A System of User-Guided Biological Literature Search Engine

Meng Hu
EECS, Case Western Reserve University
Meng.Hu@case.edu

Jiong Yang
EECS, Case Western Reserve University
Jiong.Yang@case.edu

Abstract

Efficiently finding most relevant publications in large corpora is an important research topic in information retrieval. The number of biological literatures grows exponentially in various publication databases. The objective of the study in this paper is to fast locate useful publications from large biomedical document collections based on users' preferences.

In this paper, a new iterative search paradigm is introduced which integrates biological background knowledge in organizing the results returned by search engines, and utilizes user feedbacks to filter irrelevant documents. A term weighting scheme based on Gene Ontology is introduced to improve similarity measurement of documents in biomedical domain. A prototype text retrieval system has been built based on this iterative search approach. Experimental results show that the system can filter a large number of irrelevant documents while keep most of the relevant documents with limited user interactions.

1 Introduction

Text retrieval is an important problem in information retrieval. Searching for relevant publications from large literature corpora is a frequent job to biologists and biomedical researchers. With the abundance of biomedical publications available in digital libraries in recent years, efficient text retrieval becomes a more challenging task. For example, PubMed [1] now contains over 14 million publications. It is crucial to efficiently and accurately identify those documents most relevant to users' interests from such large document collections.

It has been recognized that one limiting factor of the traditional search engine technology is the low precision of the results returned. When users search by a few keywords, a large number of matched results could be returned. Users spend a significant amount of time to browse these results to find out those documents they are truly interested in. Keyword-based search is currently the most commonly employed search strategy in biomedical digital libraries. The publications returned by keyword searches may not be organized properly, forcing the users to browse thousands of publications. In most cases, it is impossible for users to manually read every returned entry, thus leads to loss of many truly relevant publications.

Many efforts have been done to improve the efficiency and effectiveness of literature retrieval in public domain and biomedical discipline. For example, document ranking is introduced for indexing entries in large literature collections. PageRank [2] and HITS [3] are both citation-scoring functions for evaluating the importance of documents. [4] presented a method to rank documents in MEDLINE using the differences in word content between MEDLINE entries related to a topic and the whole of MEDLINE. On the other hand, text

Copyright 2005 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.
Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

categorization has been studied to organize the search results. In [5], a machine learning model based on text categorization is built to identify high-quality articles in a specific area of internal medicine. SOPHIA [6] is an unsupervised distributional clustering technique for text retrieval in MEDLINE.

In this paper, a new iterative searching paradigm which aims to solve the above problem by incorporating biological background knowledge and user feedbacks is proposed. The iterative approach works as follows. First a set of documents returned by the keywords-based search is organized in a clustering manner, then users interact with system to provide objective evaluations on a small set of representative documents from these document clusters. Biological background knowledge descried in a controlled vocabulary is integrated to help the document clustering process. Next the system takes advantage of user feedbacks to refine the document set by filtering those user-rated irrelevant documents. Users can stop the iterative search at any time if the number of remaining documents is small enough for them to review, or the search process terminates automatically if a pre-defined number of remaining documents is reached. In this system, the number of documents that users examined is significantly reduced and the size of retrieved document set could also shrink with the help of the pruning process. This approach is particularly useful when labeling text is a labor-intensive job and when there is a large amount of results returned for a keywords-based search.

Since our text retrieval system focuses on the biological domain, we believe the background knowledge in this area could benefit the document clustering process, and add explanatory power to the organization of documents. The background knowledge we exploit in this paper is Gene Ontology [8]. Gene Ontology is a structured, controlled vocabulary that describes gene products in terms of their associated biological processes, cellular components, and molecular functions. We consider Gene Ontology as a hierarchical organization of biological concepts, and incorporate this hierarchical structure in measuring the similarity between biological publications. Users' evaluation on representative documents is utilized to prune the document set. Documents in clusters whose representatives are evaluated as relevant by users are kept for the next iteration.

Document clustering is one of the research areas most relevant to this paper. In [7] a core ontology WordNet is integrated in text clustering process as background knowledge. Concepts in the core ontology are compiled into the representations of text documents. However, their methods may not work for specific biomedical domain, and also the formal concept analysis used for conceptual clustering is known to be slow and impractical in real applications. Therefore, in this paper, a new term weighting scheme based on biomedical ontology is proposed to improve the similarity metric of biological publications.

The remainder of this paper is organized as the following. In Section 2, the terminology and metrics are formally defined and the methodology of our system is described in details. Experimental results are presented in Section 3. Last, we concluded our work in Section 4.

