A Document Clustering and Ranking System for Exploring MEDLINE Citations

Yongjing Lin, M.S.[1], Wenyuan Li, Ph.D.[1], Keke Chen, Ph.D.[3], Ying Liu, Ph.D.[1,2,*]

Laboratory for Bioinformatics and Medical Informatics,

[1]Department of Computer Science

[2]Department of Molecular and Cell Biology

University of Texas at Dallas, Richardson, TX 75083

[3]Yahoo!, 2811 Mission College Blvd., Santa Clara, CA 94086

*: To whom correspondence should be addressed
Ying Liu, Ph.D.
Department of Computer Science
P.O. Box 830688; MS EC31
University of Texas at Dallas
Richardson, TX 75083-0688

Tel: 972-883-6621
Fax: 972-883-2349
E-mail: ying.liu@utdallas.edu

Abstract

Objective: A major problem faced in biomedical informatics involves how best to present information retrieval results. When a single query retrieves many results, simply showing them as a long list often provides poor overview. With a goal of presenting users with reduced sets of relevant citations, this study developed an approach that retrieved and organized MEDLINE citations into different topical groups and prioritized important citations in each group.

Design: A text mining system framework for automatic document clustering and ranking organized MEDLINE citations following simple PubMed queries. The system grouped the retrieved citations, ranked the citations in each cluster, and generated a set of keywords and MeSH terms to describe the common theme of each cluster.

Measurements: Several possible ranking functions were compared, including citation count per year (CCPY), citation count (CC), and journal impact factor (JIF). We evaluated this framework by identifying as "important" those articles selected by the Surgical Oncology Society.

Results: Our results showed that CCPY outperforms CC and JIF, i.e., CCPY better ranked important articles than did the others. Furthermore, our text clustering and knowledge extraction strategy grouped the retrieval results into informative clusters as revealed by the keywords and MeSH terms extracted from the documents in each cluster.

Conclusions: The text mining system studied effectively integrated text clustering, text summarization, and text ranking and organized MEDLINE retrieval results into different topical groups.

Introduction

MEDLINE is a major biomedical literature database repository that is supported by the U.S. National Library of Medicine (NLM). It has now generated and maintained more than 15 million citations in the field of biology and medicine, and incrementally adds thousands of new citations every day [1]. Researchers can no longer keep up-to-date with all the relevant literature manually, even for specialized topics. As a result, information retrieval tools play essential roles in enabling researchers to find and access relevant papers[2]. Frequently, biomedical researchers query the MEDLINE database and retrieve lists of citations based on given keywords. PubMed, an information retrieval tool, is one of the most widely-used interfaces to access the MEDLINE database. It allows Boolean queries based on combinations of keywords and returns all citations matching the queries. Many advanced retrieval methods, such as GoPubMed[3] and Textpresso[4], also use natural language processing methods (i.e., entity recognition and part-of-speech tagging) to better identify documents relevant to a query [2]. Even with these improvements, significant challenges remain to efficient and effective utilization of *ad hoc* information retrieval systems such as PubMed[2].

Information Retrieval

Information retrieval methods attempt to identify, within large text collections, the specific text segments (such as full text articles, their abstracts, or individual paragraphs or sentences) whose content pertains to specified certain topics or to users' expressed needs[2, 5]. Such topics or needs are often stated in user-defined queries. Information retrieval systems typically employ one of two popular methodologies -- the Boolean model and the vector model. The Boolean model, used by virtually all commercial information retrieval systems, relies on Boolean logical operators and classical set theory. Documents searched and user queries both comprise sets of

terms, and retrieval occurs when documents contain the query terms. The vector model, on the other hand, represents each document as a vector of index terms (such as keywords). The set of terms is predefined, for example, as the set of all unique words occurring across all documents in the overall corpus. A weighting scheme, such as term frequency inverse document frequency (TFIDF), assigns a value to each term occurring in each document. [6] A similarity metric determines how well a document matches a query, calculated, for example, by comparing the deviation of angles between each document vector and the original query vector, where the query is represented as the same kind of vector as the documents[7].

Challenges for PubMed Information Retrieval

The goal of PubMed, like all other search engines, is to retrieve citations considered relevant to a user query. Modern search engine developers have devoted great effort in optimizing retrieval result rankings, hoping to place the most relevant ones at the top of the ranking list. Nevertheless, no ranking solution is perfect, due to the inherent complexity of ranking search results[8]. One aspect of this complexity derives from widely different possible query types. Narrow topic queries, or specific topic queries, retrieve relatively small numbers of citations from MEDLINE. For example, for the query "*BRCA*", PubMed returns 722 citations[9]. On the other hand, broad topic queries, or general topic queries, return large numbers of citations (thousands or more) from MEDLINE. For example, for the query "*Breast cancer*", PubMed returns 96,292 citations[9]. Manual reading, summarizing, or organizing such large numbers of articles is overwhelming. The vast majority of MEDLINE users show poor patience with large retrieval sets from broad topic queries. Such users commonly browse through the first screen or even the first ten results hoping to find the right answers for their queries [10]. PubMed provides some sorting mechanisms to rank citations, such as sorting by publication date, author name, or

journal. Furthermore, a pre-calculated set of PubMed citations that are closely related to a user-selected article can also be retrieved. The related articles will be displayed in ranked order from most to least relevant, with the "linked from" citation displayed first[11]. However, the "Related Articles" are citations which are related to a selected MEDLINE citation, and may or may not be relevant to a user's original query. Therefore, to help users identify papers of interest more easily and quickly, will require more advanced rankings and additional information about the relevance of citations to a submitted query.

Another difficulty in ranking search results is that the relevance of a citation is a subjective concept. In fact, the same set of keywords may abstract different user needs according to the context in which the user formulates the query [8, 10]. For example, both a researcher interested in finding genetic-study related papers and another researcher interested in finding the latest cancer treatments might issue a MEDLINE query "Breast cancer". Despite the differences in their initial interests, both researchers might seek out papers on breast cancer recurrences. PubMed provides some filtering functions to reduce the number of citations retrieved. However, even advanced PubMed queries with Boolean logic cannot always properly structure the search results [3].

In a recent review paper, Jensen et al. (2006) [2] indicated that presenting information retrieval results to users constitutes a sentinel problem in biomedical literature retrieval. When a single query retrieves a large number of citations, simply showing them as a long list often provides a poor overview. Furthermore, personal efforts and experience are necessary to extract the desirable biological knowledge from the retrieved literature. Finding elegant and accurate ways to extract the desired information could help users, particularly novice users, select and analyze a focused, reduced, relevant set of citations [12, 13].

Using Clustering and Ranking to Boost Information Retrieval

Biologists urgently require efficient systems to help them find the most relevant and important articles from the expanding biological literature [12, 14, 15]. One approach to this dilemma is to apply two post-processing techniques, clustering and ranking, to organize the retrieved documents into different topical groups based on semantic information. In this way, users can select, analyze, and focus on only a reduced set of citations in one or more topical groups of interest. Several projects have developed effective and efficient clustering technologies for Web search result organization[16-19]. Vivisimo[20] and Eigencluster[21] are working and available demonstrations of search-and-cluster engines. However, how to use both clustering and ranking technologies to improve the search result presentation has not been well studied[10].

Next-generation information retrieval tools should take advantage of clustering and ranking technologies[10]. Given a set of documents as input, clustering techniques group them into subsets based on similarity. Not only is clustering useful when applied in the presence of broad queries, but it also can improve the search experience by labeling the clusters with meaningful keywords or sentences [8] – a very useful alternative to a long, flat list of search results. Therefore, clustering can boost user queries by extracting and displaying hidden knowledge from retrieved texts. Ranking is a process which estimates the quality of a set of results retrieved by a search engine. Traditional information retrieval has developed Boolean, probabilistic, and vector-space models for ranking retrieved documents based on their contents.

