

Improving the Precision of Web Search for Medical Domain using Automatic Query Expansion

Vinay Kakade

vkakade@cs.stanford.edu

Madhura Sharangpani

smadhura@cs.stanford.edu

Department of Computer Science
Stanford University,
Stanford, CA-94305.

ABSTRACT

In this paper, we have analyzed three ways of improving precision of search results of web search for medical domain for short queries using automatic query expansion from the point of view of a lay-person who is seeking more information about a particular medical term. We are using Google to perform the web search. The three approaches are: (1) retrieving synonyms of the query term using UMLS (2) retrieving most frequently occurring medical terms using pseudo feedback (3) retrieving synonyms of the query term using WordNet. The final query is generated by *OR*ing the query term with the set of terms obtained using one of the three approaches and then *AND*ing the query so expanded with its semantic type retrieved using UMLS.

We have tested our system for a total of 30 queries and our results indicate that for such kind of search, the first & third approaches improve the precision of results by 13% as compared to query search using Google, while the second approach does not give better results, but in fact reduces precision by 32%.

Keywords

Automatic query expansion, pseudo feedback retrieval, semantic search, web search

1. INTRODUCTION

1.1 Need for Semantic Search

In the past few years, there has been an enormous growth in the World Wide Web. Search engines have been created for efficiently extracting information from the web and have kept evolving into a better form. However, most of today's search engines, including Google, have a limitation that they do not take into consideration the "meaning" of the query. For example, for the query *hypertension*, a relevant

document containing the phrase *high blood pressure* but not the word *hypertension* will not be retrieved. Also, the amount of data available on the web is increasing exponentially and too much information will create confusion unless we find an efficient way to filter the available information based on the intent of user query. The algorithms used in keyword or link based search fail to do so because they are not able to fully understand the query based on only the keywords in the query sentence.

However, we can modify the user query based on its meaning to narrow down the search space and improve precision of results. In particular, if the search is focused on a specific domain, modeling the query using domain specific knowledge and adding domain specific meanings of given word to the query, the search results might improve significantly.

1.2 Web Search and Search in Medical Domain

Some peculiarities of web search are that 85% users of web search engines look over one result screen only and that too mostly above the fold. So, it is very important to improve the precision of topmost search results. Also, since most of the queries in the web search are short, it is important to focus on short queries.

Today, Google is the most popular and efficient search engine for web search. But for a specialized domain like medical, the topmost results obtained using Google may not always be relevant. A good example of this is that for term *fever*, the first document retrieved is a sports related document and only one result out of the top 10 is relevant. Another good example is the query *acidity* for which Google retrieves documents regarding soil acidity, wine acidity etc. as a part of its top 10 results. It is thus

necessary to refine the query so that the results focus of a subset of all documents available in Google database, which is relevant to medical field.

1.3 Our Approach

Hence in this paper, we focus on improving the precision of top 10 results obtained using Google for short (one word) queries in medical domain.

To achieve this, we have tried three different approaches:

- *Automatic Query Expansion using UMLS*
Here, we expand the given query by *OR*ing it with its synonyms obtained using UMLS and finally *AND*ing it with its semantic type obtained using UMLS.
- *Automatic Query Expansion using Pseudo Feedback Retrieval*
In this approach, we perform retrieval in two phases. In the first phase, we give the query to Google, and retrieve top k documents¹. Then we extract most frequently occurring medical terms in these top k documents by applying different heuristics and *OR* it with the query term. Finally, the query so expanded is *AND*ed with the semantic type obtained using UMLS.
- *Automatic Query Expansion using WordNet*
This approach is same as the first one, except that we are using WordNet instead of UMLS to get the synonym of the query term.

2. RELATED WORK

Much work has already been done in the fields of automatic query expansion and pseudo-feedback retrieval for a general corpus as well as a specialized domain like medical.

Query expansion is a technique for improving the effectiveness of and achieving better performance for information retrieval systems by expanding the original query with some salient relevant terms. Different approaches to query expansion that have already been implemented are probabilistic query expansion [5], manual query expansion using user relevance feedback [12], thesaurus based query expansion [13]. Efforts have also been made to implement combined query expansion approach to improve performance [6].

¹ All our results are with $k=40$

Automatic query expansion process emphasizes completely automatic approaches to understanding and retrieval of large quantities of text [4]. The focus is on query expansion by adding a set of relevant terms to given query from known relevant documents in case of routing and from just the top k retrieved documents in the case of ad-hoc query expansion [3]. Advantage of automatic query expansion is that it does not burden the user with the task of manually adding relevant terms to query thus saving the user time and if done efficiently can perform as efficiently as best manual query expansion method with little computational overhead [7]. Till today, UMLS has been used extensively for automatic query expansion in medical domain. William Hersh, Susan Price and Larry Donohoe have implemented query expansion using relationships and definitions in UMLS Metathesaurus [13]. Alan and Thomas have experimented use of MetaMap programs for associating UMLS Metathesaurus concepts with original query [1].