2 System and Methods

We have developed a prototype system to help users to retrieve useful biological literatures from a large amount of publications. The users will provide the keywords as input and interact with the system during the retrieval process. In this prototype system, Gene Ontology is utilized as the background knowledge to organize documents, and the user feedbacks are used to refine the retrieved documents. Finally, the system returns a small set of documents that are considered as most relevant to users' preference. In this section, we formally defined some terminologies. The methodology of our system is described and the three main steps in the system are explained in details.

2.1 Pre-Processing

In order to improve the response time of the system, pre-processing is done before users interacting with the system. Every time a document is imported to the database, the pre-processing described below is conducted.

During pre-processing phase, Gene Ontology, which is originally described in a DAG (directed acyclic graph), is transformed to a tree hierarchy. If a term has multiple parents, it will have multiple instances in the transformed GO tree because it has different paths to the root term, which is important for the feature weighting discussed in a later section. For example, term "RNA transport"(GO:0050658) has two parent terms: "nucleic acid transport"(GO:0050657) and "establishment of RNA localization"(GO:0051236). Therefore, "RNA transport" has two instances in the transformed GO tree: one is at level 8 as a child of "nucleic acid transport", and the other one is also at level 6 as a child of "establishment of RNA localization".

After the transformation of the Gene Ontology structure, the occurrences of GO terms are collected from the documents. The synonyms of GO terms defined in Gene Ontology are also considered equally as GO terms themselves. That is to say, if a synonym of a GO term appears in a document, the GO term is also considered occurred in the document. For instance, when searching for "peroxisome targeting sequence binding", "PTC binding" is also searched. By searching all documents, the number of occurrence of each GO term in each document is collected. Other statistical information are also collected at the same time, such as the length of every document, occurrence of every other word in each document, etc. Non-informative words, such as "the", "we", are removed from the documents based on a given English stop-word list.

2.2 Feature Selection and Weighting

Traditionally documents are considered as a bag of words, and are represented by a set of feature words. Feature selection is the process to select the set of words to represent documents. It benefits the clustering and classification by reducing the feature space and eliminating noisy features. In our system, the mutual information as defined in [9] is used as the criteria for feature selection. 2000 words with the most mutual information throughout all the documents in each iteration are selected as the feature terms. For example, if in the first iteration, the system returns 5000 documents matching users' keywords out of 100,000 documents, 2000 words with the most mutual information in these 5000 documents will be selected as feature words. Besides this, a set of GO terms is also chosen as feature terms. A feature level is selected in the transformed GO tree, and all distinct GO terms at this level serve as the feature terms.

The 2000 words with most mutual information and all the GO terms at the feature level in GO tree form the feature set. In our prototype system, level 8 in Gene Ontology, which contains around 3500 GO terms, is selected as the feature level.

After obtaining the feature words to represent documents, we construct a vector of real numbers for every document by assigning each feature term a numerical weight. The weight of a term is dependent on two factors: the importance of the term throughout all the documents and the strength of the term in a particular document. Therefore, the weight of term t consists of two parts: the global weight and the local weight. The global weight(gw) of a term t is defined as $\frac{|D|}{df(t)}$, where $|D|$ is the total number of documents in database, and $df(t)$ is the number of documents that contain term t.

A definition of the local weight of a term t in a document d based on Poisson distribution ([10]) is given as below:

$$lw(t) = 1/(1 + exp(\alpha \times dlen) \times \gamma^{f(t,d)-1}) \quad (1)$$

where $\alpha = 0.0044$, $\gamma = 0.7$, $dlen$ is the length of document d, and $f(t, d)$ is the frequency of term t in d.

For those feature terms obtained by the most mutual information, their weights in a document are just the multiplication of the global weight and the local weight: $tw(t) = gw(t) \times lw(t)$. A more complex weighting scheme is used for those feature terms from Gene Ontology. The original term weight computed from the above will be distributed and aggregated based on Gene Ontology structure. The weight of a term not at the feature level is distributed or aggregated to its ancestor or descendant terms at the feature level. If the term is at a lower level than the feature level, its weight is aggregated to all ancestors of this term in the feature level. If the term

is in a higher level than the feature level, its weight is uniformly distributed to its children level by level until the feature level is reached. After obtaining the term weight vector for each document, the similarity between two documents is defined as the cosine similarity of their term weight vectors.