Clustering and ranking are closely related, but few studies have deeply explored this relationship for biomedical literature retrieval. Some claim that clustering and ranking form a mutually reinforcing relationship. A good ranking strategy can provide a valuable information basis for clustering, and conversely, a good clustering strategy can help to rank the retrieved

results by emphasizing hidden knowledge content not captured by traditional text-based analyses. In addition, clustering algorithms can be used to extract, on the user's behalf, knowledge which goes beyond the traditional flat list [10].

Text Mining Systems to Improve PubMed Retrieval

Clustering of MEDLINE abstracts has been studied for gene function analysis[22-27] and concept discovery[28, 29]. A few systems have been proposed to present PubMed retrieval results in a user-friendly way other than a long list, most of which are based on pre-defined categories. GoPubMed [3] categorizes PubMed query results using to Gene Ontology (GO) terms. Textpresso[4] is an information retrieval system that operates on a collection of full text papers on *Caenorhabditis elegans.* It classifies the papers into about 30 high-level categories, some of which are derived from GO. Another system, XplorMed [30], maps PubMed query results to eight main Medical Subject Headings (MeSH) categories and extracts keywords and their co-occurrences. All such past systems have focused on grouping PubMed results into predefined categories, using classification techniques in which an *a priori* taxonomy of categories is available, rather than clustering techniques. Clustering differs from classification in that the categories are part of the discovered output, rather than predefined at input. When no pre-imposed classification scheme is available, automatic clustering may critically benefit users by organizing large retrieval sets into browsable groups [10]. By comparison, although current systems can classify search results into pre-defined categories, within each category, PubMed results still consist of long lists without importance-related ranking..The need exists for a system to help biomedical researchers in quickly finding relevant, important articles related to their research fields.

This paper describes a text mining system that automatically clusters PubMed query results into various groups where each group contains relevant articles, extracts the common topic for each group, and ranks the articles in each group. To the authors' knowledge, it is one of the first systems that integrates several text mining techniques, namely, text clustering, text summarization, and text ranking. The conceptual clustering component of the system takes an inductive machine learning approach[31, 32]. There are two steps in conceptual clustering: the first is an aggregation phase, which clusters documents into different groups, while the second is a characterization phase, which obtains the description of each cluster. The proposed system applies text clustering for the first phase, with text summarization and text ranking for the second phase.

System Framework

Figure 1 provides a high-level overview of the system, which proceeds through five phases: a) query submission and document retrieval; b) preprocessing the retrieved documents; c) determining the number of clusters and partitioning the document set into clusters; d) extracting the common topic for each cluster; e) identifying the most important and relevant articles in each cluster.

Query Submission and Document Retrieval

The system starts with submission of a query to the PubMed website. The PubMed search queries used in the current study are the 10 PubMed queries provided by Bernstam et al. (2006)[12] without field restrictions (see Online Supplemental Materials at www.jamia.org). The retrieved documents (abstracts) from each query are stored as single XML-format files. Since each retrieved PubMed document comprises one abstract, the authors use the words *document*,

abstract, and *article* interchangeably in this manuscript. Each XML file is then parsed, with the title, abstract, and MeSH term fields retained for further analysis. If a MEDLINE record does not have an abstract, but has an "otherabstract"[33], then the "otherabstract" was used as the study's version of the abstract. Lacking both, the title of the record was analyzed.

Preprocessing

The preprocessing phase plays a critical role in the subsequent clustering and concept extraction steps. Abstracts were broken during preprocessing into *tokens* which, in this paper, mean single words (or terms). Word stemming truncated suffixes so that words having the same root (e.g., activate, activates, and activating) collapse to the same single word for frequency counting. Our work applied the Porter stemmer for this task. [34]. Stop lists were used to filter out non-scientific English words. We developed a stop list based on an online dictionary of 22,205 words [25].

The standard term frequency-inverse document frequency (TFIDF) function was used [6] to assign weights to each word in each document. Then each document was modeled as an *N*-dimensional TFIDF vector, where *N* is the number of distinct words in all of the abstracts. Formally, a document was a vector ($tfidf_1$, $tfidf_2$, … , $tfidf_N$), where $tfidf_i$ is the tfidf value of word *i*. Then a document-by-word matrix was built, in which each row represented a word, and each column represented a document. The values in the matrix are the TFIDF values. If a word did not appear in a document, then zero appeared in the corresponding cell in the matrix (Figure 2a) [26].

Text Clustering

The document-by-word matrix was normalized using cosine normalization [26] (Figure 2b) and then used as input for the clustering step. Document clustering has been widely studied in the

text mining research area. Common methods take one of two approaches: document partitioning and hierarchical clustering. Hierarchical clustering methods organize the document set into a hierarchical tree structure, with clusters in each layer [35]. However, the clusters do not necessarily correspond to a meaningful grouping of the document set. [36] By contrast, partitioning methods can produce clusters of documents that are better than those produced by hierarchical clustering methods. Comparative studies have shown that partitioning algorithms outperform hierarchical clustering algorithms, and suggested that partitioning algorithms should be well-suited for clustering large document datasets -- due not only to their relatively low computational requirements, but also to comparable or even better clustering performance [37]. Therefore, the current study employed partitioning algorithms for clustering PubMed query results. One major drawback of partitioning algorithms is that they require prior knowledge of the number of clusters in a given data set. In our system, we applied the authors' recently proposed new algorithm, *Spectroscopy* [38, 39], to estimate the number of clusters in a document set.

Spectroscopy: Spectroscopy is a novel algorithm which can effectively predict the clustering characteristics of a text collection before the actual clustering algorithm is performed [38, 39]. It applies the techniques of spectral graph theory to data sets by investigating only a small portion of the eigenvalues of the data. Since spectral techniques have been well-studied and constitute a mature field in computation, there are a number of applicable efficient computational methods. Particularly in the case where we are only interested in a small number of eigenvalues and the term-document text data is rather sparse, numerical computation software such as LANSO[40] and ARPACK [41] can obtain results efficiently. (Please see JAMIA online data supplement at www.jamia.org for listing of pseudo-code of the spectroscopy algorithm.)

Document Clustering: Once the number of clusters is estimated, we apply the CLUTO[42] software to cluster a set of documents. We use the bisecting K-means technique because it performs better than the standard K-means approach.[43] CLUTO is a software package for clustering low and high dimensional data sets and for analyzing the characteristics of various clusters.[36] CLUTO has been shown to produce high quality clustering solutions in high dimensional data sets, especially those arising in document clustering. It has been successfully used to cluster data sets in many diverse application areas including information retrieval, commercial data, scientific data, and biological applications.[35]

Topic Extraction

For a given cluster of documents, our system generates summary sentences, a set of informative keywords, and a set of key MeSH terms, which can be used to describe the topic of that cluster.

To extract a summary sentence, the system uses a multi-document summary software, MEAD, which generates summaries using cluster centroids produced by a topic detection and tracking system[44]. Although MEAD can select sentences that are most likely to be relevant to the cluster topic, the summary sentences may not include all informative terms, therefore, they may not be able to precisely describe the topic of a cluster containing a large number of articles. In order to help users understand the topic of a cluster easily, the system also provides a set of keywords and a set of key MeSH terms that are specific and highly descriptive for a given cluster of documents.

