Improved biomedical document retrieval system with PubMed term statistics and expansions

Huian Li

University Information Technology Services,
Indiana University,
Indianapolis, IN 46202, USA
E-mail: huili@iupui.edu

Jake Yue Chen*

Department of Computer and Information Science
Purdue University School of Science,
Indiana University School of Informatics,
Indianapolis, IN 46202, USA
E-mail: jakechen@iupui.edu
*Corresponding author

Abstract: Large biomedical abstract databases such as MEDLINE enable users to search for large bodies of biomedical knowledge quickly. In this study, we describe a new framework to improve the performance of MEDLINE document retrieval. We first analysed and built a normalized term frequency distributions for 1.8 million terms by sampling from 1,500,000 MEDLINE abstracts. Then, we developed a statistical model to identify significantly observed terms ('gists') in a document as additional document keywords to help improve document retrieval precisions. To improve document recalls, we integrated several biological ontologies that can expand user queries with semantically compatible terms. The framework was implemented in Oracle 10*g*.

Keywords: term statistics; ontology expansion; information retrieval; gists.

Reference to this paper should be made as follows: Li, H. and Chen, J.Y. (2009) 'Improved biomedical document retrieval system with PubMed term statistics and expansions', *Int. J. Computational Intelligence in Bioinformatics and Systems Biology*, Vol. 1, No. 1, pp.74–85.

Biographical notes: Huian Li is a staff member at the high-performance computing group of the Research and Academic Technology Services Division at Indiana University. He holds a MS Degree in Bioinformatics from Indiana University School of Informatics in 2003. He has conducted bio-computing and bioinformatics development work since then.

Jake Y. Chen is an Assistant Professor of Informatics and Computer Science, at Indiana University – Purdue University Indianapolis. He is the founding Director of Indiana Center for Systems Biology and Personalised Medicine and an IEEE Senior Member. He holds both MS and PhD Degrees in Computer Science from the University of Minnesota. His research interests include bio-computing, networks and systems biology, and translational bioinformatics.

1 Introduction

Biomedical research publications contain a large amount of textual information. For example, the MEDLINE database (http://www.ncbi.nlm.nih.gov/entrez/) currently contains 15 million abstracts, or more than 1.5 billion English terms, from various types of biomedical papers published in the past several decades. These abstracts represent current accumulated knowledge, which is not captured elsewhere in relational databases. Many researchers routinely search these abstracts in order to review the scientific progress of their fields and develop new perspectives. However, even with PubMed (http://www.ncbi.nlm.nih.gov/entrez/; Schuler et al., 1996), a popular web-enabled software system based on MEDLINE, it is still primarily a trial-and-error process to search and retrieve relevant scientific literature. This is because large numbers of biomedical texts covering similar research topics often obscure the PubMed document retrieval and ranking engine, making it challenging for biologists to control how to define good search results.

There are two conventional models to perform biomedical document retrievals. The first model, *full-text based retrieval*, uses all the terms from all MEDLINE abstracts to build term indices (Salton and McGill, 1983). When a user query matches any indexed term, all documents containing the indexed term will be retrieved, sometimes rank-ordered by a term occurrence score. The second model, *keyword-based content retrieval*, quite popular in data mining (Srinivasan, 2004; Rindflesch et al., 1999), uses a manually created list of keywords or MeSH terms as indices.

To measure the effectiveness of document retrieval method, *recall* and *precision* measures are often used. *Recall* measures the percentage ratio of the number of relevant records retrieved to the total number of relevant records in the database. *Precision* measures the percentage ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. It is not difficult to observe that a full-text based document retrieval method has high recall (all text with a particular term appearance will be returned) but low precision (many retrieved text may be unrelated to the query term). A keyword based document retrieval method improves precision by retrieving only documents containing keywords; however, the use of keywords also lowers recall, because manually created keywords are usually far from enough to cover all key concepts of an paper.

