Using UMLS-based Re-Weighting Terms as a Query Expansion Strategy

Weizhong Zhu, Xuheng Xu, Xiaohua Hu, Il-Yeol Song, and Robert B. Allen

Abstract—Search engines have significantly improved the efficiency of bio-medical literature searching. These search engines, however, still return many results that are irrelevant to the intention of a user's query. To improve precision and recall, various query expansion strategies are widely used. In this paper, we explore the three widely used query expansion strategies - local analysis, global analysis, and ontology-based term re-weighting across various search engines. Through experiments, we show that ontology-based term re-weighting works best. Term re-weighting reformulates queries with selection of key original query terms and re-weights these key terms and their associated synonyms from UMLS. The results of experiments show that with LUCENE and LEMUR, the average precision is enhanced by up to 20.3% and 12.1%, respectively, compared to baseline runs. We believe the principles of this term re-weighting strategy may be extended and utilized in other bio-medical domains.

Index Terms—query expansion strategy, pseudo feedback, term re-weighting, UMLS

I. Introduction

THERE are many cases where the search results contain a large number of irrelevant results or may contain only some of the aspects of topics requested by the users. In many cases, novice users simply do not know how to construct efficient and effective queries. Even experienced users don't always create efficient and effective queries when searching an unknown domain. Query expansion helps users solve this problem. Query expansion adds critical terms beyond original query terms to improve the precision and/or recall. A search tool may add terms to the original query automatically or provide high-level information about the collections to the users and suggest the user to refine the original query. In this research, three query expansion strategies – local analysis, global analysis, and ontology-based term re-weighting - integrated with the UMLS (Unified Medical Language System) are compared. These methods are applied to the Ad Hoc Retrieval task of the TREC 2004 Genomics task.

The rest of this paper is organized as follows: Section 2 reviews biomedical ontology, while Section 3 reviews various techniques used in query expansion strategies. Section 4 discusses three major query expansion strategies. Section 5 presents experimental design environments. Section 6 presents experimental results and discussions. Section 7 presents conclusions and future work.

Manuscript received Dec 30, 2005.

Weizhong Zhu is with the College of Information Science and Technology, Drexel University, Philadelphia, PA, 19104 USA (e-mail: wz32@ drexel.edu).

Xuhen Xu is with the College of Information Science and Technology, Drexel University, Philadelphia, PA, 19104 USA (e-mail: xx28@ drexel.edu).

Xiaohua Hu is with the College of Information Science and Technology, Drexel University, Philadelphia, PA, 19104 USA (e-mail: thu@ cis.drexel.edu, Hu's work is supported partially by the NSF Career grant IIS-0448023 and PA Dept of Health Grant #239667).

Il-Yeol Song is with the College of Information Science and Technology, Drexel University, Philadelphia, PA, 19104 USA (e-mail: song@ drexel.edu).

Robert B. Allen is with the College of Information Science and Technology, Drexel University, Philadelphia, PA, 19104 USA (e-mail: rba@ drexel.edu).

II. The Bio-medical Ontology

A. MeSH Terms

MeSH (Medical Subject Headings) is a type of controlled vocabulary. MeSH consists of sets of descriptors in a hierarchical structure that allows searching at various levels of specificity. The roots of the hierarchical structures in MeSH are very broad headings such as "Anatomy" or "Mental Disorders." Lower levels of the hierarchy include more specific terms, such as "Ankle" and "Conduct Disorder." Altogether, there are 22,997 descriptors in MeSH. There are also thousands of cross-references that assist in finding the most appropriate MeSH Heading, for example, Vitamin C is cross-referenced as Ascorbic Acid.

B. UMLS

The Unified Medical Language System (UMLS) combines medical knowledge sources and associated language tools developed by the National Library of Medicine, including UMLS Metathesaurus, UMLS Semantic Network and SPECIALIST Lexical Tools [21]. The UMLS Metathesaurus contains information about concepts and biomedical terminology in several languages. It was developed automatically from 73 families of controlled vocabularies and 117 classification systems. The purpose of UMLS Semantic Network is to provide a consistent semantic categorization of all concepts represented in the UMLS Metathesaurus and to provide a set of useful relationships between these concepts. All information about specific concepts is found in the Metathesaurus; the Network provides information about the