Our system adopted a method that represents the relation between term set and document set as a weighted graph, and uses link analysis techniques like HITS (Hyperlink-Induced Topic Search)[45] to identify important terms. HITS, first proposed by Kleinberg[45], was originally used to rate Web pages for their "authority" and "hub" values. "Authority value" estimates the value of the content of a web page. "Hub value" estimates the value of a web page's links to other pages. The higher the authority value, the more important the web page is. The higher the hub value, the more connected the Web page is. These values can be used to rank Web pages. Authority and hub values are defined in terms of one another in a mutually recursive way. A page's authority value is computed as the sum of the scaled hub values of pages that point to it. A page's hub value is the sum of the scaled authority values of the pages it points to[46]. Therefore, HITS detects high scoring hub and authority Web pages using a reinforcement principle. This principle states that a Web page is a good authority if it is pointed to by many good hubs and that a good hub page points to many good authorities. The algorithm constructs a graph of nodes representing Web pages and the edges between them (representing hyperlinks) and each node receives an authority score and a hub score[47].

In order to extract informative keywords, a bipartite graph can be built between terms and documents as shown in Figure 2c. All keywords are represented as term nodes on the left-hand side of the bipartite graph, which have edges connecting to document nodes on the right-hand side of the bipartite graph (Figure 2c). "Authority" terms and "hub" documents can be discovered by the HITS algorithm. Then the reinforcement principle can be stated as "A term should have a high authority if it appears in many hub documents, while a document should have a high hub value if it contains many authority terms"[48]. Therefore, it is reasonable to infer that documents containing many "authority" terms must be "hubs" and core documents, while those

terms occurring in many “hub” and core documents must be “authority” and keywords. For a given cluster where documents are homogeneous and central to a topic, the HITS algorithm is effective in discovering keywords and core documents. It has been shown that the HITS algorithm is efficient enough for a Web search engine and therefore it is fast enough for the current setting.

Similarly, our approach also represents the relation between a MeSH term set and a document set as a weighted graph and applies the HITS algorithm to identify the important MeSH terms. (Please see the online supplemental materials at www.jamia.org for the pseudo-code of the project’s HITS implementation.)

Document Ranking

In the document ranking step, the goal is to identify articles that are important as well as relevant to the topic of a cluster. Our approach focuses on the citation count per year of a given article. A highly cited article has affected the field more than an article that has never been cited. Therefore, it is reasonable to consider the citation count as an important factor in ranking the articles. Bernstam et al. [12] compared eight ranking algorithms, simple PubMed queries, clinical queries (sensitive and specific versions), vector cosine comparison, citation count, journal impact factor, PageRank, and machine learning algorithms based on polynomial support machines. They concluded that citation-based algorithms are more effective than non-citation-based algorithms in identifying important articles. Our approach uses citation count per year instead of simple citation count because an article that was relatively unimportant and published several decades ago can, over time, accumulate more citations than would an important article that was published very recently. We compared our ranking algorithm based on citation count per year with simple citation count and journal impact factor. Article citation count and journal impact factor were

obtained from the Science Citation Index (SCI®) and the Journal Citation Report (JCR) [49]. If an article does not have a citation count in SCI, then its citation count and citation count per year are taken as zero. If a journal does not have a journal impact factor in JCR, then the journal impact factor of the articles published in that journal is taken as zero. In general, the system's ranking strategy is: the higher the citation counts an article has, the more important it is; the larger the citation count per year an article has, the more important it is; the larger the journal impact factor a journal has in which an article was published, the more important this article is.

Experiments

Gold standard test set

We used the Society of Surgical Oncology (http://www.surgonc.org) Annotated Bibliography (SSO_AB) as a gold standard. The SSO_AB is maintained by the Society of Surgical Oncology (SSO) and is grouped into 10 categories, each regarding a kind of cancer. Each category was compiled by a single expert and reviewed by a panel of experts on that particular topic [12, 14]. The articles in SSO_AB are chosen by experts as important. The latest edition of SSO_AB is dated October 2001 because maintaining the annotated bibliography requires a great amount of human effort. It contains 458 unique articles cited by MEDLINE. Publication dates range from March 1969 to September 2001. Therefore, in this study, we restricted the PubMed query to this date range. A perfect ranking algorithm should return the SSO_AB articles at the top of the result set.

System performance evaluation

To evaluate the effectiveness of our system, we applied the Hit curve algorithm.[12] The Hit curve function, h(n), measures the number of important articles among the top *n* ranked results. If there are *k* important articles, then the ideal Hit curve will be a straight line with a slope of 1,

for 1<n<k-1, which becomes horizontal for n>k, after all k important articles have been retrieved.[12] For this paper, we chose to measure the number of important articles among the top 10, 20, 40, 60… ranked articles. The Hit curve provides an intuitive representation of an algorithm's performance for a given query, and can be averaged over a number of different queries.[12]

Results and discussion

As mentioned before, our aim was to implement a text mining and ranking system that allows the user to analyze the documents in a conceptually homogeneous way, as well as choose the most important and relevant documents. We tested all the 10 categories (10 types of cancers) defined by SSO_AB. In this paper, we only present the "breast cancer" results. The results of the other 9 cancers are included as Online Supplementary Materials at www.jamia.org. The size of the breast cancer result set was 77,784 with 65 unique important articles in SSO_AB.

Preprocessing Results

After the preprocessing stage for the breast cancer results, we obtained a term set with size of 55,712, and a sparse matrix with size of 1,379,417. Note that the document set is of size 77,784, which implies that each document has, on average, about 18 unique terms left after the preprocessing.

Clustering and Knowledge Extraction Results for Breast Cancer

The clustering procedure returned 6 clusters for the breast cancer set. Each cluster refers to a set of abstracts that are related by terms that co-occur among the different abstracts. The sentence summaries are very long, and not informative. Therefore, in this paper, only the top-ranked

keywords and MeSH terms are presented (Table 1). Cluster E has the largest number of important articles selected by SSO.

We extend the idea of the HITS (Hyperlink-Induced Topic Search) algorithm [45] in extracting keywords and MeSH terms, in which the ranking is based on the relationships among terms and documents. One of the limits of the HITS algorithm is that it relies on “global" information derived from all the vectors in the dataset, which is more effective for datasets consisting of homogeneously distributed vectors. However, the retrieved documents returned from PubMed consist of multiple distinguishable topics. We integrated the HITS algorithm with a clustering technique. Articles of the same topic are grouped into clusters such that the top ranked terms and documents can be identified efficiently. Table 1 illustrates that the clustering and topic extraction strategy performs well. The six clusters our system derived represent six distinct topical groups, which are revealed by the top-ranked keywords and MeSH terms:

1. Cluster A shows one common topic: molecular genetic studies of breast cancer, especially the genes *BRCA1*, *BRCA2*, p53, and p21.
2. As revealed by the top MeSH terms, such as “antineoplastic combined chemotherapy protocols”, “drug therapy”, and the chemotherapy drugs, paclitaxel, epirubicin, and cisplatin, we determined that Cluster B contains the articles which are related to chemotherapy.
3. The articles in Cluster C report research focusing on the role of hormones and growth factors, such as epidermal growth factor (EGF) and transforming growth factor (TGF) in breast cancer development, and on tamoxifen, an anti-estrogen (anti-hormonal) drug as a treatment for breast cancer [50, 51]. Estrogen promotes the growth of breast cancer cells, and tamoxifen blocks the effects of estrogen on these cells, slowing the growth of the patient’s

cancer cells that have estrogen receptors. As adjuvant therapy, tamoxifen helps prevent the original breast cancer from returning and also helps prevent the development of new cancers in the other breast.

4. The common topic of the articles in cluster D is population studies, including mammographic screening for breast cancer among different ethnic group. Cluster-related keywords include “American”, “African”, and “Hispanic”.
5. The articles in cluster E focus on recurrences of breast cancer and follow-up studies.
6. Cluster F represents the set of articles which report on treatment of breast cancer using monoclonal antibodies. MAb is one kind of monoclonal antibody which can target epidermal growth-factor receptors [52]. The anti-vascular endothelial growth factor (VEGF) antibody, which has been approved by Food and Drug Administration (FDA) for the treatment of colon cancer, is also able to achieve similar progress in the treatment of locally advanced breast cancer [53].