Several techniques have been developed to improve both recall and precision for biomedical document retrievals. Automated expansion of keyword list in keyword-based document retrieval systems has been explored to improve precision (Magnini and Prevete, 2000). However, unless the keyword list can be generated automatically and systematically from biomedical ontological systems, the level of precision for biomedical document retrievals will remain low. In full-text based document retrieval systems, adding term weight is a common method to improve overall precision and have been adopted by, for example, PubMed (Wilbur and Yang, 1996). Many other systems have also adopted methods that use inverse document frequency to calculate a term's weight score from the following formula (Salton and Young, 1973):

$$tf \times \log(N / df) \tag{1}$$

where tf is the number of times a term T appears in a query document, N is the total number of documents from a document collection, and df is the total number

of documents with term *T* in them from the document collection. The calculated term weight score is generally high for rare terms such as 'bcl2', and low for common terms such as 'patient'.

Although using term weight scoring methods can improve precision for full-text based document retrieval, these calculated scores can be arbitrary and may not reflect the true significance of their occurrence in a document. For example, the common term 'patient' may appear more than once per document in a document collection related to Alzheimer disease clinical trials. Using the simple term frequency method, one may conclude that the term 'patient' is more *significant* than the term for a novel Alzheimer disease drug, which may appear less than once per document. Using term inverse document frequency score may also fail, if the document collection is not sufficiently large, e.g., the term 'patient' may occur several times in only a few documents out of a biased sample of documents, inflating its weight score. Therefore, an advanced method that considers true statistical significance of terms based on sampling large biomedical texts need to be developed in order to improve precision for content-based biomedical document retrievals.

In this work, we developed methods and a system to improve precision and recall of biomedical document retrievals. The paper is organised as follows. First, we collected frequency distributions of 1.8 million biomedical terms by sampling each term's occurrence from 1.5 million MEDLINE abstracts. Second, we used these term statistics to derive a *p*-value for each term that occurred in a given biomedical document. Only significantly occurring terms that satisfy a certain threshold are retained and put into a document 'gist', an automatically generated list of keywords. Third, we addressed the problem of recall by allowing query term expansions, using three separately integrated data sources: gene ontology (Wheeler et al., 2000), cancer thesaurus (http://nciterms.nci.nih.gov/), and OMIM (http://www.ncbi.nlm.nih.gov/omim/). Fourth, we incorporated the above methods into a web-based software system (http://discover .uits.indiana.edu:8340/cgi-bin/TextMining/search_p.pl) built on the Oracle 10*g* text mining platform (http://www.oracle.com/technology/ industries/life_sciences/index. html). Fifth and lastly, we discuss the significance of our work.

2 Term frequency distributions and statistical significance

Essential to improving precision is a new method that requires the collection of each biomedical term's statistical distribution found in large biomedical document samples. To accomplish this, we fetched 396 compressed MEDLINE files, medline03n0001.xml.gz to medline03n0396 gz, from the NCBI website. Each file contains 30,000 abstracts in XML format. For this work, we used about 1.5 million MEDLINE abstracts from 50 randomly selected files as the document *population*. This corresponds to approximately 10% representative subset of the total 15 million abstracts found in MEDLINE.

We used the following technique to perform random document sampling. In each sampling process, we randomly select 10,000 MEDLINE abstracts from the document population. To guarantee 'randomness', we used two C program functions, srand48() and drand48(): the former initialises the seed, and the latter returns non-negative double-precision values uniformly distributed in the interval [0, 1). We seed the random number generator function with the Unix function time(). During each

sampling process, every term that occurs in the 10,000 MEDLINE abstracts is extracted and its occurrence is then counted. Title words and MeSH terms are also considered as part of an abstract and counted. The result is written to a file including two columns, i.e., term symbol and its raw counted frequency. The total number of terms for each sample is also recorded. Normalised term frequency is calculated as raw counted term frequency divided by the total number of terms in one random sample.