Figure1 illustrates an example of the distribution and aggregation process. A part of the Gene Ontology hierarchy is shown in Figure1. The two numbers beside each term at the feature level are the original weights computed for a document and the final weights after distribution and aggregation, respectively. If the second level in this figure is selected as the feature level, then only ”Transport”, ”Secretion” and ”Establishment of RNA localization” will serve as the feature terms when computing the document similarity. In this case, although the term ”Establishment of RNA localization” never appears in the document, the weights of its children terms will be aggregated to the second level. Therefore, term ”Establishment of RNA localization” will gain weight of 0.25 from its children terms ”RNA Transport” and ”establishment of pole plasm mRNA localization”. However, the weights of ”Amide Transport”, ”Ion Transport” and ”Boron Transport” are not aggregated to the second level, because their ”Transport” is a substring of its children terms, and the occurrences of ”Transport” has already been counted. Meanwhile, the weight of term ”establishment of localization”, which locates in the first level, is distributed uniformly to its children terms. Therefore, the final weight of feature terms ”Transport”, ”Secretion” and ”Establishment of RNA localization” in this document will be 0.76, 0.16 and 0.33 respectively.

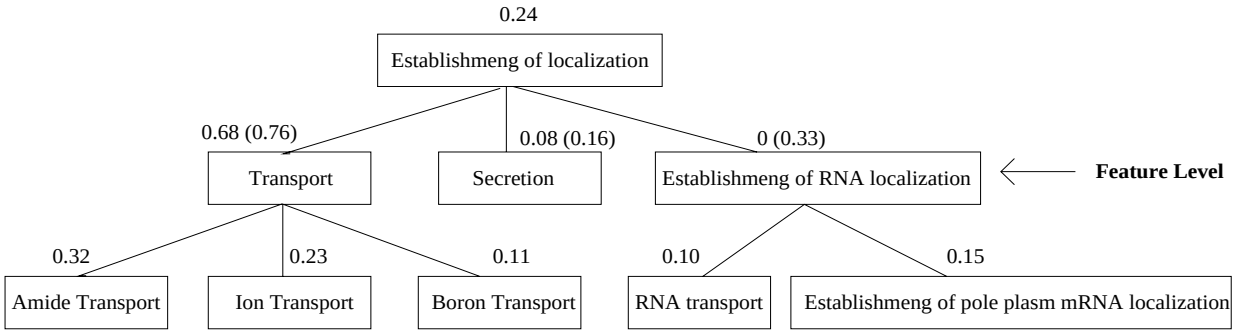

Figure 1: Distribution and Aggregation of term weights

2.3 Clustering and Representative Selection

Document clustering has been considered as an important tool for browsing and navigating large document collections. In our prototype system, after users input the keywords to search, a set of documents is returned from the document corpus by exact keyword matching. To organize these documents in a meaningful way, these documents are clustered into groups according to their mutual similarities. Traditional document clustering methods only consider the distribution of words in documents, but ignore the fact that prior knowledge could be important in organizing the documents. In stead of measuring the document similarity directly by the distribution of words, our idea is to compile the background knowledge provided by biological lexicon into the similarity measurement, which is described in the earlier section.

In our system, Bi-Section-KMeans clustering method ([7]) is used for clustering purpose, which has been shown to perform as good as other clustering algorithm, but much faster than others in document clustering. Bi-Section-KMeans is essentially a variant of KMeans clustering algorithm, which keeps partitioning the largest cluster until the desired number of clusters is reached.

After obtaining the document clusters, one representative document is selected from each cluster. In our prototype system, the centroid document of each cluster, which has the maximum average similarity to all other documents in the cluster, is chosen as the representative document. The user will review all the representative documents and rate each one as relevant or non-relevant. In each iteration, documents are clustered and repre-

sentative documents are selected. The number of clusters is a parameter of the system and can be set by users. Users will read the representatives and provide their evaluations. The system will use their evaluations to refine the document set, then reduce the number of documents. Documents in those clusters whose representative documents are rated as "relevant" by users are then kept for next iteration. By looking at a small number of documents in each iteration, users save a significant amount of time from manually reading all search results.

3 Experimental Results

A prototype search system is implemented in Perl based on the methodology proposed in this paper. 100,000 abstracts from PubMed, which are stored as plain text files in a 7200 rpm hard drive, are used to test our prototype system. These abstracts serve as the document universe in our experiments. In this section, experimental results are presented to demonstrate the effectiveness and efficiency of our proposed method.

The following experimental method is conducted for evaluating the prototype system. First we search a set of keywords, referred to as reference keywords, by exact keywords matching, then a set of documents are returned for this search query. This set of documents are considered as the benchmark and serve as the reference result set. Then some keywords are removed from the reference keywords to generate a reduced keyword set. Naturally, the reduced keyword set will result in a larger document set, which is referred to as initial document set. The system organizes these documents by document clusters, and users will review the representatives selected from these document clusters in each iteration. Finally the system will return a set of documents after several iterations. In our experiments, recall is used to evaluate the search performance of our prototype system. We denote the set of documents obtained by searching reference keywords as D_r and the set of documents our prototype system returns by taking the reduced keywords as input is denoted as D_o. The recall is defined as $\frac{|D_o \cap D_r|}{|D_r|}$.