The clustering and topic extraction results for the other nine types of cancers appear in the Online Supplemental Materials at www.jamia.org. For all nine cancer types, the algorithm grouped articles into distinct clusters with specific topics.

Document Ranking Results

Figure 3 presents, using Hit curves, the document ranking results for the breast cancer document set (six clusters), In Figure 3, the x-axis represents the number of ranked documents, while the y-axis represents the numbers of SSO_AB articles found in the top ranked document list. All the ranking algorithms, CC, CCPY, and JIF, carried to the end would finish with all important articles found. To provide an upper bound, we also plotted the hit curve of an ideal ranking strategy, which would rank all the important articles at the top of the result list.

Therefore, the ranking algorithm corresponding to the hit curve that is "closest" to the ideal hit curve was the best algorithm[12].

Figure 3 shows that, as a ranking function, CCPY (Citation Count Per Year) outperforms CC (Citation Count) and JIF (Journal Impact Factor) as the CCPY hit curve is the "closest" to the ideal hit curve. In cluster E, which has the largest number of important articles (41 important articles), out of the 40 top ranked articles, 7 were important if ranked by CCPY, while only 3 were important if ranked by CC, and no important article appeared in the top ranked documents if ranked by JIF. Similar results occurred in the other clusters. The document ranking results for the other nine types of cancers (see Online Supplementary Materials at www.jamia.org) also show that, as a ranking function, CCPY outperforms CC and JIF.

In order to show how much CCPY outperforms CC, for each cancer type, we calculated the average ranking of the important articles as:

$$\overline{R} = \frac{\sum_{i=1}^{n} r_i}{n} \qquad (1)$$

where $\overline{R}$ is the average ranking, n is the number of important articles identified by SSO, and r_i is the ranking of article i ranked by either CC or CCPY. For example, there were 65 important articles in the "breast cancer" retrieval result (n = 65). After all the "breast cancer" documents were ranked either by CC or CCPY, we found the ranks of all the 65 articles and calculated the average ranking $\overline{R}$. Then, an improvement rate (IR) was calculated as:

$$IR = \frac{\overline{R}_{CC} - \overline{R}_{CCPY}}{\overline{R}_{CC}} \qquad (2)$$

where $\overline{R}_{CC}$ is the average ranking of the important articles ranked by CC and $\overline{R}_{CCPY}$ is the average ranking of the important articles ranked by CCPY. Table 2 lists these results. For all ten

different types of cancers, CCPY improved the ranking compared with CC (from ~23% to more than 46%). For each cancer type, two important articles (their PMIDs are listed) with the largest IRs were identified. The CC rank, CCPY rank, and the year when the article was published are also listed. We note that these articles were published relatively recently (closer to 2001 within the study's PubMed query date range from March 1969 to September 2001). The ranking information of 65 important "Breast cancer" articles is shown in Table 3. The ranking information of the other 9 cancer types appears in the Online Supplementary Materials at www.jamia.org.

Bernstam et al. (2006)[12] reported that citation-based algorithms are more effective than non-citation-based algorithms in identifying important articles. The current study showed that for our purposes, citation count per year worked better than simple citation count. We hypothesized that an article which was relatively unimportant and published several decades ago may accumulate more absolute citations than a more important article published just recently.

The importance of a paper to a field varies over time. A citation decay pattern has been discovered in bibliometric studies of published scientific literature[54, 55]. Burton and Kebler coined the term, "Citation half-life", with respect to scientific and technical literature[56]. Citation half-life can be defined as the number of years required to encompass the most recent 50% of all references made[54]. A paper with a longer half-life might have more enduring value than a paper with a shorter half-life[54, 57]. Therefore, it seems that citation half-life may be a better measure for identifying important articles than simple citation count or citation count per year. However, citation half-life is highly related to the journal publication frequency, journal age, language, country of publication, length of the paper, and subject category[54, 57]. Papers that are longer and in sciences or subfields that are growing fast are more likely to be cited over a longer period, and

thus have longer citation half-lives[57]. How to use the citation half-life information for document ranking requires further investigation.

Algorithm Computational Efficiency

Retrieval and mining of large document sets is computationally intensive. However, by choosing efficient clustering, summarization and ranking algorithms, the system studied performed acceptably fast. The study utilized a machine with the Windows XP operating system, a 3.0GHz CPU, and a 2.0GB RAM. Table 4 shows the computing time for each of the five phases of system operation. Only the "breast cancer" data set computing time results are shown because that analysis involved the largest number of documents retrieved from MEDLINE (77,784 citations). The first system analytic phase was query submission and document retrieval. Since this phase relies on PubMed to retrieve the documents, the time is not listed in Table 4. Table 4 shows the times to conduct the other four phases, text preprocessing, document clustering, document ranking and topic extraction. The time to extract the topic for each cluster was only the time to generate the top-ranked keywords and MeSH terms, because summarization of the document set using MEAD took very long (about 40 minutes per cluster, not shown in the table). Furthermore, the sentence summary was not as informative as the top-ranked keywords or MeSH terms, so we used the top-ranked keywords and MeSH terms to represent the common topic of the documents in each cluster. From Table 4, we can see that the pre-processing phase took most of the time (45 seconds out of a total 70.06 seconds). In the future work, we plan to have a local copy of MEDLINE and index each abstract with its keywords. Then, the pre-processing time will be significantly reduced (to about 1-2 seconds). As a result, the system will cluster and rank MEDLINE abstracts in a more efficient and faster manner.

Conclusions and Future work

The text mining system we presented, which integrates several text mining techniques, namely, text clustering, text summarization, and text ranking, can effectively organize PubMed retrieval results into different topical groups. It offers users the potential to focus on reduced sets of articles for which they have greater interest, instead of reading through the long list of citations returned by a query. An additional finding of the study involved demonstrating that as a ranking function, citation count per year outperforms simple citation count and journal impact factor.

In this study, authors developed a system framework to explore MEDLINE citations to assist biomedical researchers in identifying important articles according to different topics. There are several areas in which future efforts might improve our system. One such area is the text summarization part of the system. In this study, the sentences derived from MEAD, a multi-document summarizer, were not informative compared to the keywords generated by HITS. We plan to include the Gene Ontology (GO) information in the next iteration of the text summarization process. For each group of articles, besides the informative keywords, the appropriate GO terms would also be listed. New algorithms will be designed and tested to find appropriate GO terms to represent the topic of a given group of articles. A second potential area for improvement would be to develop a new ranking function based on a combination of the CCPY, JIF and other factors. Third, an active learning system might be employed to utilize user-provided feedback to refine clustering and ranking based on users' suggestions. Fourth, a parallel and distributed algorithm might improve system performance by carrying out document clustering, ranking, and topic extraction in a parallel and distributed way. Some open-source

distributed computing infrastructure, such as hadoop[58], will be explored. Last, a Web-based software system can be developed and deployed for remote researchers to use as shown in Fig. 4.

Acknowledgement

The authors are grateful to the editor and anonymous reviewers for a number of suggestions for improvement. We thank Phil Bachman and Lavin Urbano for helpful discussions.

References

1. Shef M, Epple, A.,Werner, T. The next generation of literature analysis: Integration of genomic analysis into text mining. *Briefings in Bioinformatics.* 2005;6:287-297.

2. Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. *Nature Reviews Genetics.* 2006;7:119-129.

3. Doms A, Schroeder M. GoPubMed: exploring PubMed with the gene ontology. *Nucleic Acids Research.* 2005;33:W783-W786.

4. Muller H-M, Kenny EE, Sternberg PW. Textpresso: an ontology-based information retrieval and extracting system for biological literature. *PLoS Biol.* 2004;2:1984-1998.

5. Shatkay H. Hairpins in bookstacks: information retrieval from biomedical text. *Briefings in Bioinformatics.* 2005;6:222-238.