Figure 1 shows the distribution of summed *sample total term frequency* binned from 1,000 random samples. The *x*-axis represents the sample total term frequency range bins and the *y*-axis represents the count of samples in a given frequency range bin. The nice bell curve (near-normal distribution) is evidence that the sampling process is unbiased, because *sample total term frequency* demonstrates the statistical characteristics of a random variable. We also determined the total unique terms extracted from the 1,000 random samples to be 1,819,228 (not shown in the figure). Each biomedical term's occurrence frequency is different, ranging from millions to a few dozen. For example, In Table 1, we listed several common terms with the highest term frequencies. These terms usually appear many times in almost every abstract. Apparently, they are the worst candidates to be selected as document keywords.

Figure 1 A histogram showing distribution of summed sample total term frequency binned from 1,000 random samples (see online version for colours)

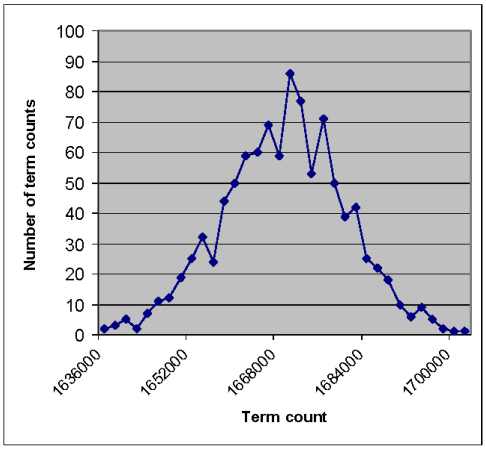


Table 1 Common terms with highest term frequency (The frequency shown is summed over all 1,000 samples)

Term	*Frequency*
Of	72454545
The	71317305
And	50872006
In	41774072
To	26561892
With	18202942
Was	13970537
For	13446468

Table 1 Common terms with highest term frequency (The frequency shown is summed over all 1,000 samples) (continued)

Term	*Frequency*
The	12933436
Were	12008644
That	10798401
By	10750327

We further show the term frequency distributions of all the extracted 1,819,221 terms in Figure 2. Here, we use the *x*-axis to represent the term frequency range bins (summed over 1,000 samples, shown in log scale) and the y-axis to represent count of terms in a given frequency range bin. We can divide the distribution curve into three portions: the left portion (frequency < 10), the middle portion (frequency between 10 and 100,000), and the right portion (frequency > 100,000). The right portion represents *abundant terms* with very high frequency (some examples were shown in Table 1). The middle portion follows a nice power law distribution, and represents *regular terms*. The left portion is somewhat interesting: their distribution is quite different from the rest of the regular terms. The total number of unique terms in this range is huge (at 844,567), or approximately 46% of all the words. By examining these terms in detail (Table 2), we found that they are all belong to a class of '*peculiar terms*', which are misspelled (e.g., 'B!nding' instead of 'Binding'), extremely rare (e.g., '$1/day'), or synthesised by the author (e.g., 'c-Series'). Additional examples of abundant terms and regular terms are shown in Table 2. The general trend is that as the term frequency increases, particularly moving from peculiar term frequency range to the regular term frequency range, terms become more comprehensible and general in biomedical documents.

Figure 2 Distribution of frequencies of all terms in 1,000 samplings (see online version for colours)