One reference keywords set used is "metabolism", "expression", "regulation", "phenotype", "protein", "mRNA" and "yeast". By doing an exact keyword matching on this set of keywords, 300 documents are returned from our testing document universe. Then we use the following three reduced keyword sets: "regulation, Phenotype and yeast", "metabolism, expression, regulation, Phenotype, protein and mRNA" and "regulation, mRNA and yeast" as the input keyword sets of our system. Each of the three reduced keyword sets will result in thousands of documents by exact keywords matching. In this experiment, the system was set to terminate when the number of remaining documents reaches half of the initial result document set. The number of document clusters was set to 10 in each iteration.

The results show that our prototype system can identify over 70% of the benchmark documents while removing thousands of irrelevant documents in several iterations. Since we reduced the size of the result document set to half, but achieved a recall over 50%, the precision of the results was also improved compared to the initial results returned by exact matching on the reduced keywords. Similar results were obtained by other keyword sets such as "protein, kinase, enzyme, synthetase, DNA and ligase" and "nucleotide binding, promoter, enzyme, expression and regulator". To evaluate the robustness of our system against different input size, we also chose keyword sets to vary the size of initial document set, which is returned by exact keyword-matching on the reduced keyword set. The experimental results show that the recall varied insignificantly around 68%, although the response time rose with the increase of initial document size.

One parameter of our system is the number of document clusters in each iteration. We tested the performance of our system under different settings of this parameter. The results show that if the cluster number is not set too small, the system performed almost steadily, and was able to identify 70% of the reference document set. The reason for this observation is that a partitioning clustering algorithm is used in our system, and in each iteration of the clustering process, it only splits the larger cluster. When the size of the larger cluster is not too large, users tend to have the same evaluations on two clusters split from one larger cluster. Therefore the system performs robustly when the number of clusters is not set too small. However, the number of clusters can not be

Table 1: Performance on keyword set

	Iterations	Response Time	Recall
Test Set 1	4	600 s	74%
Test Set 2	5	645 s	69%
Test Set 3	4	570 s	70%

set too large in practice, because this parameter is actually the number of representatives users will review in each iteration. A reasonable setting of the number of clusters is from 5 to 15.

4 Conclusions

In this paper, a new iterative search paradigm is proposed. In our approach, document clustering is adopted to organize documents, and the user feedbacks are used to refine the retrieved documents. A new term weighting scheme is defined based on Gene Ontology, which benefits the document clustering by considering the hierarchy of biological concepts in the document similarity measurement. By this approach, users review a much smaller number of representative documents and the system filters a large number of irrelevant documents according to user feedbacks. A prototype biomedical literature search system has been built upon this iterative search paradigm. Experimental results demonstrate the effectiveness, efficiency and robustness of our prototype system.

References

[1] PubMed, available at http://www.ncbi.nlm.nih.gov/entrez/

[2] Brin S. and Page L., The Anatomy of A Large-scale Hypertextual Web Search Engine, *WWW7 Conf., 1998*

[3] Kleinberg J., Authoritative Sources in A Hyperlinked Environment, *9th ACM-SIAM Symposium on Discrete Algorithms, 1998*

[4] Suomela BP and Andrade MA., Ranking the Whole MEDLINE Database According to A Large Training Set Using Text Indexing, *BMC Bioinformatics. Mar 2005*

[5] Aphinyanaphongs Y., Tsamardinos I., Statnikov A., Hardin D., and Aliferis CF, Text Categorization Models for High Quality Article Retrieval in Internal Medicine, *J Am Med Inform Assoc. 2005;12*

[6] Vladimir D., David P., Mykola G., and Niall R., SOPHIA: An Interactive Cluster-Based Retrieval System for the OHSUMED Collection, *IEEE Transctions on Information Technology in Biomedicine, June 2005*

[7] Andreas H., Steffen S., and Gerd S., Text Clustering Based on Background Knowledge, *Technical Report*

[8] The Gene Ontology Consortium, available at http://www.geneontology.org.

[9] Slonim N. and Tishby N., Document Clustering Using Word Clusters via The Information Bottleneck Method, *ACM SIGIR 2000*

[10] Kim W., Aronson AR and Wilbur WJ., Automatic MeSH Term Assignment and Quality Assessment. *Proc AMIA Symp 2001*