6. Salton G, Buckley C. Text weighting approaches in automatic text retrieval. *Information Processing and Management.* 1988;24:513-523.

7. Wilbur WaY, Y. An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology text. . *Comput. Biol. Med.* 1996;26:209-222.

8. Gulli A, Signorini A. Building an open source meta search engine. *Proceddings of 14th International World Wide Web Conference, Chiba, Japan.* 2005:1004-1005.

9. PubMed. (**http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed**) accessed on Feb. 14th, 2007.

10. Gulli A. *On two web IR boosting tools: clustering and ranking* [Ph.D thesis], University Degli Stud di Pisa, Italy; 2006.

11. **http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helppubmed.section.pubmedhelp.Searching_PubMed**.

12. Bernstam EV, Herskovig JR, Aphinyaphongs Y, Aliferis CF, Sriram, MG, Hersh WR. Using citation data to improve retrievl from MEDLINE. *JAMIA.* 2006;13:96-105.

13. Hersh WR, Hickam D. How well do physicans use electronic information retrieval systems? A framework for investigation and systematic review. *JAMA.* 1998;280:1347-1352.

14. Aphinyanaphongs Y, Tsamardinos, I., Statnikov, A., Hardin, D., Aliferis, C.F. Text categorization models for high-quality article retrieval in internal medicine. *JAMIA.* 2005;12:207-216.

15. Cohen AM, Hersh, W.R. A survey of current work in biomedical text mining. *Briefings in Bioinformatics.* 2005;6:57-71.

16. Zamir O, Etzioni, O. Web document clustering: a feasibility demonstration. *Proceedings of the 19th International ACM SIGIR Conference on Research and Development of Information Retrieval (SIGIR'98).* 1998:46-54.

17. Leuski A, Allan, J. Improving interactive retrieval by combining ranked list and clustering. *Proceedings of RIAO, College de France.* 2000:665-681.

18. Zeng H-J, He, Q.-C., Chen, Z., Ma, W.-Y., Ma, J. Learning to cluster web search results. *Proceedings of the 27th International ACM SIGIR Conference on Research and Development of Information Retrieval (SIGIR'04).* 2004:210-217.

19. Cheng D, Kannan, R., Vempala, S., Wang, G. A divide-and merge methodology for clustering. *ACM Transactions on Database Systems.* 2006;31:1499-1525.

20. Vivisimo. **http://vivisimo.com**.

21. Eigencluster. **http://eigencluster.csail.mit.edu/about.html**.

22. Chaussabel D, Sher, A. Mining microarray expression data by literature profiling. *Genome Biology.* 2002;13:R55.

23. Glenisson P, Coessens, B., Van Vooren, S., Mathys, J., Moreau, Y., De Moor, B. TXTGate: profiling gene groups with text-based information. *Genome Biology.* 2004;5:R43.

24. Homayouni R, Heinrich, K., Wei, L., Berry, M.W. Gene clustering by latent semantic indexing of MEDLINE abstracts. *Bioinformatics.* 2005;21:104-115

25. Liu Y, Navathe SB, Civera J, Dasigi V, Ram A, Ciliax BJ, Dingledine R. Text mining biomedical literature for discovering gene-to-gene relationships: a comparative study of algorithms. *IEEE/ACM Transactions on Computational Biology and Bioinformatics.* 2005;2:62-76.

26. Liu Y, Navathe SB, Pivoshenko A, Dasigi V, Dingledine R, Ciliax BJ. Text Analysis of MEDLINE for Discovering Functional Relationships among Genes: Evaluation of Keyword Extraction Weighting Schemes. *International Journal of Data Mining and Bioinformatics.* 2006;1:88-110.

27. Yang J, Cohen, A.M., Hersh, W.R. Functional Gene Group Summarization by Clustering MEDLINE Abstract Sentences. *AMIA Annu Symp Proc.* 2006:1151.

28. Iliopoulos I, Enright, A.J., Ouzounis, C.A. Textquest: document clustering of Medline abstracts for concept discovery in molecular biology. *Pac Symp Biocomput.* 2001:384-395.

29. Rindflesch TC, Tanabe, L., Weinstein, J.N., Hunter, L. EDGAR: extraction of drugs, genes and relations from the biomedical literature. *Pac Symp Biocomput.* . 2000(517-528).

30. Perez-Iratxeta C, Perez AJ, Bork P, Andrad MA. Update on XplorMed: a web server for exploring scientific literature. *Nucleic Acids Research.* 2003;31:3866-3868.

31. Fattore M, Arrigo P. Topical clustering of biomedical abstracts by self-organizing maps. Paper presented at: The fourth international conference on bioinformatics of genome regulation and structure, 2004.

32. Fisher DH. Knowledge acquisition via incremental conceptual clustering. *Machine Learning.* 1987;2:139-172.

33. **http://www.nlm.nih.gov/bsd/licensee/elements_descriptions.html**.

34. Porter M. An algorithm for suffix stripping. *Program.* 1980;14:130-137.

35. Zhao Y, Karypis G, Fayyad UM. Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery.* 2005;10:141-168.

36. Zhao Y, Karypis G. Empirical and theoretical comparison of selected criterion functions for document clustering. *Machine Learning.* 2004;55:311-331.

37. Zhao Y, Karypis G. Evaluation of hierarchical clustering algorithms for document datasets. Paper presented at: CIKM'02, 2002; McLean, VA.

38. Li W, Ng, W., Liu, Y., Ong, K. Enhancing the effectiveness of clustering with spectra analysis. *IEEE Transcations on Knowledge and Data Engineering.* 2007;To appear.

39. Li W, Ng W.-K., Ong K.-L., Lim E.-P. A spectroscopy of text for effective clustering. Paper presented at: PKDD, 2004.

40. LANSO. *Dept. of Computer Science and Industrial Liason Office, University of California at Berkeley.*

41. Lehoucq R, ,Sorensen, DC, Yang C. ARPACK User's Guide: solution of large-scale eigenvalue problems by implicitly restarted Arnoldi methods. Paper presented at: SIAM, 1998; Philadelphia, PA.

42. Karypis G. CLUTO: a clustering toolkit. *Technical report, Department of Computer Science, University of Minnesota.* 2003.

43. Steinback M, Karypis G, Kumar V. A comparison of document clustering techniques. *KDD Workshop on Text Mining.* 2000.

44. Radev DR, Jing H, Stys M, Tam D. Centroid-based summarization of multiple documents. *Information Processing and Management.* 2004;40:919-938.

45. Kleinberg J. Authoritative sources in a hyperlinked environment. Paper presented at: 9th Annual ACM-SIAM symposium on Discrete Algorithms, 1998; San Francisco, CA.

46. Kleinberg J, Lawrence, S. The structure of the web. *Science.* 2001;294:1849.

47. Janssens F, Moor, Bart De. Application of HITS algorithm to detect terms and sentences with high saliency scores. *Katholieke Universiteit Leuven, Department Elektrotechniek, Technical report ESAT-SISTA/TR 04-29.* 2003.

48. Zha H. Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. *SIGIR 2002.* 2002:113-120.

49. Science Citation Index. Journal citation reports. Philadelphia: Thomson-ISI. 2005.

50. Pollak M. IGF-I physiology and breast cancer. *Recent Results Cancer Res.* 1998;152:63-70.

51. Yee D. Targeting insulin-like growth factor pathways. *British Journal of Cancer.* 2006;94:465-468.

52. Sridhar SS, Seymour L, Shepherd FA. Inhibitors of epidermal-growth-factor receptors: a review of clinical research with a focus on non-small-cell lung cancer. *Lancet Oncol.* 2003;4:397-406.

53. Wu Y, Zhong, Z., Huber, J. et al. Anti-vascular endothelial growth factor receptor-1 antagonist antibody as a therapeutic agent for cancer. *Clin Cancer Res.* 2006;12(21):6573-6584.