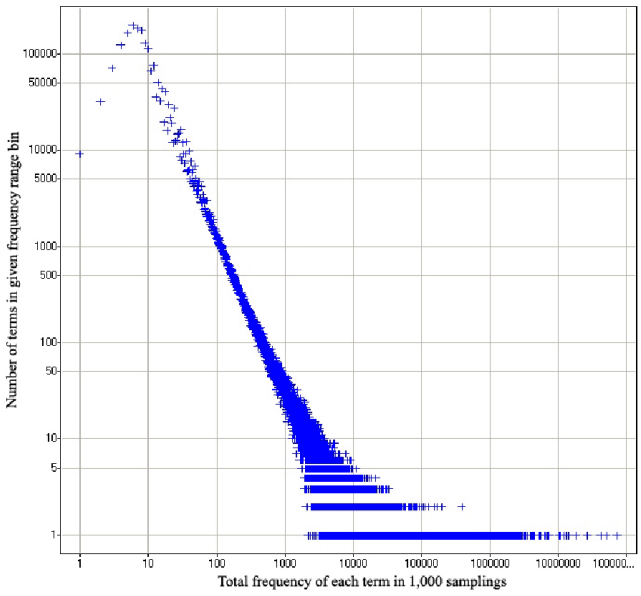
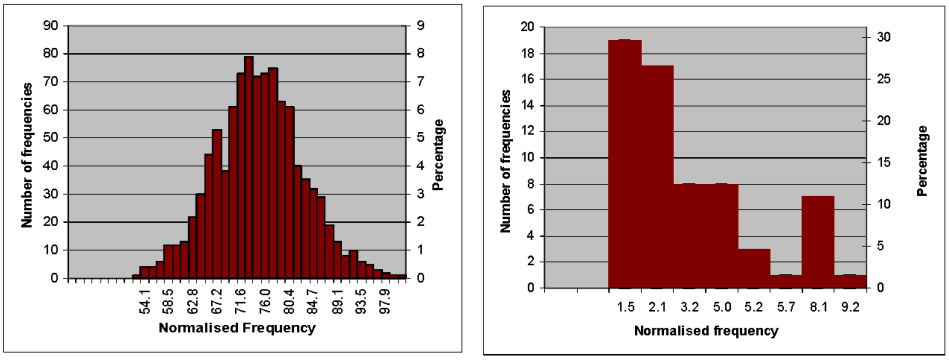


Table 2 Example terms in different frequency ranges

[1, 10)	[10, 100)	[100, 1K)	[1K, 10K)	[10K, 100K)
$1/day	$905/patient	0.5-fold	1-hour	10-fold
1.5-4-kHz	2”-modified	1-aminobenzotriazole	APS	ADP-ribose
2ndly	B-carotene	2nd-order	Generation	Body
B??hl	c-ANCA-positive	Mucosectomy	Alphabeta	Acceptor
B!nding	Decency	Normoalbuminuria	alphavbeta3	Elegans
c-Series	Earring	Oncocytomas	Burnetii	Hydration
early-growth	failure-rate	Propoxy	butyrylcholinesterase	Mammary
steine	gamelike	radio-therapy	Calreticulin	Percutaneous
p-hexyl	Kings	sham-operation	Camptothecin	Sputum
mulitpotent	z-test	Sgk	Campus	Zidovudine

We further investigated the term distributions between *low-abundance regular terms* (G_0) and *normal-abundance regular terms* (G_1). In Figure 3(a) and 3(b), we showed two examples of *normalised* term frequency distribution, 3(a) for the term ‘bind’ of G_1 type and 3(b) for the term ‘sgk’ of G_0 type. The normalised term frequency is calculated as term frequency over total terms in any given sample in a *parts-per-million* (*ppm*) scale. The *x*-axis represents normalised term frequency range bins and the *y*-axis represents the count of all samples which contains the given term in a specific frequency bin.

Figure 3 Normalised frequency distribution of the term (a) ‘binds’ and (b) ‘sgk’



(a) (b)

For the term ‘bind’, the term frequency distribution is near-normal (Figure 3(a)). We can read that its mean $\mu_{\text{bind}} = 74.64$ and its standard deviation $\sigma_{\text{bind}} = 8.08$. This near-normal distribution is typical for terms of the G_1 type. For the term ‘sgk’ of type G_0, however, the distribution is different. Because the term is rarely used (the total frequency is 64, still significantly larger than 10 expected of peculiar terms discussed early), its term frequency is quite small and sometimes zero in many document samples. Figure 3(b) shows the frequency distribution of the term ‘sgk’, which is not a ‘normal distribution’.

