A Bacterial Textual Processing and Retrieval System

Tyne Liang* Yu-teng Chang Dian-song Wu

Department of Computer and Information Science
National Chiao Tung University
Hsinchu, Taiwan 30050

***email: tliang@cis.nctu.edu.tw**

***responsible for all correspondences.**
Fax: 886-3-5721490 and telephone: 886-3-5131365

Abstract

In this paper, a web-based bacteria textual processing and retrieval system is presented with the purpose to support biological researchers a unified retrieval access as well as to ease their data management. The system contains two main parts, namely, thesaurus construction module and retrieval module. Three new thesauri are built on the basis of statistical approaches and they are verified with real corpora to be useful for document indexing, categorization and retrieval. On the other hand the proposed unified retrieval module simplifies users' access task to deal with various kinds of databases either at local sites or remote sites. It is also embedded with ranking function for relevance judgment as well as on- line information extractors, such as indexer, bacteria predictor and the pattern extractor, for data management. All these proposed methods can be easily adaptable to other domains.

Keywords: thesaurus creation, bacteria, data processing, retrieval system, information extraction.

1. Introduction

Nowadays most biological information sources become autonomous and distributed across heterogeneous platforms. In addition to different types of call interfaces and query interfaces, the semantic heterogeneity between the many data-sources and analysis tools increases the user burden to undergo their information request. Hence there is growing need for a unified retrieval system for users to simply their information requests. In past few years, several famous systems were proposed to ease access to multiple sources. For example, TAMBIS, CANCERLIT, MELISA … , et al. [1, 2, 3, 5, 6]. It is found that most of these successful retrieval systems are generally incorporated with a thesauri system to facilitate document indexing and retrieval, yet manual thesauri construction is essentially time-consuming [5, 7. 8, 10, 12].

In this paper a bacteria retrieval and processing system is presented with the purpose to ease thesaurus construction as well as users' information requests. This system is

incorporated with an automatic thesaurus construction module as well as a unified retrieval module. The thesaurus construction is based on statistical methods by using large-scale bacteria corpora which are automatically collected from MEDLINE [8, 10]. Except the original MeSH thesaurus [8], three new thesauri are generated, namely, MeSH term clusters, significant bacteria descriptors and verb patterns. Meanwhile the usage of these thesauri is verified by comparing our indexing system with PubMed [8], a web-based search engine by a thesaurus-based indexing scheme [1]. From the experimental results it is observed that the newly created thesauri indeed facilitate document indexing as well as document categorization.

On the other hand the proposed unified retrieval module simplifies users' access task to deal with various kinds of databases either at local sites or remote sites. Unlike the PubMed which is lack of ranking functions on retrieved documents, the retrieval module is embedded with a ranking scheme and allows users to browse their corresponding information detail, such as the Taxonomy [8] and CCRC records [13] at their disposal. In addition, any retrieved paper can be processed by the on-line information extractors such as indexer, bacteria predictor and the pattern extractor, so that its important information can be extracted and stored into structured database. All these proposed methods can be easily adaptable to other domains. It is believed that the implementation of such kind of system will benefit both the information scientist in the context of knowledge discovery and at the same time provide an efficient biological data management and query resolution tools for researchers in microbial strain researchers community.

2. The Proposed System

As in Figure 1, our proposed system consists of three parts: source module, bacteria thesauri and the retrieval system. In source module, we collect the data from PubMed, Taxnomy, databases in NCBI [8] and the bacteria data in Culture Collection and Research Center (CCRC) database of Food Industry Research and Development Institute, Taiwan [13]. After collecting data, we index the data and store in local database. In bacteria domain thesauri construction we construct three thesauri include MeSH term clusters, bacteria descriptors thesaurus and verb pattern database. The MeSH term clusters to support the article ranking and query expansion. The bacteria descriptors thesaurus stores the related MeSH terms for each bacterium and will be used as bacteria prediction at document categorization. The verb patterns database stores the MeSH term patterns of interesting verbs to facilitate the template query search. Finally the information retrieval system provides an unified interface to users to retrieve the information as requested.