1-4244-0134-8/06/$20.00 © 2006 IEEE

set of basic semantic types, or categories, which may be assigned to these concepts, and it defines the set of relationships that may hold between the semantic types. The 2004AA release of the Semantic Network contains 135 semantic types and 54 relationships. The semantic types are the nodes in the Network, and the relationships between them are the links. UMLS provides symbolic relationships in 5 million pairs of concepts and statistical relations in 6.5 million pairs of co-concepts. The symbolic relations mainly include hierarchical and associative relationships. The associative relationships are represented as relationships between semantic types of concepts. Statistical relations are computed by determining the frequency with which concepts in specific vocabularies co-occur in records in a database. For instance, there are co-occurrence relationships for the number of times concepts have co-occurred as key topics within the same articles, as evidenced by the Medical Subject Headings assigned to those articles in the MEDLINE database.

III. Background

Many query expansion methods have been proposed to improve the precision and/or recall by searches. There are two general strategies for query expansion. One is ontology-based; the other is statistical. Previous studies showed that expanding a query with synonyms or hyponyms has a limited effect on biomedical information retrieval performance [4-7]. There are several reasons for that: ontological methods use no notion of weight; they do not consider actual documents but use only prior knowledge. Some authors [2, 11, 16] explored different weighting methods but did not report how the weights were used in conjunction with the ontology.

Statistical methods focus on documents and they can be further divided in two main sub-categories: global analysis and local analysis [14]. Global analysis analyzes the whole collection of documents to extract co-occurrence of related terms. Global analysis methods include term clustering, latent semantic indexing, and similarity thesauri. One of the major drawbacks of global analysis methods is that the methods require semantic similarity and disambiguation of terms.

Local analysis extracts highly-related terms from the relevant documents retrieved by an initial query or from data mining results. Xu and Croft [15] introduced Local Context Analysis which uses the top documents returned by an initial query but selects the terms based on co-occurrence with query terms. This approach, because it usually requires less human intervention, assumes that the certain numbers of top documents returned by the initial query are actually relevant ("pseudo-relevance feedback"). However, these methods are not robust because it is almost impossible for all search engines or mining methods to return only relevant documents. So studies on a hybrid of global analysis and local analysis may be more promising. Our approaches broadly investigate local analysis, global analysis and ontology-based term re-weighting as described in the next section. Our study should provide heuristics on how to combine these methods.

IV. Query Expansion Strategies

A. Local Analysis

This method determines the expanded terms based on pseudo-relevance feedback. By examining the top N documents retrieved from the initial query, Latent Semantic Indexing (LSI) and Association Rule (AR) algorithms are used to seek the top co-terms of the original terms from the top N (N=300) retrieved documents. Co-terms are weighted according to the total term frequency in the top N retrieved documents. The processes of local analysis are depicted in Fig. 1.

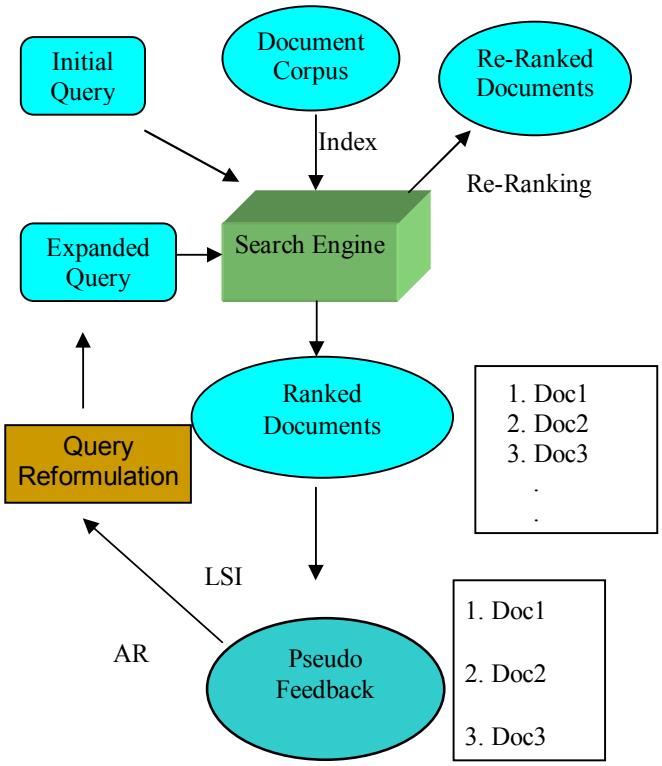

Fig. 1. Query expansion procedures of local analysis.