We store the entire term frequency distribution for each term in a relational database schema. Except for peculiar terms, we classify biomedical terms (both abundant terms and regular terms) into G_0 and G_1 types according to the term's total frequency. If the term belongs to the G_1 type, we note in our database that a normal distribution function using mean and standard deviation is available; otherwise, we note in the database that an empirical cumulative distribution function should be used instead.

The calculation of term significance using *p*-value is done at the document querying stage as follows. When user issues a query search term, K_0, K_1, …, K_s, we retrieve these query terms against the MEDLINE documents stored in our text management relational database (Oracle 10*g*). The document collection, D_0, D_1, …, D_n, found in the database will be forwarded to the term significance evaluator system module. For each document D_i in the collection, we find and record each term T_0, T_1, …, T_m, and their counted term occurrence in the given document in $TF_{i,j}$ (where $i = 0, 1, \ldots, m,$ and $j = 0, 1, \ldots, n$). DF_i denotes the total number of terms occurred in the document D_i. Note that the MeSH terms and the title are counted as part of the document. We calculate the *p*-value for term T_i of the type G_1 using:

$$P_{i,j} = \text{NormDist}\left(\frac{TF_{i,j}}{DF_j}, \mu_i, \sigma_i\right), \quad i = 0, 1, \ldots, m \text{ and } j = 0, 1, \ldots, n. \tag{2}$$

If the term T_i is of type G_0, we calculate its *p*-value using:

$$P_{i,j} = \text{EmprDist}\left(\frac{TF_{i,j}}{DF_j}\right) \quad i = 0, 1, \ldots, m, \text{ and } j = 0, 1, \ldots, n \tag{3}$$

To implement the above algorithm, we downloaded a Perl module from CPAN (http://search.cpan.org/dist/Statistics-Distributions/), which is used to calculate statistical distributions. We report each p-value using natural logarithm–a standard practice in bioinformatics software–and set the lower-bound of *p*-value to be –999. Therefore, all term *p*-values would range between 0 (least significant) and – 999 (most significant).

To illustrate how term occurrence *p*-value can improve precision of biomedical document retrieval, we first tested an arbitrary PubMed paper titled 'Progress in cancer gene therapy' (PubMed ID = 10522756). In Table 3, we listed most significant terms in the document, in descending orders of term *p*-value. By only reading this list of automatically generated keywords (which we call 'gist'), we can tell the paper is related to 'gene' and, 'imminogenetics' particularly. It talks about 'progress', 'drug resistance' of 'oncogenes' and 'tumour suppressor genes'. Then, we performed a query using our text mining software (See Section 4) with the query term 'oncogenes progress' (parameters: theme option on, expansion off). All top ranked documents appear related to advances of different tumour genes, and the document just analysed is ranked 4th.

Table 3 Terms of significance in the abstract 'Progress in cancer gene therapy'

Gene	–999.000
Immunogenetics	–999.000
synergies	–999.000
Oncogenes	–492.515
Progress	–347.353
Naked	–324.959
Strategies	–306.843
Drug Resistance, Neoplasm	–250.421
Genes, Tumour Suppressor	–236.575
Drug Resistance, Multiple	–218.411
Vectors	–210.553

3 Thesaurus-based query term expansions

We further investigate how to improve document recall, particularly when user interests vary among different domains. For example, a bioinformatician typing a query 'Alzheimer' may be interested in retrieving literature that discusses Alzheimer-related genes, whereas a physician typing the same query may be interested in new patient treatments. In this case, using gene ontology and clinical ontology thesaurus to expand the term 'Alzheimer' may be appropriate.