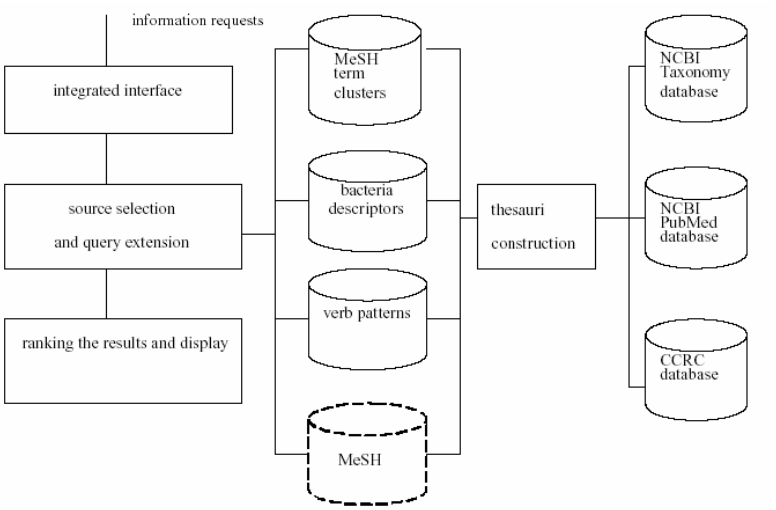

Figure 1: The architecture of proposed system.

2.1 Thesaurus construction

In this paper the bacteria thesaurus include MeSH term clusters, bacteria descriptors and verb patterns and they will be used at content detection and information filtering. The thesaurus construction is based on statistical approaches by using a large-scale of corpus. It is believed that such construction not only support full automatic thesaurus creation but also ease knowledge management for biologists in the course of information search.

2.1.1 MeSH term clusters creation

MeSH term clusters are used for automatic indexing and it is generated from the corpus related to bacteria. The corpus is generated by sending 30354 bacteria names gathered from NCBI Taxonomy database downto to level 11 (http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi). Then 267448 MEDLINE articles are retrieved from PubMed database (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed) and they contain 20742 MeSH terms (http://www.nlm.nih.gov/mesh/meshhome.html) after stemming process. Then the frequency of each stemmed MeSH term is recorded and a term-to-document matrix is built. The value in the matrix is calculated as equation (1) where freqi,j is the frequency of term i in document j and maxl freql,j is the maximum term frequency in document j.

$$f_{i,j} = \frac{freq_{i,j}}{\max_l freq_{l,j}} \quad (1)$$

The construction is mainly based on the latent semantic indexing (LSA) which is a statistical model of word usage that permits comparison of the semantic similarity between textual information. In this paper we use *Singular Value Decomposition (SVD)* to realize our LSA scheme due to its simple computation. The implementation is as below:

(1) Construct a Term – Document matrix X

(2) Transform matrix X into production of three matrix T, S, D by SVD

$$X = T \times S \times D^T \quad (2)$$

(3) Find a suitable value K from matrix S, and K is the new rank for matrix T, S, D .Reduce the original matrix T, S, D and get three new matrixes T_m, S_m, D_m.

(4) Do the production of T_m, S_m, D_m, and we get a new matrix X'.

(5) The value in the new matrix X' represents the importance for each term in each document.

The result of SVD is a reweighed term-document matrix A' and then the similarity between terms can be calculated by cosine measure.

$$similarity\ \ (t_i, t_j) = \frac{\vec{v_i} \bullet \vec{v_j}}{\left|\vec{v_i}\right\|\vec{v_j}\right|} \qquad (3)$$

where $\vec{v_i}$ is the vector of term i and $\vec{v_j}$ is the vector of term j. $\vec{v_i} \bullet \vec{v_j}$ is the inner product of vector term t_i and term t_j and $\left|\vec{v_i}\right|, \left|\vec{v_j}\right|$ are the length of vector term t_i and t_j. Finally for each term we select the top 30 terms with highest similarity to form a MeSH term sets.

2.1.2 MeSH-term indexing experiments

Three experiments were implemented to justify the performance of the proposed indexing scheme. In the first experiment we compare the index terms similarity between MEDLINE indexing and the SVD-based method. The testing corpus of 257448 articles is collected from PubMed by sending 30355 bacteria names which are from Taxonomy with taxonomy hierarchy downto level 11 (http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi). Each article will be indexed by using the constructed MeSH term sets. The indexing is done in the following:

(1) collect the MeSH terms from each article
(2) calculate MeSH term weight in the article by equation (4)
(3) select the top weight terms as many as the PUBMED does.

$$weight\ \ (t_i, d_j) = [\ \sum_{q \in I}^{30} freq\ (t_q, d_j) \times s(t_i, t_q)\ ] \times idf\ (t_i) \qquad (4)$$

where I is the set of t_i, $s(t_i, t_q)$ is the similarity of t_i and t_q, $Freq(t_q, d_j)$ is term frequency of t_j in d_j and $idf(t_i)$ is the inverse document frequency of t_i in the corpus.