Pseudo relevance feedback is usually implemented using classical Rocchio Algorithm [2] and efforts have been made to study its effectiveness in different fields like cross-language retrieval [9].

Use of WordNet for measuring semantic similarities between words has been proved to be effective [11]. Ray Richardson and Alan Smeaton experimented semantic distance measurement between concepts or words and using this word distance to compute similarity between a query and a document [10].

3. AUTOMATIC QUERY EXPANSION USING UMLS

3.1 Introduction to UMLS

The National Library of Medicine's Unified Medical Language System is an extensive vocabulary for medical literature. There are three UMLS Knowledge Sources, the UMLS Metathesaurus, the UMLS Semantic Network, and the UMLS SPECIALIST Lexicon, out of which only the UMLS Metathesaurus is being used by our system. The Metathesaurus contains semantic information about various biomedical concepts, their semantic types, their synonyms and relationships among them. It contains information from 98 different vocabularies arranged by meaning into concepts and a total 2,046,022 strings in 15 languages [14].

3.2 Process of query expansion.

3.2.1 Adding the query synonyms to the query

We expand the query using synonyms of the query term obtained using UMLS Metathesaurus. Specifically, we *OR* the original query term with its synonyms obtained from UMLS. But simply *OR*ing the query term with its synonyms, will only increase the recall and may decrease the precision of top 10 results significantly. Moreover, Google follows PageRank algorithm to rank the documents, and PageRank ranks the documents depending on the number of inlinks of the document which is independent of the query [8]. For example, for the query *fever*, only one of the top ten results displayed by Google is relevant. The query expanded using synonyms is (*fever OR "increased body temperature" OR hyperthermia OR pyrexia OR "temperature elevation"*). The expanded query does not give better performance than original query and the top most results remain more or less the same as these top result documents have higher Page Rank than the documents which contain only *increased body temperature* or only *hyperthermia*. So, generally, only *OR*ing the query term with its synonyms will not have a significant impact on the precision of top 10 search results.

3.2.2 *AND*ing the expanded query with the semantic type of original query term

UMLS Metathesaurus also provides semantic type for every medical term. For example, the semantic type of *fever* is *Sign or Symptom*. If we *AND* the semantic type of the query term with the expanded query, precision of the top 10 search results increases in many cases. The reason being that since the semantic type like *disease or syndrome* or *sign or symptom* gives a better description of the query term, it is effective in filtering out documents not related to medical domain. However, by doing this we might narrow our search space to a set of documents which do not include some very relevant documents that would otherwise be retrieved by Google.

Sometimes, the semantic type returned by UMLS is too specific, and affects the system performance. For example, the semantic type of the term *maxillary* is *body part, organ or organ component*. The system gives poor performance for such semantic type, as many documents that contain the word *maxillary* and are very relevant to the query do not contain the

semantic type of *maxillary* and hence are filtered out.

4. AUTOMATIC QUERY EXPANSION USING PSEUDO FEEDBACK RETRIEVAL

Automatic query expansion via relevance feedback is an effective technique commonly used to add relevant words to the query. However, it is difficult to obtain relevance feedback by manual evaluation by casual user [7]. Hence for such users, a most commonly used approach is *ad-hoc* or *blind* feedback that takes the form of *pseudo* feedback where actual input from the user is not required.

A small set of documents retrieved in initial phase of searching is *assumed* to be relevant without any intervention by the user. These assumed set of documents are further used in relevance feedback process, which applies algorithms like the Rocchio Method [2] to construct an expanded query which is then used to retrieve a refined set of documents. An obvious drawback of this method is that it assumes that the initial set of documents retrieved are relevant which may not always be true, in which case words added to the query are more likely to be irrelevant and the quality of documents retrieved in second phase is likely to be poor.

Our system implements relevance feedback for the initial set of documents retrieved using Google. For top 50 results given by Google for the user submitted query, we extract cached pages of the documents, index the documents using *Jakarta Lucene*, then extract the most frequently occurring terms from these documents to expand the original query, and finally *AND* this expanded query with the semantic type of the original query term.

We initially tried expanding the query by adding to it only the frequently occurring terms from the set of documents retrieved. The major drawback in this approach was that our system focuses on Medical domain, and top 10 documents retrieved by Google for short, one word medical queries are not always related directly to the query. For example consider the query *acidity*. The top 3 documents retrieved by Google for this query are (1) Soil acidity and Liming, (2) The acidity of Forest Soils and (3) The acidity of Wine, which are not at all related to the *medical* term *acidity*. Hence when relevance feedback was applied to the given query, the resultant expanded query was (*disease*) *AND* (*acidity OR has OR for OR many*).