To build thesaurus, we used several public biological data integrated from different sources. The first data source is the NCI cancer thesaurus (http://nciterms.nci.nih.gov/) obtained from National Cancer Institute. It contains the working cancer-related vocabulary, covering clinical, translational and basic research as well as administrative terminology. There are 35,299 terms and 40,748 relationships. These terms connected by term relationships form a Directed Acyclic Graph (DAG) structure, which we provided software to visualise (see next section).

A second data source that we integrated as thesaurus is Gene Ontology (GO) (Wheeler et al., 2000). The GO project provides consistent descriptions of gene products in different databases using controlled terminologies. It provides a standard vocabulary and a set of relationships among approved vocabulary terms. There are 18,385 GO terms and 26,249 GO relationships for the GO version that we used. Similar to the NCI thesaurus, GO terms and GO relationships also have a DAG structure.

The third data source that we integrated is Online Mendelian Inheritance in Man (http://www.ncbi.nlm.nih.gov/omim/) It is a curated database of human genes and genetic disorders, which concerns mainly on inherited, or heritable, genetic diseases (http://www.ncbi.nlm.nih.gov/omim/). There are 9,213 OMIM records in the OMIM version of the database. OMIM records are crosslinked, but do not have tree-like relationship.

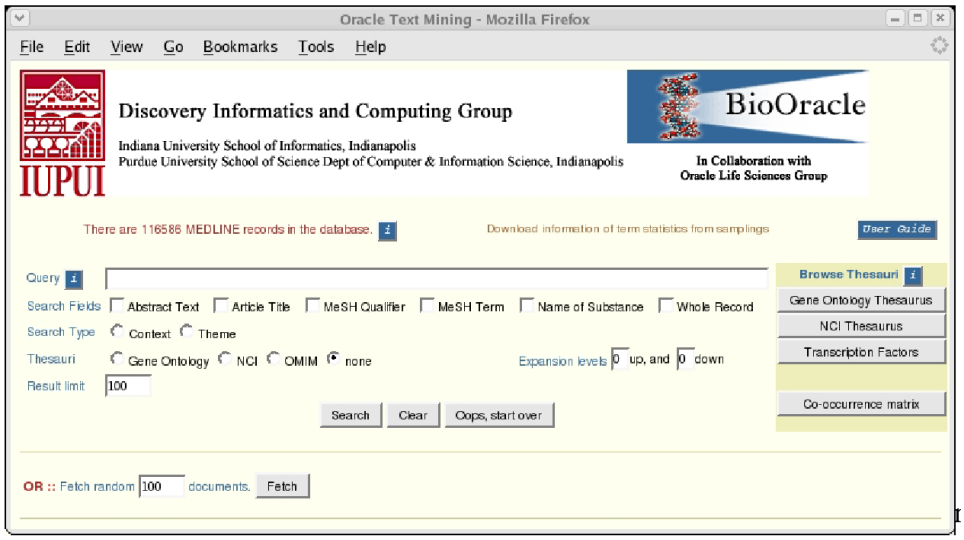
To perform expanded keyword, the user can select a particular thesaurus to expand query term using, for example, the OMIM thesaurus. Prior to submitting the query term to the text search engine, the query engine will add the expanded gene symbols found to

be related to the query term such as 'Alzheimer'. The rest of search will therefore be conducted based on all the initial and expanded query terms.