Then the similarity between the MeSH terms indexed by PubMed and our index method is calculated by equation (5):

$$similarity\ \ (C_{Med}, C_{LSA}) = \frac{C_{Med} \cap C_{LSA}}{C_{Med} \cup C_{LSA}} \qquad (5)$$

C_{LSA}: the index set by our LSA result
C_{Med}: the index set by MedLine
$C_{Med} \cap C_{LSA}$: intersection of C_{Med} and C_{LSA}
$C_{Med} \cup C_{LSA}$: union of C_{Med} or C_{LSA}

Figure 2 is the similarity distribution between MEDLINE index and our index method.

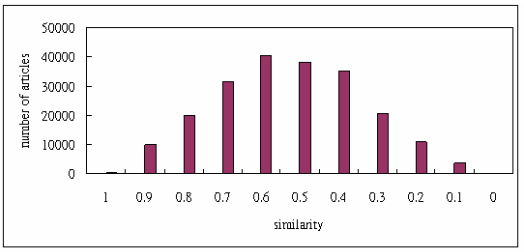

Figure 2: Similarity between MEDLINE-index and LSA-index

The second test is to investigate the rank order of the bacteria names appearing in documents. The testing documents are generated by selecting those documents, out of 257448 articles, which contain bacteria names as their index terms. Then we found that there is about 81% of the testing documents in which the bacteria names appear in the top five indexing terms.

Table 1: The prediction results of index ranking.

select top n index terms	1	2	3	4	5
number of articles	43.6%	45%	51.5%	65.2%	81.3%

Finally the relevance check was implemented with a small testing corpus due to manual cost. First we randomly select thirty

bacteria names as queries input, then we retrieved the top five articles ranked by the weighting function. A biologist was asked to manually check document relevancy. From this small sample test we found that there are 71% of the 150 retrieved articles relevant to user's queries.

2.1.3 Bacteria descriptors finding

The bacteria descriptor thesaurus is constructed to find the significant terms correlated to each bacterium so that a new document in bacteria domain can be categorized into appropriate bacteria class.

Concerning that there are more than twenty thousand bacteria names listed in NCBI bacteria Taxonomy, we use all the twenty bacteria names appearing to the Taxonomy level 1. The twenty bacteria names are "Aquificae", "CFB", "Chlamydiae", "Chloroflexi (green non-sulfur bacteria)", "Chrysiogenetes", "Cyanobacteria (blue-green algae)", "Deferribacteres", "Dehalococcoides group", "Deinococcus-Thermus", "Dictyoglomi", "Fibrobacteres", "Firmicutes (Gram-positive bacteria)", "Fusobacteria", "Nitrospirae", "Planctomycetes", "Proteobacteria (purple bactera and relatives)", "Spirochaetes", "Thermodesulfobacteria" , "Thermomicrobia" and " Thermotogae".

The significant term finding is based on the weighting scheme which both the term concentration and distribution of document space are taken into account [4]. The weight of term t_i in bacterium B_k is calculated as below:

$$weight(t_i, B_k) = \frac{\sum_{j=1}^{n} weight(t_i, d_j, B_k)}{n \times 2^{Entropy(t_i)}} \quad (6)$$

where weight (t_i, d_j, B_k) is the weight of t_i of d_j in B_k and is calculated as equation (7), n is number of documents in B_k, and entropy of t_i is calculated as equation (8).

$$weight(t_i, d_j, B_k) = \alpha \frac{df(t_i, B_k)}{sum\_df(B_k)} + \beta \frac{tf(t_i, d_j, B_k)}{sum\_tf(d_j, B_k)} + \gamma \frac{cf(t_i, d_j, B_k)}{sum\_cf(d_j, B_k)} + \delta \frac{length(t_i)}{\max\_length(d_j)} \quad (7)$$

$\alpha = 0.4 \quad \beta = 0.3 \quad \gamma = 0.2 \quad \delta = 0.1$

$df(t_i, B_k)$: number of documents that t_i occurring in B_k.
$tf(t_i, d_j, B_k)$: the frequencies of t_i occurring in title of d_j in B_k.
$cf(t_i, d_j, B_k)$: the frequencies of t_i in the abstract of d_j in B_k.
$length(t_i)$: the length of term t_i.
$sum\_df(B_k)$: number of documents in bacterium B_k.
$sum\_tf(d_j, B_k)$: sum of occurrence of all terms occur in title of d_j in B_k.
$sum\_cf(d_j, B_k)$: sum of occurrence of all terms occur in abstract of d_j in B_k.
$max\_length(d_j)$: maximum length of terms in d_j.