54. Tsay M-Y. Library journal use and citation half-life in medical science. *Journal of American Society for Information Science.* 1998;49(14):1283-1292.

55. Peterson CJ. Citation analysis of astronomical literature: comments on citation half-lives. *Publications of the Astronomical Society of the Pacific.* 1988;100:106.

56. Burton RE, Kebler, R.W. The half-life of some scientific and technical literature. *American Documentation.* 1960;11:18-22.

57. Abt HA. Why some papers have long citation lifetimes. *Science.* 1998;395:756-757.

58. Hadoop. **http://lucene.apache.org/hadoop/**.

Table 1. “Breast cancer” document clustering and topic extraction results.

Clusters	# of important articles	Top Keywords	Top MeSH Terms
A	5	BRCA1, BRCA2, p53, ERBB, cell, BCL, DNA, tumor, cyclin, apoptosis, neu, mutant, p21, carcinoma, tumour, receptors, ovarian, suppressor, oncogene, chromosome, antibody, exon, ras, kinase	"mutation" "neoplasm proteins" "transcription factors" "dna, neoplasm" "molecular sequence data" "receptor, erbb-2" "base sequence" "tumor suppressor protein p53" "proto-oncogene proteins" "brca2 protein" "genes, brca1" "genes, p53" "ovarian neoplasms" "genes, tumor suppressor" "brca1 protein" "gene expression regulation, neoplastic" "immunohistochemistry" "dna mutational analysis"
B	1	IGF, IGFBP, insulin, cell, MCF, receptors, mitogen, plasma, serum, paclitaxel, affinity, estrogen, hs578t, tamoxifen, autocrine, mda, kinase, cancer, ligand, tumor, phosphorylation, apoptosis, circulating, TGF	"antineoplastic combined chemotherapy protocols" "fluorouracil" "cyclophosphamide" "doxorubicin" "methotrexate" "paclitaxel" "antineoplastic agents, phytogenic" "adult" "neoplasm metastasis" "middle aged" "aged" "antineoplastic agents" "epirubicin" "cisplatin" "antibiotics, antineoplastic" "infusions, intravenous" "taxoids" "vincristine" "treatment outcome" "dose-response relationship, drug" "drug therapy, combination" "clinical trials" "chemotherapy, adjuvant"
C	8	estrogen, receptors, tamoxifen, cell, estradiol, MCF, hormone, antiestrogens, progesterone, TGF, steroid, EGF, aromatase, tumor, endometrial, oestrogen, women, postmenopausal, progestin, androgen, PgR, mammary, cytosolic, pS2	"receptors, estrogen" "estradiol" "tamoxifen" "receptors, progesterone" "tumor cells, cultured" "estrogens" "estrogen antagonists" "cell division" "rats" "neoplasms, hormone-dependent" "menopause" "cell line" "antineoplastic agents, hormonal" "mice" "mammary neoplasms, experimental" "rna, messenger" "uterus" "progesterone" "kinetics" "cytosol" "aged"
D	8	women, mammography, mammogram, mammographic, cancer, fat, American, aged, BSE, pregnancy, interventions, lesion, OCs, disease, oral, African, abortion, dietary, diagnosed, birth, deaths, Hispanic, cervical, younger	"mammography" "mass screening" "adult" "aged" "risk factors" "middle aged" "age factors" "united states" "comparative study" "adolescent" "incidence" "questionnaires" "neoplasms" "male" "aged, 80 and over" "case-control studies" "risk" "breast" "breast diseases" "socioeconomic factors" "cohort studies" "sensitivity and specificity"
E	41	metastases, axillary, recurrence, lymph, bone, carcinoma, survival, SLN, DCIS, lesion, tumor, mastectomy, disease, metastasis, metastatic, ductal, cytology, excision, cell, cancer, women, lobular, nodal	"lymphatic metastasis" "neoplasm recurrence, local" "aged" "adult" "combined modality therapy" "middle aged" "mastectomy" "follow-up studies" "neoplasm staging" "retrospective studies" "prognosis" "axilla" "lymph node excision" "aged, 80 and over" "male" "lymph nodes" "mastectomy, segmental" "survival rate" "time factors" "carcinoma, intraductal, noninfiltrating" "radiotherapy dosage" "carcinoma, ductal, breast" "carcinoma"
F	2	cell, antigens, tumor, antibody, carcinoma, MMP, epithelial, MAb, CSF, MCF, CD34, MUC1, mammary, receptors, membrane, cadherin, lymphocytes, marrow, UPA, TNF, HLA, VEGF, kinase	"mice" "tumor cells, cultured" "antibodies, monoclonal" "antigens, neoplasm" "molecular sequence data" "cell line" "amino acid sequence" "mice, nude" "rna, messenger" "cell division" "immunohistochemistry" "base sequence"

Table 2. The comparison of CCPY and CC for important article ranking

category	# of Important article[1] (Total # of articles retrieved[2])	Average Ranking[3]		Improvement Rate (%)[6]	Example[7]			
		CC[4]	CCPY[5]		PMID	CC Rank	CCPY Rank	Year Published
Breast Cancer	65 (77784)	4976.5	3069.844	38.3132	11157042	1655	381	2001
					11230499	1847	432	2001
Colorectal Cancer	39 (53686)	6938.838	4857.378	29.99723	11006366	49	10	2000
					11309435	1829	515	2001
Endocrine Cancer	72 (46981)	4402.014	3389.435	23.00264	10973383	1182	261	2000
					10458257	771	210	1999
Esophageal Cancer	34 (16359)	995.4688	641.2188	35.58625	10080844	3	0	1999
					11547741	41	9	2001
Gastric Cancer	47 (33938)	2178.298	1463.489	32.815	11547741	101	22	2001
					10080844	19	5	1999
Hepatobiliary Cancer	68 (76616)	7338.657	4123.269	43.8144	10636102	308	49	2000
					10973388	341	55	2001
Lung Cancer	42 (74189)	3413.585	2595.488	23.96593	10694600	1979	587	2000
					9187198	10	3	1997
Melanoma	46 (33074)	2968.891	1761.13	40.68054	11504745	63	8	2001
					11504744	133	26	2001
Pancreas Cancer	61 (25241)	2061.633	1131.933	45.09531	11297271	1099	238	2001
					11258776	1329	290	2001
Soft Tissue Sarcomas	22 (3193)	205.7619	110	46.54015	11230464	117	16	2001
					11230466	296	61	2001

[1]: The number of important articles in each type of cancers defined by SSO_AB;

[2]: The number of articles retrieved from MEDLINE;

[3]: Please refer to equation (1) for Average Ranking calculation;

[4]: CC: Citation Count;

[5]: CCPY: Citation Count per Year;

[6]: Please refer to equation (2) for Improvement Rate calculation;

[7]: For each type of cancer, two articles (PMIDs were listed) with largest improvement rate are identified.

Table 3: “Breast cancer” important article ranking results. There are 65 important articles.