4 Software system implementation

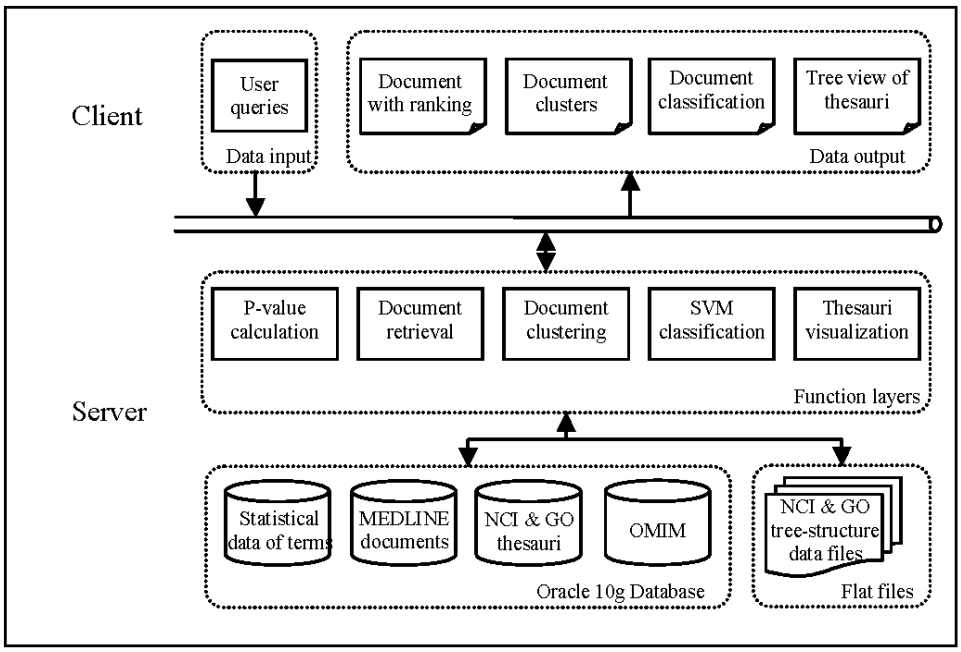
We developed a Web-based biomedical document management software system. This system currently contains a random collection of 116,586 indexed biomedical paper abstracts, which we imported from MEDLINE. The software system is built with several new features of Oracle 10*g* database server, including build-in clustering algorithms, thesauri capability, and text mining cartridge. A snapshot of web user interface of this system is shown as Figure 4. Users can interact with the system by entering query terms in the search box, and choose different search parameters. Users can also click the '*i*' button wherever available to obtain additional information about the search features. Furthermore, users can click on a thesaurus (either GO or NCI) and browse the relationship of terms in a tree-like visualisation environment.

Figure 4 A snapshot of the web user interface to the biomedical document management system (se online version for colours)



In Figure 5, we show a diagram of the system architecture. At the server side reside all the data management and functional modules. The Oracle 10*g* database provides document data management support and support the operations of other server-side functional modules. A few flat text files outside of the database are exported from the database server to support visualisation of document trees. SVM classification, SVM categorisation, document clustering, and cluster visualisations are bundled software system features supporting biomedical document management. At the client side, user queries can be input into web browser input boxes, and query results from the server can be displayed directly in users' web browser.

Figure 5 System architecture



We process document data into our systems using the following procedure. First, we download collections of MEDLINE documents in the XML format. Second, we load all the abstracts into the Oracle 10g database. Third, we use Oracle text mining utilities to extract abstracts, tokenise terms in the abstracts, and index them for each document. Fourth, we apply our method to derive a *p*-value for each index word in a MEDLINE document, using the term's frequency distribution from our experiment described in the previous sections. Fifth, we expand terms using thesaurus data when available to improve recall of documents, whose topics are semantically related to users' query terms. After these preparations, the documents stored in the database are immediately available for retrieval, classification, and clustering.

We provide in our system a link to one gzipped text file which includes the information of all regular terms collected from 1,000 samplings. The file has two columns. The first column lists each individual term, while the second gives the average frequency (in the format of ppm) of the corresponding term. We hope that the information can be beneficial to other researchers and provide some insight to biomedical literature.