In this study, only TF-IDF based Vector Space information retrieval model has been used to test local analysis. Relevance feedback for the Vector Space Model can be represented by the Rocchio formula [10]:

$$q_{i+1} = \alpha q_i + \frac{\beta}{|Dr|}\sum_{d_j \in Dr} d_j - \frac{\gamma}{|Dn|}\sum_{d_j \in Dn} d_j \quad (1)$$

For expanded query q_{i+1}

q_i: the initial query

Dr: a set of relevant documents among retrieved documents

Dn: a set of non-relevant documents among retrieved documents

α,β,γ: tuning constants

In this model the information in relevant documents is treated more as more important than the information in non-relevant documents ($\gamma << \beta$). This study applied pseudo-relevance feedback in that the top N retrieved documents are supposed to be all relevant: γ equals 0.

Latent Semantic Indexing (LSI) begins with a matrix of terms by documents (row: terms, column: documents) [3, 8, 13]. This matrix is then analyzed by Singular-Value Decomposition (SVD) to obtain the latent semantic structure. SVD decomposes a term-document matrix into three separate matrices, a term-by-concept matrix, a concept-by-concept matrix and a concept-by-document matrix. In this research, the dimension of the concept-by-concept matrix is set to 50. SVD can be thought of as deriving a set of uncorrelated indexing variables or factors; each term and document is represented by its vector of factor values. For instance, the value of the cosine between two row vectors reflects the extent to which two terms have a similar pattern. We use this technique to compare the similarities between the terms in the matrix extracted from the documents.

An Association Rule (AR) identifies collections of data attributes that are statistically related in the underlying data [1, 9]. It finds associations and/or correlation relationships among data items. Similar to the LSI-based algorithm, a matrix of terms-by-documents is generated and then analyzed. An association rule is of the form B -> A, where B and A are disjoint conjunctions of attribute-value pairs. Here we take C concept as input, and then we calculate all B-> C rules. Last, we generate all A-> B rules. We acquire the related terms to initial query terms based on ARs.

The steps to expand the query terms based on LSI and AR algorithms are the following:

Step 1: Derive a term matrix based on the top N (N=300) retrieved documents for one initial query returned by the search engine.

Step2: Run LSI or AR algorithms to acquire the related terms to original query terms. We choose the top M (M=5) co-terms of the original terms from the top N (N=300) retrieved documents.

Step3: Expand the query terms including the related terms acquired above.

B. *Global Analysis*

In global analysis, terms to be added are extracted from all the documents of the whole collection. The initial query will be expanded by UMLS co-concepts of the original key terms with the same semantic types. UMLS provides co-concepts and related co-occurred frequencies for many of the medical terms appeared in MEDLINE during the past years. For instance, term "transgenic mice" in a query has a most frequent co-concept "mice, knockout" in UMLS with the same semantic type. Term "mice, knockout" will be added to the initial query. The key term "transgenic mice" in the initial query is extracted by LingPipe [19], an efficient tool for bio-medical name entity extraction. The processes of global analysis are depicted in Fig. 2.

