antimicrobial peptides related publications. The "Clustering Procedure and Results" section will describe the details about how those articles are retrieved and processed and this section will introduce PubMed search options, features bars, and results display options.

PubMed search is similar to Google or Yahoo search engines. There is a query box at the top of PubMed website. Users can type a word or phrase into the query box, and then click on the Go button to start the search. They can also combine search words with connector words: "AND", "OR" or "NOT" (PubMed Basics 2004).

There are five features bars below the query box: Limits, Preview/Index, History, Clipboard, and Details. Limits provides pull-down menus for Volume, Publication Type, Languages, Subsets (special areas of interest), Ages, Human or Animal, Gender, Entrez Date (the date when the item was added to the database), and Publication Date. Each selection has multiple choices available. Preview allows users to add terms to the query box and see the number of search results and index allows users to view terms within a field. A field can be specified as language, issue, journal, author, and so on. History displays most recent queries and ranks queries by search numbers. The Clipboard allows collecting of selected citations from one or more searches for printing, saving, e-mailing or ordering. The Details feature describes query information such as query translation (the corresponding MeSH headings for the search term), database, and number of results (PubMed Basics 2004).

The search results of PubMed will be displayed in Summary format by default. Users can select from 35 formats options to display the query result, such as Abstract, Citation, MEDLINE, or XML. In this article, only titles and abstracts of antimicrobial peptides related papers were used in clustering. Hence, Abstract format was selected to display and collected the search results.

DOCUMENT CLUSTERING: CONCEPTS AND TECHNIQUES

Clustering in text mining involves various concepts and techniques. This section will only focus on concepts and techniques that were used in this research. In particular, selected concepts and techniques of text preprocessing, retrieval model, and clustering that were applied in this paper are discussed in sequence.

Text Preprocessing: Tokenization, Stop-words, and Stemming

The goals of text preprocessing are: to represent full-text documents in a suitable format and to optimize the performance of text mining algorithms by discarding irrelevant data (Mathiak and Eckstein, 2004). Text documents are first divided into a set of *index terms* or *keywords*. This division process is called *tokenization*. These index terms are then used to represent full-text documents, which refer to the titles and abstracts in our case. Different index terms have varying importance and this difference is expressed using *weight*s. Each keyword in a document is associated with a weight. In this paper, a self-developed C++ program was used to tokenize antimicrobial peptides-related article titles and abstracts into keywords.

Normally, tokenization will result in thousands or even tens of thousands of keywords. Dealing with such a high dimensionality is a formidable task for present data analysis tools. Stop-words and stemming are two prevalent keywords reduction methods. Regardless of topics and research areas, there are always common words that occur frequently in all documents, such as articles, prepositions, and conjunctions. This type of words is so called *Stop-words* and are irrelevant for the purpose of retrieval. There are free stop-words lists on the Internet. A *stem* is the portion of a word which is left after the removal of its affixes, i.e., prefixes and suffixes. Porter's stemming algorithm (Porter 1980) has been widely used and was selected by this project to further reduce the dimensionality.

Retrieval Model: Vector Space Model

After converting full-text documents into a set of keywords and reducing indexing structure, retrieval models can be set up. In information retrieval, retrieval models can be classified into three categories: *Boolean model*, *vector space model*, and *probabilistic model*. *Boolean model* considers index terms are either present or absent in a document, and hence can not recognize partial matches. *Probabilistic model* is based on the probabilistic principle (Baeza-Yates and Ribeiro-Neto, 1999). For different experiments, probabilistic model and vector model may have different performance. Nevertheless, it has been shown that the vector model is expected to outperform the probabilistic model with many cases. Due to the reasons above, vector space model was chosen in this paper.

According to *Vector space model*, a document can be represented as a vector:

$$<(d_{r1}, w_1), (d_{r2}, w_2), (d_{r3}, w_3),\ldots(d_m, w_n)>,$$

where d_{ri} denotes a keyword i used to describe the document r, and w_i denotes the weight of the keyword i, which can be determined by frequency of use. A collection of n documents can be represented by a *term-document matrix*. An entry in the matrix corresponds to the weight of a term in that document; zero means the term doesn't exist in the document or has no significance in the document (Baeza-Yates and Ribeiro-Neto, 1999).