5 Discussion

In this work, we developed novel methods that can improve precision and recall of content-based biomedical text retrieval and mining. Our first method, used to improve precision, relied on our development of a comprehensive database of biomedical term statistical distributions derived from over 1.5 million randomly sampled MEDLINE abstracts. Therefore, given a query document, we were able to use the term's a priori

frequency distribution to estimate a term occurrence *p*-value to assess the significance of observing each term in the document. This method enables us to accurately and quantitatively rank-order all the terms that occur in a document, produce document gist automatically, and use gists instead of short keyword list or the whole text to represent, retrieve, classify, and cluster biomedical documents. In the future, we plan to make quantitative comparisons on how well recall and precision are improved over existing methods. Our second method, used to improve recall, incorporated multiple biological data sources as thesaurus to help expand user query term. This is especially important because, during query expansions, users can control the thesaurus types and expansion levels according to his/her disease biology areas. Therefore, the retrieved documents are going to contain less generic information for a researcher interested in 'breast cancer', but more specific information such as 'BRCA' gene automatically expanded from Gene Ontology thesaurus.

We also demonstrated that it is possible to incorporate our method into a novel biomedical document management system using Oracle 10*g* database as a platform. This system, with a large number of abstracts, can perform document search, clustering, and classification tasks with good level of interactivity. In the future, we plan to expand the system to allow integrated search of document's metadata and to accommodate all the 15 million MEDLINE abstracts.

There are remaining challenges to improve document retrieval beyond single term queries. For example, a meaningful biomedical query phrase may consist of multiple words. For the moment, our method deals with multi-term query phrases by averaging the *p*-values of all the terms from the query phrase. However, the more accurate practice is to collect statistics of such phrases via sampling process. We plan to modify our method to address this issue in the future. For another example, thesaurus is critical to keyword based document retrieval. Therefore, it would be useful to introduce several important thesauri into the database to enable users expand keywords using thesaurus from their scientific domain.

Acknowledgements

We thank Stephanie Burks from Indiana University Research and Academic Computing for maintaining Unix servers and Oracle 10g database servers for this work. This work is partially supported by funding from Indiana University – Purdue University RSFG grant award made to Dr. Jake Yue Chen. We also thank Susie Stephens, Raf Podowski and Charlie Berger from Oracle Corporations for their help launching this project.

This work was supported in part by systems obtained by Indiana University through its relationship with Sun Microsystems Inc. as a Sun Centre of Excellence.

References

Magnini, B. and Prevete, R. (2000) 'Exploiting lexical expansions and Boolean compositions for web querying', *ACL'2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval*.

Rindflesch, T.C., Hunter, L. and Aronson, A.R. (1999) 'Mining molecular binding terminology from biomedical text', *Proceedings of the American Medical Informatics Association Symposium*, pp.127–131.

Salton, G. and McGill, M.J. (1983) *Introduction to Modern Information Retrieval*, McGraw-Hill, New York.

Salton, G. and Young, C.S. (1973) 'On the specification of term values in automatic indexing', *Journal of Documentation*, Vol. 29, No. 4, pp.351–372.

Schuler, G.D., Epstein, J.A., Ohkawa, H. and Kans, J.A. (1996) 'Entrez: molecular biology database and retrieval system', *Methods Enzymol*., Vol. 266, pp.141–162.

Srinivasan, P. (2004) 'Text mining: generating hypotheses from MEDLINE', *Journal of the American Society for Information Science and Technology*, Vol. 55, No. 5, pp.396–413.

Wheeler, D.L., Chappey, C. *et al*. (2000) 'Database resources of the national center for biotechnology information', *Nucleic Acids Res.*, Vol. 28, No. 1, pp.10–14.

Wilbur, W.J. and Yang, Y. (1996) 'An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts', *Comput. Biol. Med.*, Vol. 26, pp.209–222.

Websites

CPAN, http://search.cpan.org/dist/Statistics-Distributions/

Entrez PubPed http://www.ncbi.nlm.nih.gov/entrez/

NCI Terms http://nciterms.nci.nih.gov/

Online Mendelian Inheritance in Man OMIM (TM), McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD), URL: http://www.ncbi.nlm.nih.gov/omim/

Oracle Life sciences Platform http://www.oracle.com/technology/ industries/life_sciences/index.html