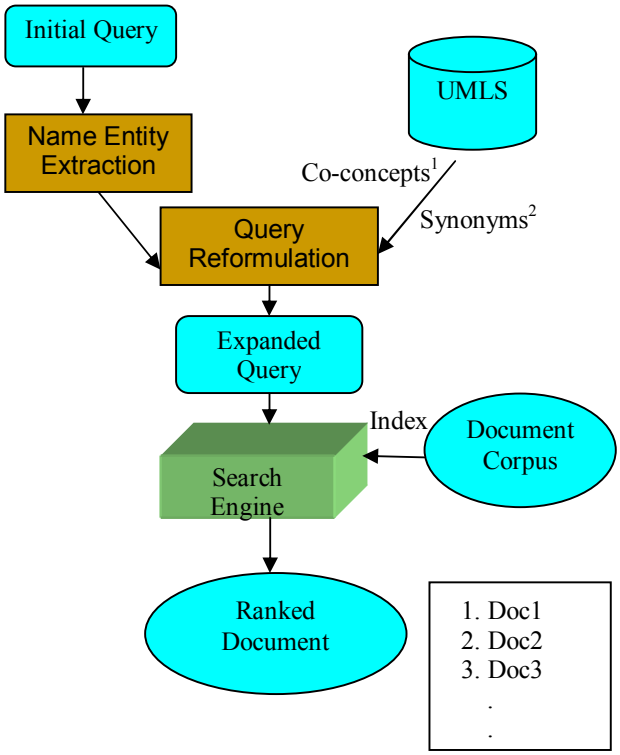

Fig. 2. Query expansion procedures of global analysis and term re-weighting strategies. Here, *1* notes global analysis and *2* notes term re-weighting.

C. *Term Re-weighting*

This approach enhances the weights of key original terms or co-terms according to their relative importance in queries. Many medical terms have very complex symbolic relationships among them. For instance, the human gene Ferroportin-1 has many synonyms, such as SLC40A1, Ferroportin 1, FPN1, HFE4, IREG1, Iron regulated gene 1, Iron-regulated transporter 1, MTP1, SLC11A3, and Solute carrier family 11. There are two options to expand such terms. One approach is just adding all the synonyms of Ferroportin-1 to the original query. Previous study [6] showed limited success when query expansion terms were generated from UMLS and equally weighted. Another method, which we also explored, boosts the weight of the original query term (e.g., Ferroprotin-1).

Specifically, we employed two principles for determining the weights. The first is that if an original term has a higher term frequency in the initial query with a specific major UMLS semantic type, the term should be given a higher weight in the expanded query. The second one is that if a key original term or an expanded term has a pre-selected major UMLS semantic type, its preferred MeSH term synonym defined in UMLS will be expanded and given a higher weight. The selection of key original terms is decided by the tagging of the Name Entity Extraction tool, LingPipe. In TREC 2004 Genomics Ad Hoc Retrieval Task, most topics discuss the

functionality of certain genes or proteins. So, the major UMLS semantic type for these topics is selected as "Amino Acid, Peptide", or "Protein". For instance, protein "NEIL1" is located in the initial query and we used a boost score of 4. Moreover its preferred MeSH term "NEIL1 protein human" extracted from UMLS is expanded and a boost score of 8 is assigned. Generally, the preferred MeSH terms in UMLS are used to index the MEDLINE abstracts. Thus, it is reasonable to give them a higher weight. The processes of term re-weighting are depicted in Fig. 2.

We studied this strategy for query expansion using two search engines, LUCENE and LEMUR, to ensure the generality of our findings.

With LEMUR, TF-IDF based Vector Model and Okapi BM25 Model were selected to verify the term re-weighting approach. Ranking algorithms for both of the models are calculated by Okapi term frequency (TF) formula [12, 20]. Based on Okapi query TF formula, key terms or phrases in queries are re-weighted simply by increasing their query term frequency according to assigned boost scores.

LUCENE search engine supports customized term boost. The boost factor, $boost_t$, is a part of the LUCENE rank algorithm [17], which is computed as follows:

$$sim(q,d) = sum_t(tf_q * idf_t / norm_q * tf_d * idf_t / norm_d^t * boost_t) * coord_q^d \quad (2)$$

where

sim(q,d): similarity score between query q and document d
sum_t: the sum weight scores for all terms t
tf_q: the square root of the frequency of t in the query
tf_d: the square root of the frequency of t in d
idf_t: log(numDocs/docFreqt+1) + 1.0
numDocs: the number of documents in index
docFreqt: the number of documents containing t
$norm_q$: sqrt(sumt((tfq*idft)^2))
$norm_d^t$: square root of number of tokens in d in the same field as t
$boost_t$: the user-specified boost for term t
$coord_q^d$: the number of terms in both query and document / the number of terms in query

V. Experiment Design and Implementation

The goal of TREC 2004 Genomics Ad Hoc Retrieval Task is to find all the relevant documents to the 50 topics from the whole corpus. The structure of this task was a conventional searching task based on a 10-year subset of MEDLINE (about 4.5 million documents and 20 gigabytes in size with NLM XML format) and 50 topics derived from information needs obtained via interviews of biomedical researchers. There was no training data, although sample topics and relevance judgments were available [18].