PMID	Citation Count	Year Published	CC Rank	CCPY Rank	Improvement Rate (%)
11157042	140	2001	1655	381	76.98
11230499	132	2001	1847	432	76.61
11304779	92	2001	3405	916	73.10
10760307	222	2000	723	205	71.65
10684910	387	2000	231	68	70.56
10893286	204	2000	854	254	70.26
11208879	66	2001	5572	1665	70.12
11409797	61	2001	6128	1913	68.78
10683002	171	2000	1182	370	68.70
10659874	485	2000	140	44	68.57
10893287	164	2000	1276	407	68.10
10751498	163	2000	1288	411	68.09
11230466	52	2001	7398	2481	66.46
11420508	43	2001	9416	3374	64.17
10376613	286	1999	432	160	62.96
10764427	88	2000	3654	1408	61.47
10335782	378	1999	245	97	60.41
9887158	403	1999	213	86	59.62
10764431	72	2000	4858	1981	59.22
10904085	67	2000	5381	2216	58.82
11157012	32	2001	12677	5261	58.50
10320383	203	1999	858	357	58.39
10489948	220	1999	735	306	58.37
10334518	180	1999	1085	455	58.06
10768705	61	2000	6146	2588	57.89
10784640	56	2000	6750	2958	56.18
11254867	26	2001	15266	6918	54.68
11304777	24	2001	16787	7707	54.09
10561339	133	1999	1830	853	53.39
10561205	111	1999	2490	1207	51.53
10787083	39	2000	10493	5172	50.71
9747868	2021	1998	10	5	50.00
11032585	38	2000	10609	5317	49.88
10658521	83	1999	3936	2010	48.93
10477433	73	1999	4725	2469	47.75
9469327	233	1998	659	353	46.43
9605801	1721	1998	13	7	46.15
10901741	30	2000	13654	7355	46.13
11075239	29	2000	14122	7615	46.08
10601383	60	1999	6259	3392	45.81
9753708	782	1998	58	32	44.83
9704717	473	1998	146	81	44.52
10493623	47	1999	8439	4887	42.09
10646888	14	2000	23794	15753	33.79
10526280	27	1999	14816	9906	33.14
9395428	825	1997	53	37	30.19
9752815	978	1998	30	22	26.67
9145676	857	1997	45	34	24.44
8931609	274	1996	479	402	16.08
8635094	239	1996	632	549	13.13
8604907	217	1996	755	658	12.85
8614420	170	1996	1190	1050	11.76
10575423	3	1999	40149	35922	10.53
7799496	461	1995	159	154	3.14
7477145	592	1995	101	98	2.97
10832826	0	2000	50133	50133	0.00
7473814	33	1995	12423	12801	-3.04
7577477	33	1995	12389	12772	-3.09
7600278	42	1995	9710	10067	-3.68
8635050	180	1995	1081	1134	-4.90
7666458	201	1995	879	927	-5.46
7908410	661	1994	85	94	-10.59
8389654	62	1993	5998	7741	-29.06
1978757	1942	1990	12	23	-91.67

Table 4. “Breast cancer” data set computing time for each phase in our system.

		# of documents	Computing Time (in seconds)
Text pre-processing		77,784	~45
Text clustering (using CLUTO)		77,784	14.28
Keyword Extraction	Cluster A	8,212	0.77
	Cluster B	6,159	0.66
	Cluster C	13,122	1.057
	Cluster D	16,292	0.97
	Cluster E	21,005	1.25
	Cluster F	12,994	1.09
MeSH term Extraction	Cluster A	8,212	0.41
	Cluster B	6,159	0.33
	Cluster C	13,122	0.58
	Cluster D	16,292	0.66
	Cluster E	21,005	0.81
	Cluster F	12,994	0.59
Document ranking	Cluster A	8,212	0.20
	Cluster B	6,159	0.11
	Cluster C	13,122	0.27
	Cluster D	16,292	0.30
	Cluster E	21,005	0.44
	Cluster F	12,994	0.28
Total			70.06

Figure Legends:

Figure 1. Overview of the system.

Figure 2. The representation of the documents. (A) Document-by-word matrix. The values in the matrix are the term frequency-inverse document frequency (TFIDF) values. (B) Normalized document-by-word matrix. The document-by-word matrix is normalized using cosine normalization. (C) Bipartite graph representation of the normalized document-by-word matrix.

Figure 3. The Hit curves of the six clusters (A, B, C, D, E and F) derived from "breast cancer" data set. The x-axis represents the number of ranked documents, while the y-axis represents the SSO_AB articles found in the top ranked document list. (e.g. in cluster E, using CCPY as the ranking function, among the top 40 ranked document list, 7 of the documents are important articles.) CC- Citation Count; CCPY – Citation Count Per Year; JIF – Journal Impact Factor. As a ranking function, CCPY outperforms CC and JIF. To provide an upper bound, we also plot the hit curve of an ideal ranking strategy, which ranks all the important articles at the top of the result list. Therefore, the ranking algorithm corresponding to the hit curve that is "closest" to the ideal hit curve is the best algorithm.

Figure 4. An example of PubMed search result clustering and ranking

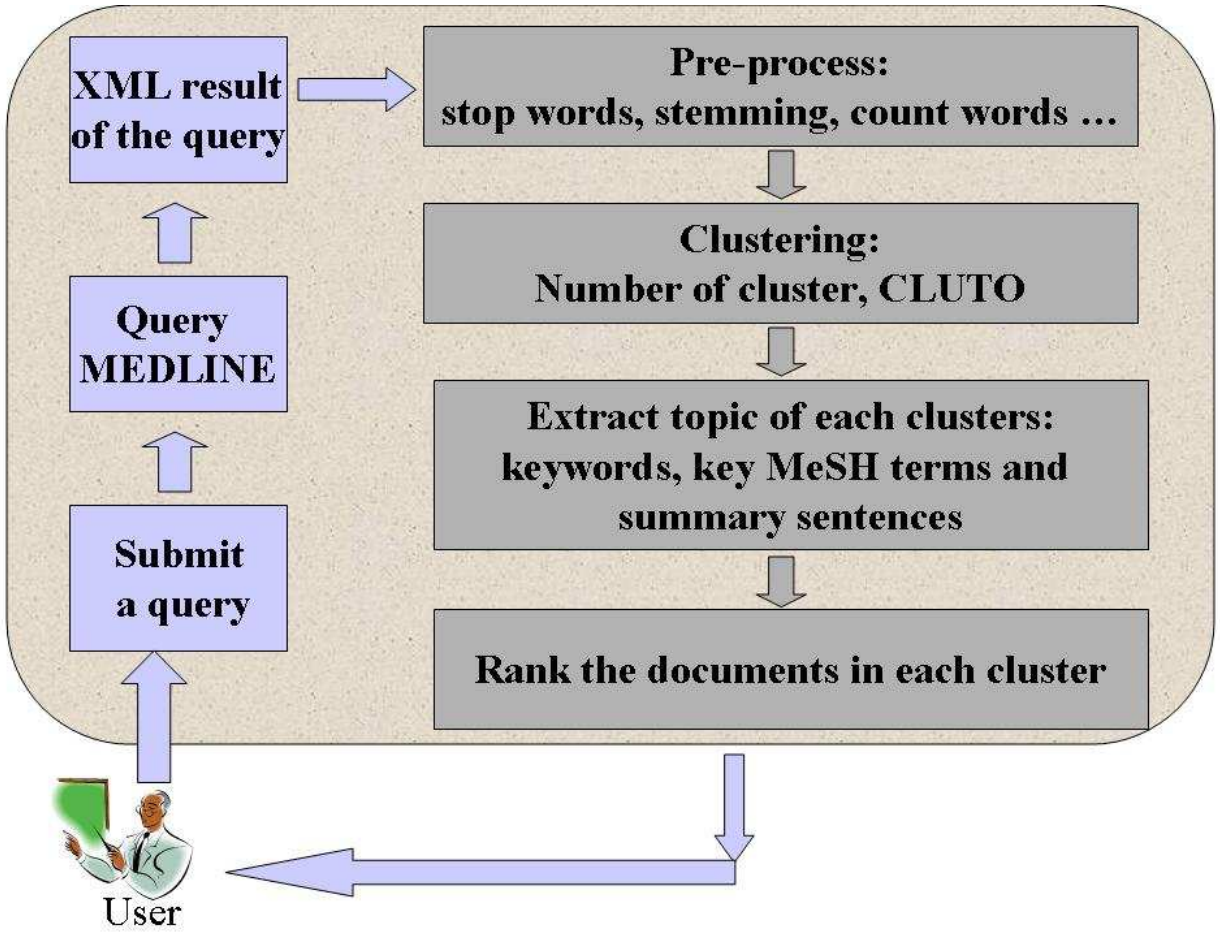
XML result of the query
Pre-process: stop words, stemming, count words …
Clustering: Number of cluster, CLUTO
Query MEDLINE
Extract topic of each clusters: keywords, key MeSH terms and summary sentences
Submit a query
Rank the documents in each cluster
User

Figure 1.