$$Entropy(t_i) = \sum_{k=1}^{n} \left[\frac{df(t_i, B_k)}{\sum_{i=1}^{n} df(t_i, B_k)} \times \log \frac{\sum_{i=1}^{n} df(t_i, B_k)}{df(t_i, B_k)} \right] \quad (8)$$

In the end the terms with top 50 weights are selected for each bacterium.

2.1.4. Bacteria prediction experiments

The performance of the bacteria descriptor finding is verified with a real test data set which are the 3206 articles listed in the reference in Taxonomy. Each article will be indexed with the system indexing scheme and will be calculated its *weight(B_k, d_j)* for each bacterium B_k as Equation (9):

$$weight(B_k, d_j) = \sum_{i \in A} [weight(t_i, d_j) \times weight(t_i, B_k,)] \quad (9)$$

where A is the index set of d_j by the proposed indexing scheme.

During the test we record the top five bacteria names for each test article. As shown in Table 2, 80% of references are categorized correctly by the system at the first try. In other words, their first predicted bacteria names are just as the same as the names with which they are categorized in the Taxonomy reference list. 88% references are categorized correctly when their bacteria names appear in the top-five list.

Table 2: The prediction accuracy by the proposed weighting scheme.

rank	1	2	3	4	5
prediction accuracy	80%	82%	84%	85%	88%

2.2 Verb pattern extraction

The verb pattern extraction is implemented with the aim to support template search which is close to be natural-language-like queries since there are some verbs playing important roles during information requests in molecular domains. In the proposed system, twenty verbs that are important in biology are used as indexing keywords. They are "induce", "inhibit", "infect", "transform", "reduce", "react", "develop", "growth", "enhance", "correlate", "bind", "link", "treat", "culture", "prevent", "supply", "produce", "mutate", "synthesize" and "express". Using these verbs and MeSH hierarchy, we could build patterns database which supports sentence-like retrieval. In the proposed retrieval system term expansion can be done on the basis of MeSH tree structure.

For example, if a user inputs the sentence "find the antibiotic resistance reacted on gram-negative bacteria," the system will find the pattern that verb_type="induce", factor1="antibiotic resistance" and factor2="gram-negative bacteria". Then system will retrieve the articles that contain the sentences with the patterns. If the retrieved articles are not enough, system will expand the factor according to MeSH hierarchy.

3. The Unified Retrieval System

Figure 3 is the retrieval system flowchart in which the query processor accepts user's queries and transforms them into internal forms. The local database module is to retrieve data in the local database which contain the taxonomy bacteria data and PubMed articles we retrieved in advance. The remote database agent is to retrieve the data in PubMed databases.

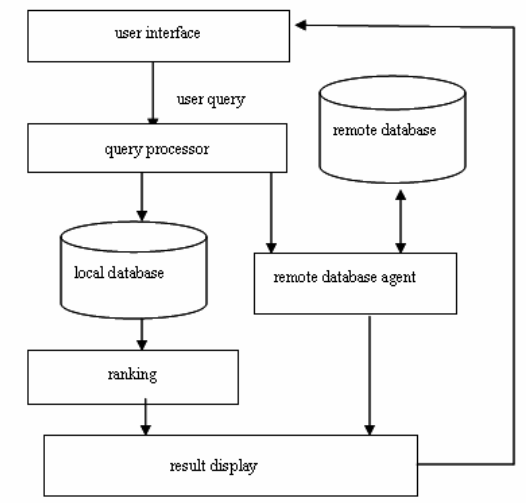

Figure 3: The architecture of retrieval. system.

The menu-based user interface supports user to select the function of local database search, remote database search, advanced search and database update and maintenance. User could input his queries to search the articles in our local database or the remote database that we specified. If user selects the database update and maintenance, user could input article and operates the function of indexing, prediction and pattern extraction. As shown in Figure 4 the local database search supports multiple attribute search with easy Boolean expression.