Researchers have developed various weighting schemes to calculate weights of terms. A typical term weight is *tf-idf* weighting: $w_{ij}=tf_{ij} \times idf_i$, where tf_{ij} is *term frequency* across the entire corpus: $tf_{ij} = f_{ij} / \max\{f_{ij}\}$ and f_{ij} is the frequency of term i in document j; idf_i is the *inverse document frequency* of term i: $idf_i = \log_2(N/df_i)$, N is the total number of documents and df_i is the document frequency of term i, i.e. the number of documents containing term *i*. A term in a document with higher *tf-idf* weight is regarded as more indicative than terms with lower *tf-idf* weights. The reasoning behind *tf-idf* weighting is that a term occurring frequently in a document but rarely in the rest of the collection is considered to be important. Experiments have shown that *tf-idf* weighting works well in many applications (Baeza-Yates and Ribeiro-Neto, 1999). In this paper, the frequencies of each index term within each abstract were counted using SQL and *tf-idf* weights were computed and stored in a term-document matrix.

Clustering Methods

As an unsupervised classification method, clustering has been extensively studied and many algorithms have been developed. Major clustering approaches can be categorized into partitioning, hierarchy, density-based, grid-based, and model-based algorithms (Han and Kamber, 2000).

In life science, clustering methods have been investigated and applied by researchers in recent years. Weiss, White, and Apte (2000) described a lightweight document clustering method that is capable of operating in high dimensions and grouping them into several thousand clusters. The method uses a reduced indexing view of the original documents and has been evaluated on a database of over 50,000 customer service problem reports that were reduced to 3,000 clusters and 5,000 exemplar documents. Iliopoulos, Enright, and Ouzounis (2001) presented an algorithm for large-scale document clustering of biological text. The algorithm is based on statistical treatment of terms, stemming, go-list, unsupervised machine learning and graph layout optimization. Benjamin, Wang, and Ester (2003) proposed to use frequent itemsets, which comes from association rule mining, for document clustering. Each cluster is identified by some frequent itemsets. By focusing on frequent items, the dimensionality of the document set is drastically reduced. Zhao and Karypis (2004) provide an overview of the various issues of clustering in various areas within life-sciences. This overview describes the various types of clustering algorithms, discusses issues related to clustering gene expression datasets, and introduces a clustering toolkit - CLUTO.

In this research, k-means is selected to cluster documents due to its popularity and simplicity. K-means clustering algorithm (McQueen 1967) partitions data objects into k disjoint subsets, where k is predefined. The k-means algorithm consists of four steps. In the first step, k objects are randomly selected to be the centers of the k clusters. The second step computes similarities between each data object and these k centers. Each object is then assigned to the most similar center. Thus the initial k clusters are established. In the third step, the k centers are re-computed using data objects in each cluster. In the last step, each data object is reassigned to the most similar center. This process stops when no more new assignment occurs (Han and Kamber, 2000). Similarities between objects can be measured using different functions and this paper used SPSS (SPSS 12.0) k-means function to do the work.

CLUSTERING PROCEDURE AND RESULTS

Following previous discussion, text documents (Titles and Abstracts) were first gathered using PubMed system on Oct 25, 2004. The titles and abstracts were then preprocessed by removing stop-words and stemming. Because there are still about 26,000 index terms left after stop-words and stemming operations and SPSS k-means function cannot handle such a high dimensionality, a heuristic rule was applied to further reduce the number of distinct index terms. Our assumption is that the most or least frequent words are not important for calculating TF/IDF value. Thus terms with less than 200 or larger than 2000 frequency were considered as less important and dropped from index terms. For each index term remains, *tf*, *idf*, and *tf-idf* weights were calculated. Based on *tf*, *idf*, and *tf-idf* weights, a term-document matrix was created. Finally, the k-means function of SPSS 12.0 for Windows was used to do clustering analysis. Different numbers of clusters were explored and three representative clustering results were reported, i.e. clusters number equals to 10, 20, and 50. The following procedure summarized the whole process:

Clustering Procedure:

Input: The 4608 (as of Oct 25, 2004) articles from PubMed according to the user defined query criteria: "Antimicrobial Peptide" and the Number of clusters for k-means algorithm.

Output: A set of Clusters, each of which contains a group of articles.

Step 1 Extract the distinct words from "abstract" and "title" of all these articles using a self-coded C++ program and store them into a word list table in the database. There are 45,735 distinct words in total.
Step 2 Drop off the numbers, abnormal characters and commonly used 1000 stop-words (Edict virtual language centre 2005) appeared in the word list.
Step 3 Use Porter stemming technique to narrow down the size of the word list. After step 2 and 3, the number of distinct words is reduced to 26,495.
Step 4 For each word, calculate both the frequency of that word appeared in each article and the frequency in all articles. Figure 1 shows the frequency of each word appeared in all the articles.
Step 5 We assume that the most or least frequent words are not important for calculating TF/IDF value. Less important words are defined as having frequency less than 200 or larger than 2000.
Step 6 Calculating the TF/IDF weights and creating the term-document matrix represented the relationships among the documents and words. Table 1 displays a small portion of the term-document matrix, in which each document is represented by a TF/IDF weight vector of terms. Index terms and documents are represented using numerical IDs and an entry in the matrix corresponds to the TF/IDF weight of a term in that document.