A. Corpus Indexing

Two search engines were used to index corpus, LEMUR (v. 4.0) and LUCENE (v. 1.4.3). LUCENE is a high performance, scalable information retrieval library. It provides a simple, but powerful, core that supports indexing and searching capabilities. LEMUR is a tool set designed to facilitate research in language modeling and information retrieval. It supports the construction of basic text retrieval systems using language modeling methods, as well as traditional methods such as those based on the Vector Space Model and the statistical models, such as Okapi BM25 Model. Vector Space Model and Okapi BM25 Model are both implemented in LEMUR. It also provides parsers to index TREC format documents. For index, the XML format data on TREC 2004 Genomics was transformed into general TREC format and then was inverted indexed. For LUCENE, the original XML data set was transferred into HTML format. Only the content from fields such as "TITLE", "ABSTRACT", "MESH", and "CHEMICAL SUBSTANCE" is kept in the HTML files.

B. Query Preprocessing and Construction

First, a query is tokenized and filtered by stop-word lists. Secondly open source software, LingPipe (v. 2.0.0), is used to extract "named entities" from topics. An example of a topic tagged by LingPipe is shown as follows:

```
<TOPIC>
<ID><sent>1</sent></ID>
<TITLE>
<sent><ENAMEX id="0" type="protein_molecule">Ferroportin-1
</ENAMEX> in <ENAMEX id="1"
type="multi_cell">humans</ENAMEX>
</sent>
</TITLE>
<NEED>
<sent>Find articles about <ENAMEX id="0"
type="protein_molecule">Ferroportin-1</ENAMEX>, an
<ENAMEX id="2" type="DNA_domain_or_region">iron
transporter</ENAMEX>, in <ENAMEX id="1"
type="multi_cell">humans</ENAMEX>.
</sent>
</NEED>
<CONTEXT>
<sent>
<ENAMEX id="3" type="other_name">Ferroportin1</ENAMEX>
(also known as <ENAMEX id="4"
type="DNA_domain_or_region">SLC40A1; Ferroportin 1; FPN1;
HFE4; IREG1; Iron regulated gene 1; Iron-regulated transporter 1;
MTP1; SLC11A3</ENAMEX>; and <ENAMEX id="5"
type="protein_family_or_group">Solute carrier family 11 (proton-
coupled</ENAMEX> divalent <ENAMEX id="6"
type="atom">metal ion transporters</ENAMEX>), member 3) may
play a role in <ENAMEX id="7" type="other_name">iron
transport</ENAMEX>.
</sent>
</CONTEXT>
</TOPIC>
```

LingPipe's entity extraction is based on the Bayesian generative model that tags each token as being the beginning

of a named entity, a continuation of a named entity, or not in a named entity. The model is trained on the GENIA corpus, which closely matches a subset of MEDLINE abstracts. Roughly, LingPipe has only about 70% accuracy. Here, the label is ignored and the boundaries are used to identify the key terms or phrases in the query. For instance, if a phrase is labeled as entity type of "protein_molecule", the terms included in the phrase are key terms. Then, these key terms are selected to extract co-concepts or preferred MeSH terms from UMLS for further expansion or re-weighting. Additionally, the query will be stemmed in LEMUR before query processing.

C. Query Input to Search Engines

The TREC 2004 Genomics topics consist of three parts: title, need, and context shown in the above example. The titles describe the information need in a few abbreviated words and they are most similar to queries entered by end users. In this research, the title of each topic is selected for baseline runs. Context describes information about the environment in which queries may occur. So it may be helpful to expand the original query, the titles. But in this study, only these key terms or phases labeled by LingPipe in context information of the topics were selected to expand the original queries.

VI. Result and Discussion

A. TREC Evaluation

Recall and precision for the ad hoc retrieval task were calculated in the classic IR way, using the preferred TREC statistics of mean average precision (average precision at each point a relevant document is retrieved, also called MAP). This was done using the standard TREC approach of participants submitting their results in the format for input to Buckley's trec_eval program.

B. Experimental Results

The three query expansion strategies have been tested with Lemur search engine (Vector Space Model and Okapi BM25 Model) and LUCENE search engine (Vector Space Model). The experimental results are listed in Tables I, II, III respectively. We used the non-interpolated average precision, precision at 10 documents (P@10) and ratio change compared with baselines as our evaluation metrics.