A

	d_1	d_2	$\cdots$	d_n
t_1	1	3	$\cdots$	0
t_2	1	0	$\cdots$	5
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
t_m	2	7	$\cdots$	0

B

$$M := \begin{pmatrix} .111 & .033 & \cdots & 0 \\ .111 & 0 & \cdots & .45 \\ \vdots & \vdots & \ddots & \vdots \\ .222 & .077 & \cdots & 0 \end{pmatrix}$$

C

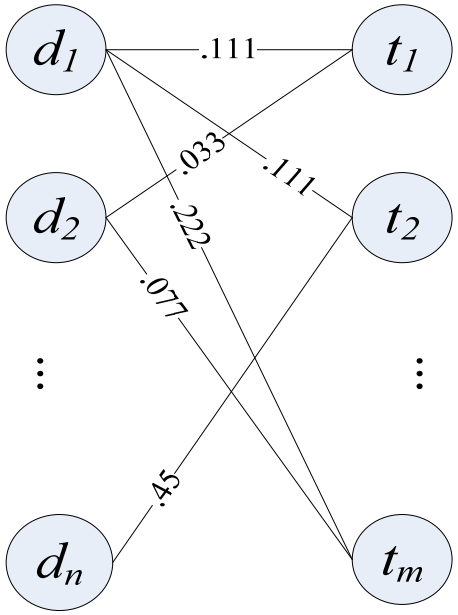

Figure 2.

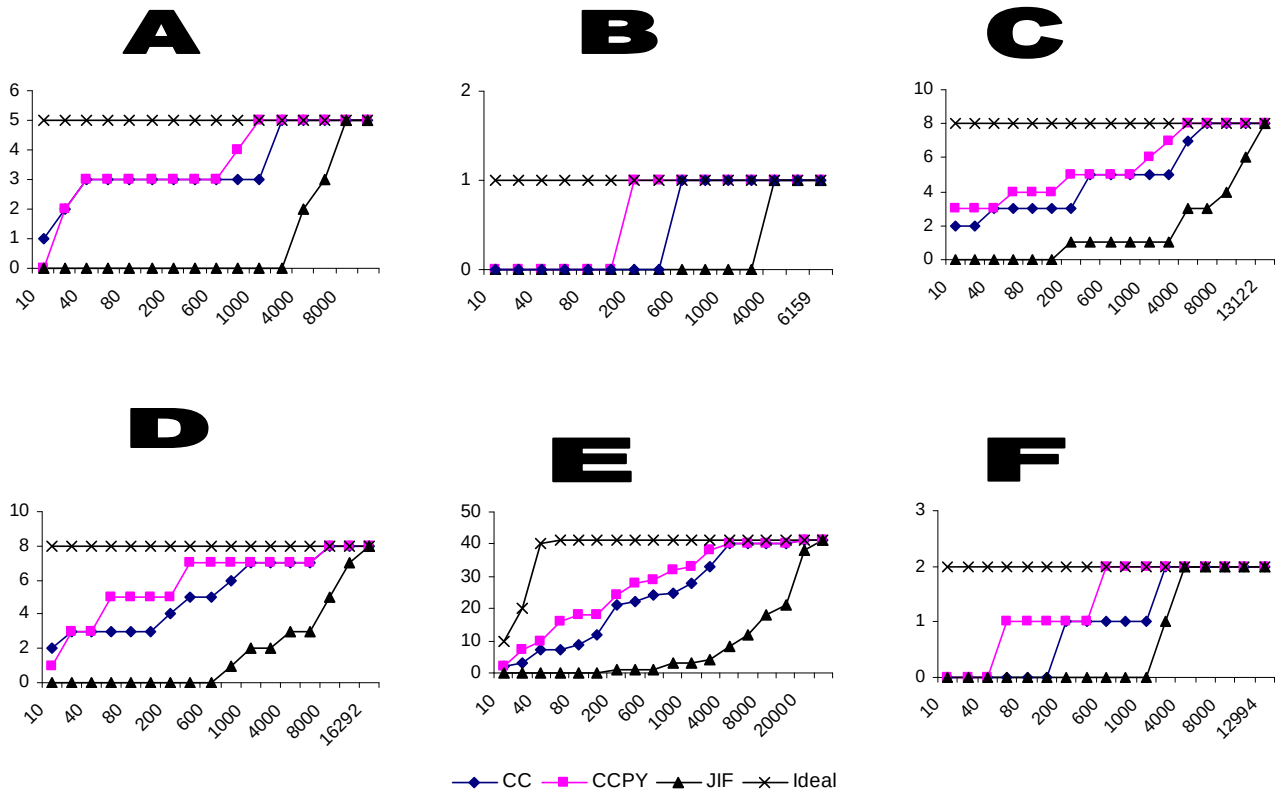

Figure 3.


PubMed Clustering and Ranking System - Mozilla Firefox
File Edit View Go Bookmarks Yahoo! Tools Help
http://localhost:3473/WebSite1/default.aspx
Go
Pubmed Clustering and Ranking System
Breast Cancer (77784)
"breast cancer"
Search
Cluster #1 (8212)
Keywords:
BRCA1, BRCA2, P53, p21
MeSH Terms:
"neoplasm proteins" "receptor, erbb-2" "tumor suppressor protein p53" "proto-oncogene proteins" "brca2 protein" "genes, brca1" "genes, p53"
Cluster #2 (6159)
Keywords:
paclitaxel
MeSH Terms:
"chemotherapy protocols" "fluorouracil" "doxorubicin" "paclitaxel" "epirubicin" "cisplatin" "vincristine" "chemotherapy, adjuvant"
Cluster #3 (13122)
Keywords:
estrogen, tamoxifen, hormone, antiestrogens,, progesterone, EGF, oestrogen
MeSH Terms:
"estradiol" "tamoxifen" "estrogens" "progesterone"
Cluster #4 (16292)
Keywords:
mammography, mammogram, dietary, American, African, Hispanic
MeSH Terms:
"mammography" "mass screening" "risk factors"
Cluster #5 (21005)
Keywords:
recurrence
Items 1 to 50 of 8212
Page 1 of 165
1. Regulation of BRCA1 and BRCA2 expression in human breast cancer cells by DNA-damaging agents.
2. Mutation analysis of BRCA1, TP53, and KRAS2 in ovarian and related pelvic tumors.
3. Promoter hypermethylation and BRCA1 inactivation in sporadic breast and ovarian tumors.
4. BRCA1 activation of the GADD45 promoter.
5. Developmental expression of Brca2 colocalizes with Brca1 and is associated with proliferation and differentiation in multiple tissues.
6. A population study of mutations and LOH at breast cancer gene loci in tumours from sister pairs: two recurrent mutations seem to account for all BRCA1/BRCA2 linked breast cancer in Iceland
7. Genetic heterogeneity in hereditary breast cancer: role of BRCA1 and BRCA2.
8. Pathology of familial breast cancer: differences between breast cancers in carriers of BRCA1 or BRCA2 mutations and sporadic cases.
9. Mutations of the BRCA1 and BRCA2 genes in patients with bilateral breast cancer.
10. Altered expression of BRCA1, BRCA2, and a newly identified BRCA2 exon 12 deletion variant in malignant human ovarian, prostate, and breast cancer cell lines.
Done

Figure 4