Step 7 Use k-means function from SPSS 12.0 for Windows (SPSS 12.0) to do cluster analysis.

END

Table 2, 3, and 4 summarize the number of articles within each cluster when numbers of clusters were defined as 10, 20, and 50. In addition, if users want to know which articles are grouped together, they can click on a specific document and get connected to the full abstract of that article. For 10 clusters case (Table 2), we show three examples of document titles for each cluster. These detailed views allow users to explore their interested subcategory of "Antimicrobial Peptide". In table 3 and 4, due to the large amount of clusters, we only use three example articles for the first two clusters to illustrate the detailed subcategory information. The words frequencies (Figure 1) and clustering results can be used in several ways: through examining the frequent words (e.g., the top 10% most frequently used terms), users may identify antimicrobial peptides research directions or focuses which are unavailable through PubMed search; by investigating clustering results, users may identify research groups in antimicrobial peptides that are related to their interests.

CONCLUSION

This article applied text mining techniques in the area of antimicrobial peptides. Specifically, text documents were gathered using PubMed search engine and preprocessed using tokenization, stop-words, stemming, and a heuristic rule. After that, a term-document matrix was created and SPSS k-means was used to cluster documents. The clustering results of this paper, with 10, 20, and 50 clusters, are not directly useful without summarization of each cluster. Summarizations of each cluster need experts' inspections. For example, experts of antimicrobial peptides can randomly select 3 to 5 documents from each cluster and determine to which subcategory the cluster should belong.

The results of this paper serve as a starting point for more advanced text mining research in antimicrobial peptides. One research direction is to investigate the applicability of novel document clustering algorithms which can deal with large number of index terms. Many existing clustering methods are incapable of handling tens of thousands of index terms. Although preprocess techniques, such as stop-words and stemming can be used to reduce some irrelevant terms, the number of index terms may still be huge. For example, in the clustering of antimicrobial peptides articles, we ended up with more than 26,000 index terms even after removing stop-words and stemming. Without consultation from antimicrobial peptides experts, we have no clue about which terms should be kept and which terms should be eliminated. Furthermore, it is time consuming for experts to go through index terms manually. Thus, it is important to design new or reconstruct existing clustering algorithms which can manage massive number of index terms automatically.

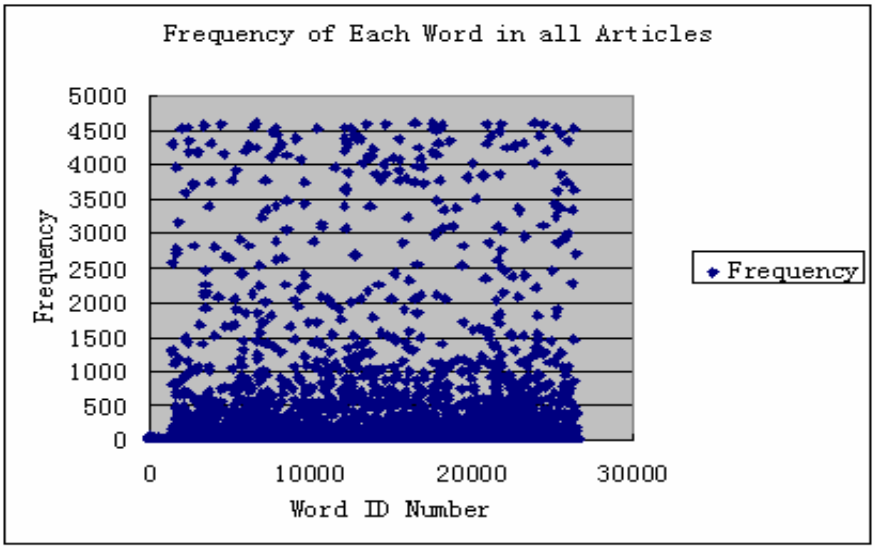

Figure 1. Words Frequencies

Word_id / Doc_id	1	2	3	4	5	6	7	…
…	…	…	…	…	…	…	…	…
11	0	0	0	0.02	0	0	0	…
12	0	0	0	0	0	0	0	…
13	0	0	0	0	0	0.01	0.02	…
14	0	0	0.02	0	0	0	0	…
15	0	0	0	0	0	0	0	…
16	0	0	0	0	0	0	0	…
17	0.02	0.04	0	0	0.02	0.01	0	…
18	0	0	0	0	0	0	0	…
19	0	0	0	0	0	0	0	…
20	0	0	0	0	0	0	0	…
…	…	…	…	…	…	…	…	…