C. Discussion

There are seven types of runs designed in the experiments, Baseline, UMLS_Global (title + co-concepts in UMLS), LSI_local (title + local co-terms generated by LSI), AR_local (title + local co-terms generated by AR), Baseline+Re-weighting (title + term re-weighting approach), Baseline+context (title + key terms or phrases in context of query), and Baseline+Context+Re-weighting (title + key terms or phrases in context of query + term re-weighting approach). The ratios of change are compared to Baselines.

The average precision of our baseline runs are better than the results of the 2004 TREC Genomics Ad Hoc task where the average precision of 47 runs submitted by 32 teams was 20.74%. Our best runs are 33.74% and 28.91% for LEMUR and LUCENE, respectively.

TABLE I
EXPERIMENTAL RESULTS FROM LEMUR SEARCH ENGINE
(VECTOR SPACE MODEL)

Run	Average Precision	P@10	Ratio Change to Baseline
Baseline	0.3024	5.04	---
UMLS_Global	0.3007	4.76	-0.6%
LSI_Local	0.3039	5.04	+0.5%
AR_Local	0.2842	4.62	-6.0%
Baseline+Re-weighting	0.3092	5.10	+2.2%
Baseline+Context	0.3231	5.04	+6.8%
Baseline+Context+Re-weighting	0.3252	4.98	+7.5%

TABLE II
EXPERIMENTAL RESULTS FROM LEMUR SEARCH ENGINE
(OKAPI BM25 MODEL)

Run	Average Precision	P@10	Ratio Change to Baseline
Baseline	0.3010	5.16	---
UMLS_Global	0.2562	4.64	-14.9%
Baseline+Re-weighting	0.3131	5.06	+4.0%
Baseline+Context	0.3107	5.54	+3.2%
Baseline+Context+Re-weighting	0.3374	5.28	+12.1%

TABLE III
EXPERIMENTAL RESULTS FROM LUCENE SEARCH ENGINE
(VECTOR SPACE MODEL)

Run	Average Precision	P@10	Ratio Change to Baseline
Baseline	0.2404	4.36	---
UMLS_Global	0.1858	3.34	-22.7%
Baseline+Re-weighting	0.2512	4.42	+4.5%
Baseline+Context	0.2488	4.82	+3.5%
Baseline+Context+Re-weighting	0.2891	4.68	+20.3%

Co-term expansion with local analysis increases average precision by only 0.5% (LSI). AR gives an even worse result and decreases the average precision by 6.0%. Here, AR mining may not be efficient. Instead, the association rules at sentence level may produce more precise term associations.

The term re-weighting strategy is applied to two runs on LUCENE, Baseline and Baseline+Context. The results show this approach increases average precision in both conditions by 4.5% and 20.3% (Table III). Apparently the term re-

weighting approach improves the average precision of the Baseline+Context than that of Baseline run. This implies context information of queries is critical to enhance the performance of retrieval. If the Baseline+Context run is treated as baseline, the term re-weighting approach will eventually improve the retrieval performance by 16.2%. The term re-weighting strategy is also applied to four runs on LEMUR which queries were both formed from Baseline and Baseline+Context (Table I and II). But these four runs are based on two different information retrieval models, Vector Space Model and Statistical Model (Okapi BM25 Model). The results indicate that Statistical Model that increases average precision by 12.1% could more empower the term re-weighting approach than the Vector Model that increase average precision by 7.5% with context taken into account. If without context, the term-reweighing approach on Statistical Model only increases average precision by 4.0% and the performance enhancement on Vector Space Model is about 2.2%. If the Baseline+Context runs are treated as baselines, through term re-weighting, Statistical Model could elevate average precision by 8.6%, but Vector Space Model could only enhance average precision only by 0.7%.

Co-concepts expanded from global analysis make the results worse. The average precision of the UMLS+Global run on LEMUR decreases by 0.6% (Vector Space Model) and 14.9% (Okapi BM25 Model) compared to Baseline run and the one on LUCENE decreases by 22.7% (Vector Space Model). Perhaps, the top-ranking co-concepts from UMLS co-occur frequently with original terms in certain context that is totally different from that of TREC Genomics topics.