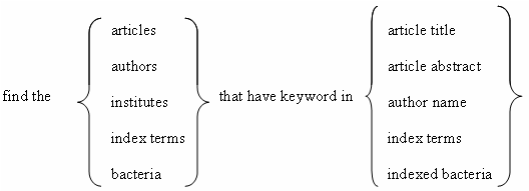

Figure 4: The query types supported by local database search.

After processing the queries, system will show the result page which lists the top 150 articles by the system ranking function. Meanwhile the system supports details information for each article as shown in Figure 5 in which the top five bacteria names and their corresponding information such as synonym name, reference, classification hierarchy,…,etc. can be also easily browsed. Similar functions are supported for

advanced search.

Figure 5: The detail article information page.

On the other hand users can use the supported remote database search agent to access the remote databases and the query types are generated based on SQL-like expression such as "**Find** the articles **from** PubMed **where** article title contains "Aquificae". After the remote database returning the query results, the presented system will parse the results and list the article titles in the result display page. Then the title, abstract, author and journal information will be extracted automatically by the proposed on-line extractor. Similarly the article can be indexed by the on-line indexer and bacteria categorization by the predictor.

Figure 6: The indexing, prediction and pattern extraction page.

4. Conclusion and Future Works

In this paper, a web-based retrieval and processing system for bacteria texts is presented. The system supports unified access to different databases, as well as the function such as full automatic parsing, indexing and categorization. Meanwhile the system is embedded with an automatic thesaurus construction by statistical approaches. Such construction will not only be useful for biologists to manage knowledge but also facilitate knowledge discovery in the field of molecular biology.

In the future, we will integrate the other bacteria information such as sequence information into our system so as to present users with more complete information. In addition, application of information extraction techniques to mine interesting biological relations will be concerned in our future direction so as to enhance the automation of knowledge base construction.

Acknowledgement: This paper is partially supported by National Science Council, R. O. C. under the contract NSC-91-2213-E-009-082.

Reference

[1] A. R. Aroson, O. Bodenreider, H. F. Chang, S. M. Humphrey, J. G. Mork, S. J. Nelson, T. C. Rindflesch, W. J. Wilbur (2000) "The NLM Indexing Initiative," *AMIA Annual Fall Symposium*, pp.17-21.

[2] J. M. Abasolo and M. Gomez, (2000) "MELISA. An ontology-based agent for information retrieval in medicine," *Proceedings of the First International Workshop on the Semantic Web*, pp. 73-82.

[3] P. G.. Baker, C. A. Goble, S. Bechhofer, N. W. Paton, R. Steven and A.Brass(1998) "TAMBIS-Transparent Access to Multiple Bioinformatics Information Sources," *In Proc. of the 6th International Conference on Intelligent System for Molecular Biology*, AAAI Press, pp. 25-34.

[4] C. C. Chang (1997) "The Study of Chinese

Textual Document Retrieval Based on Fuzzy Concept Networks," Master thesis, Nation Chiao-Tung University.

[5] H. Chen, T. Yim, D.Fye and B.Schatz (1995) "Automatic Thesaurus Generation for an Electronix Community System," *Journal of The American Society for Information Science*, Vol. 46, No. 3, pp. 175-193.

[6] B. A. Eckman, A. S. Kosky and L. A. Laroco (2001) "Extending traditional query-based integration approaches for functional characterization of post-genomic data," *Bioinformatics*, Vol. 17, No. 7, pp. 587-601.

[7] T. G. Kolda and D. P. O'Leary (1996) "A Semidiscrete Matrix Decomposition for Latent Semantic Indexing in Information Retrieval," *ACM Transactions on Information Systems*, Vol. 16, No. 4 pp.322-346.

[8] National Center for Biotechnology Information, (1999) "Entrez" http://www.ncbi.nlm.nih.gov/Entrez/.

[9] M. F. Porter, (1980) "An algorithm for suffix stripping," *Program*, Vol. 14, No. 3, pp. 130-137.

[10] D. L. Wheeler, C. Chappey, A. E. Lash, D. D. Leipe, T. L. Madden, G. D. Schuler, T. A. Tatusova and R. A. Rapp, (2000) "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Res*., Vol. 28, pp. 10-14.

[11] J. Xu and W. B. Croft, (1996) "Query expansion using local and global document analysis," *In Proc. of the ACM-SIGIR Conference on Research and Development in Information Retrieval*, Zurich, Switzerland, pp. 4-11.

[12] B. Yates and R. Neto, (1999) "Modern Information Retrieval," Addison Wesley.

[13] Food Industry Research and Development Institute, Taiwan, www.firdi.org.tw.