Table 1. The Term-Document Matrix

cluster	No. of documents	Three Examples of Document Titles from Each Cluster
1	133	l Human beta-defensin-2 in oral cancer with opportunistic Candida infection. l Effects of cytokines and heat shock on defensin levels of cultured keratinocytes. l Equine beta-defensin-1: full-length cDNA sequence and tissue expression.
2	37	l C-terminal domain of human CAP18 antimicrobial peptide induces apoptosis in oral squamous cell carcinoma SAS-H1 cells. l Protective efficacy of CAP18106-138-immunoglobulin G in sepsis. l Effect of human cationic antimicrobial protein 18 Peptide on endotoxin-induced uveitis in rats.
3	10	l Structure-based design of an indolicidin peptide analogue with increased protease stability. l The leader peptide is essential for the post-translational modification of the DNA-gyrase inhibitor microcin B17. l Characterization of the promoters of the guinea pig neutrophil cationic peptide-1 and -2 genes.
4	9	l Genetic and functional analysis of a PmrA-PmrB-regulated locus necessary for lipopolysaccharide modification, antimicrobial peptide resistance, and oral virulence of Salmonella enterica serovar typhimurium. l A small protein that mediates the activation of a two-component system by another two-component system. l PmrA-PmrB-regulated genes necessary for 4-aminoarabinose lipid A modification and polymyxin resistance.
5	673	l Immune activation of apolipophorin-III and its distribution in hemocyte from Hyphantria cunea. l Radiolabeled compounds in diagnosis of infectious and inflammatory disease. l Toll and Toll-9 in Drosophila innate immune response.
6	19	l Cryptdin 3 forms anion selective channels in cytoplasmic membranes of human embryonic kidney cells. l Activation of Paneth cell alpha-defensins in mouse small intestine. l Homodimeric theta-defensins from rhesus macaque leukocytes: isolation, synthesis, antimicrobial activities, and bacterial binding properties of the cyclic peptides.
7	46	l Effect of hepcidin on intestinal iron absorption in mice. l Hepcidin, the negative regulator of iron absorbtion l Increased hepcidin expression and hypoferraemia associated with an acute phase response are not affected by inactivation of HFE.
8	87	l Structural dissection of a highly knotted peptide reveals minimal motif with antimicrobial activity. l Factors contributing to the potency of antimicrobial cationic peptides from the N-terminal region of human lactoferrin. l Antimicrobial evaluation of nocathiacins, a thiazole peptide class of antibiotics.

9	3571	l Induction of transient ion channel-like pores in a cancer cell by antibiotic Peptide. l New indolicidin analogues with potent antibacterial activity. l Synthesis and HIV-1 integrase inhibitory activity of dimeric and tetrameric analogs of indolicidin.
10	23	l Inhibition of HIV infection by CXCR4 and CCR5 chemokine receptor antagonists. l An antiparallel beta-sheet and a beta-turn characterize the structure of antiviral HIV-1 peptide T140, as revealed by 2D NMR and MD Simulations. l Development of selective antagonists against an HIV second receptor

Table 2. Number of Clusters=10. In each cluster, three example articles with title are listed.

cluster	No. of documents & Examples of document titles from each cluster	cluster	No. of documents
1	80, e.g. l Evaluation of antimicrobial peptide nisin as a safe vaginal contraceptive agent in rabbits: in vitro and in vivo studies. l Encapsulation of nisin and lysozyme in liposomes enhances efficacy against Listeria monocytogenes. l Increased ATPase activity is responsible for acid sensitivity of nisin-resistant Listeria monocytogenes ATCC 700302.	11	85
2	20 , e.g. l In vivo evolution of X4 human immunodeficiency virus type 1 variants in the natural course of infection coincides with decreasing sensitivity to CXCR4 antagonists. l Development of selective antagonists against an HIV second receptor. l Inhibition of HIV infection by CXCR4 and CCR5 chemokine receptor antagonists.	12	8
3	52	13	64
4	31	14	11
5	46	15	5
6	191	16	81
7	30	17	13
8	91	18	2511
9	9	19	612
10	17	20	651