VII. Conclusion and Future work

In this paper we compared and explored three query expansion strategies for bio-medical domain. The term re-weighting strategy showed great potential to improve precision and recall across different search engines and information retrieval models. We explicitly showed how to find most important terms in the queries and how to apply domain ontology to extend these terms. This strategy could be utilized in other bio-medical information retrieval systems. The LSI-based local analysis only slightly improved the performance of search engines, but it showed a positive trend for further investigation. So, integration with more precise ontology-based term weighting and retrieval feedback may enhance LSI-based local analysis substantially.

References

[1] Agrawal, R., et al., Fast Discovery of Association Rules, Advances in Knowledge Discovery and Data Mining, U. Fayyad. et al., Editors, 1995, AAAI/MIT Press.

[2] Cesarano, C., d'Acierno, A., and Picariello, A., An Intelligent Search Agent System for Semantic Information Retrieval on the Internet. Proceedings of the 5th ACM international workshop on Web information and data management, New Orleans, Louisiana, USA, 2003, 111-117.

[3] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R. J., Indexing by Latent Semantic Analysis, Journal of the American Society for Information Science, 41(6) 1990, 391-407.

[4] Guo, Y., Harkema, H., and Gaizauskas, R., Sheffield University and the TREC 2004 Genomics Track: Query Expansion Using Synonymous Terms, 13th Text Retrieval Conference (TREC 2004).

[5] Hersh, W., and Hickam, D., Information Retrieval in Medicine-The Sapphire Experience, Journal of the American Society for Information Science, Dec., 1995, 46(10), 743-747.

[6] Hersh, W., Price, S., and Donohoe, L., Assessing thesaurus-based query expansion using the UMLS Metathesaurus, Proc AMIA Symposium, 2000.

[7] Leroy, G., and Chen, H., Meeting Medical Terminology Needs: The Ontology-Enhanced Medical Concept Mapper, IEEE Transactions on Information Technology in Biomedicine, 2001, 5(4), 261-270.

[8] Lindsay, R.K., and Gordon, M.D., Literature-based discovery by lexical statistics, Journal of the American Society for Information Science, 1999, 50(7), 574-587.

[9] Pratt, W. and Yetisgen-Yildiz, M., LitLinker: Capturing connections across the biomedical literature, K-CAP'03, Sanibel Island, FL. Oct. 2003, 105-112.

[10] Rocchio, J., Relevance Feedback in information retrieval, In G. Salton (Ed.), The SMART Retrieval System-experiments in Automatic Document Processing (Chap 14), Englewood Cliffs, NJ: Prentice Hall.

[11] Richardson R., and Smeaton A.F, Using Wordnet in a knowledge-based approach to information retrieval, In Proceedings of the BCS-IRSG Colloquium, Crewe, 1995.

[12] S. E. Robertson and S. Walker (2000), "Okapi/Keenbow at TREC-8," in E. Voorhees and D. K. Harman (Editors), the Eighth Text Retrieval Conference (TREC-8), NIST Special Publication 500-246

[13] Weeber, M., Vos, R., Klein, H., de Jong-Van den Berg. L.T.W., Aronson, A. and Molema, G., Generating hypotheses by discovering implicit associations in the literature: A case report for new potential therapeutic uses for Thalidomide. Journal of the American Medical Informatics Association, 2003, 10(3):252-259.

[14] Xu, J., and Croft, W., Query expansion using local and global document analysis, In Proceedings of ACM SIGIR, 1996.

[15] Xu, J. and Croft, W.B., Improving the Effectiveness of Information Retrieval with Local Context Analysis, ACM Transactions on Information Systems, 18(1), 79-112, 2000.

[16] Zhu, A., Gauch, S., Lutz, G., Kral, N., and Pretschner, A., Ontology-Based Web Site Mapping for Information Exploration, Proceedings of the Eighth International Conference on Information and Knowledge Management (CIKM '99), 188-194.

[17] http://LUCENE.apache.org/java/docs/api/index.html

[18] http://ir.ohsu.edu/genomics/2004protocol.html

[19] http://www.alias-i.com/lingpipe/

[20] http://www.cs.cmu.edu/~lemur/1.9/tfidf.ps

[21] http://www.nlm.nih.gov/research/umls/archive/2004AA/umlsdoc_2004aa.pdf