Table 3. Number of Clusters=20. Due to the large amount of clusters, only the first two clusters have three example articles with title.

cluster	No. of documents & Examples of document titles from each cluster	cluster	No. of documents
1	9 , e.g. l Fulminant Listeria monocytogenes meningitis complicated with acute hydrocephalus in healthy children beyond the newborn period. l Involvement of lipooligosaccharides of Haemophilus influenzae and Neisseria meningitidis in defensin-enhanced bacterial adherence to epithelial cells. l Management of pneumococcal meningitis.	26	20
2	4 , e.g. l The apoptotic protein tBid promotes leakage by altering membrane curvature. l Analogs of the antimicrobial peptide trichogin having opposite membrane properties. l Structure-function relationship of model Aib-containing peptides as ion transfer intermembrane templates.	27	41
3	6	28	9
4	299	29	1
5	44	30	70
6	25	31	8
7	17	32	55
8	123	33	1080
9	991	34	4
10	12	35	9
11	3	36	2
12	31	37	31
13	63	38	70
14	30	39	169
15	11	40	65
16	50	41	15
17	64	42	6
18	12	43	10
19	162	44	31
20	13	45	11
21	19	46	485
22	13	47	21
23	4	48	7
24	11	49	58
25	10	50	304

Table 4. Number of Clusters=50. Due to the large amount of clusters, only the first two clusters have three example articles with title.

REFERENCES

1. Baeza-Yates, R. and Ribeiro-Neto, B. (1999) Modern Information Retrieval. Addison-Wesley, Wokingham, UK.
2. Benjamin, C.M., Wang, F.K., Ester, M. (2003) Hierarchical Document Clustering Using Frequent Itemsets, SIAM International Conference on Data Mining 2003.
3. Boman, H. G. (Ed.) (1994) Antimicrobial Peptides, Ciba Foundation Symposium 186, John Wiley & Sons Ltd.
4. Edict virtual language centre, available at: http://www.edict.com.hk/TextAnalyser/wordlists.htm (as of Jan 27, 2005).
5. Han, J. W. and Kamber, M. (2000) Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers.
6. Iliopoulos, I., Enright, A. J., Ouzounis, C. A. (2001) TEXTQUEST: Document Clustering of Medline Abstracts for Concept Discovery in Molecular Biology, *Proc. PSB.*
7. Kogan, J., Nicholas, C., and Teboulle, M. (2003) Clustering Large and High Dimensional Data, ACM Conference on Information and Knowledge Management tutorial, November 2-8, New Orleans, Louisiana, USA.
8. Mathiak, B. and Eckstein, S. (2004) Five Steps to Text Mining in Biomedical Literature, *Proceedings of the Second European Workshop on Data Mining and Text Mining in Bioinformatics*, Italy, 47-50.
9. McQueen, J. (1967) Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematics, Statistics and Probability*, 1:281-298.
10. Nahm, U. Y. (2001) *A Roadmap to Text Mining and Web Mining*, available at: http://www.cs.utexas.edu/users/pebronia/text-mining/.
11. Porter, A. (2002) Text Mining, Review of TPAC Technologies for ONR, ASDL-Aug. 2002, available at: http://intelligent-web.org/wsm/.
12. Porter, M.F. (1980) An algorithm for suffix stripping, *Program*, **14**(3): 130-137.
13. PubMed (2005), provided by the National Library of Medicine, available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi (as of Jan 27, 2005).
14. PubMed Basics (2004), developed by NN/LM staff, funded by NLM, available at: http://nnlm.gov/nnlm/online/pubmed/pmtri.pdf.
15. SAS products and solutions (2005), Analytical intelligence: data and text mining, available at: http://www.sas.com/technologies/analytics/datamining/index.html.
16. SPSS 12.0 for Windows, SPSS Inc., http://www.spss.com.
17. Weiss, S., White, B., Apte, C. (2000) Lightweight Document Clustering, IBM Research Report RC-21684, available at: http://citeseer.ist.psu.edu/cache/papers/cs/14745/http:zSzzSzwww.research.ibm.comzSzdarzSzpaperszSzpdfzSzweiss_ldc_with_cover.pdf/weiss00lightweight.pdf.
18. Zhao, Y. and Karypis, G. (2002) Clustering in Life Sciences, technical report, Computer Science and Engineering, University of Minnesota, available at: https://wwws.cs.umn.edu/tech_reports/index.cgi?selectedyear=2002&mode=printreport&report_id=02-016.