We experimented different methods for selecting the terms to be added to the query arranging the terms in descending order of (1) *raw term Frequency (tf)*, or (2) product of *term frequency* with *inverse document frequency* of the term ($tf*idf$), but the results obtained were more or less similar. The basic reason was that since the initial set of documents were so irrelevant and unrelated to each other that except for common English terms (stop words) most other terms had a document frequency of one. In addition the run time efficiency of the system was very poor. An important observation in this case was: to get better performance, it is necessary to check whether the term being added to the query is relevant with respect to medical domain or not. Hence, we enhanced our system, by checking for every term if it is medical or not and adding only medical terms to the expanded query. We started checking whether the terms is present in UMLS database to determine its relevance with medical domain. One problem in this approach was that UMLS is a very extensive database. Hence common English terms like *has, a, color, is* are already present and have some medical meaning in UMLS. So the filtering process did not yield better results, on the other hand run time efficiency of system reduced down further by a very large extent because for every indexed term, the system was connecting to UMLS server to retrieve the concept type of the term.

The approach of using UMLS to check relevance of a term for medical domain was discarded because of its poor performance and huge time requirement. Instead, a new approach was experimented²: for each “potential” term to be added to the expanded query, we calculate its relevance ratio. The relevance ratio of a term is defined as follows:

$$\text{relevance ratio} = (\text{number of page hits for the term ANDed with the word medical}) / (\text{number of page hits for given term})$$

The basic assumption was that count of number of page hits obtained when the term is ANDed with word *medical* will give a more effective measure of relevance of a term to medical domain. This count is normalized by total number of page hits for the given term. We experimented this approach for a set of about 30 queries and observed that a ratio value of about 0.3 is good indicator that the term is related to medical domain.

² This approach was suggested by Dr. Hinrich Schutze

Although this new approach considerably improved the relevance of terms being added to the expanded query, time efficiency of the system was still as poor as the previous approach, due to significant amount of time required to obtain the count of number of page hits for (1) the individual term, and (2) the term ANDed with word *medical*, for a number of terms in the index.

To improve the system performance, we then refined the system by using an extensive stoplist, a list of most commonly occurring English terms, while building the index. We test run the system in learning mode over a set of 30 queries, where the system would check if the relevance ratio of a given term is greater that or equal to 0.3, if not then the term would be added to the stoplist. Thus the system dynamically built the stoplist. After performing several reruns, the stoplist built was sufficiently extensive. This approach considerably improved the system time requirements, since it reduced the number of terms being checked for whether they are medical terms. As a common heuristic applied for all the three basic search techniques presented in this paper, we finally ANDed the expanded query so obtained with its semantic type retrieved using UMLS. This further improved the set of documents being retrieved in the second phase.

The improvement in search performance, for certain queries, is evident from the fact that when we run the query *acidity* on the improved system, the expanded query is: *disease AND (acidity OR acids OR stomach OR neutralization)* which gives significantly better results than the initial query.

However, one significant drawback that still remains is that assigning weights to the terms added to query has significant impact on final results obtained and it is not possible to assign weights to query terms using Google. Hence, we can not effectively implement algorithms like Rocchio's algorithm for pseudo feedback retrieval since it gives emphasis on assigning weights to query and document vectors. Further, since all terms in the expanded query are given equal weight, a particular word that is a medical term but not related to the query when gets added to the query, influences the entire search. For example, the expanded query for the query term *glycosuria* is (*disease*) AND (*glycosuria OR renal OR urine OR medical*). The word *medical* gets added to the query and influences the search results so much that effectively the query is executed as (*disease*) AND (*medical*), and none of the top 10 results contain the

word *glycosuria*. Similar is the case for many other queries, where a very common term like *medical* or *infection* gets added to the query and influences the search results and none of the top 10 retrieved documents are relevant. Hence, although pseudo feedback retrieval performs well for a few queries, the average performance of pseudo feedback retrieval is not at all encouraging.

4. AUTOMATIC QUERY EXPANSION USING WORDNET

Princeton WordNet is an online lexical database whose design is inspired by current psycholinguistic theories of human lexical memory. In WordNet, english nouns, verbs, adjectives and adverbs are organized into synonym sets, each representing one underlying lexical concept.

A word (or term) sense is represented by a group of terms that are synonymous under this sense. Such a group is called a *synset*. A given word (or term) may be comprised in several synsets. The synonymy relation is implicit within each synset. Other relations are established among synsets.

Our system makes use of different senses of given query word retrieved using WordNet to expand the given query. The query so expanded is finally ANDed with semantic type of the term retrieved using UMLS to derive final query. The query expansion process using WordNet is exactly similar to that using UMLS except that the synonyms are retrieved using WordNet database.

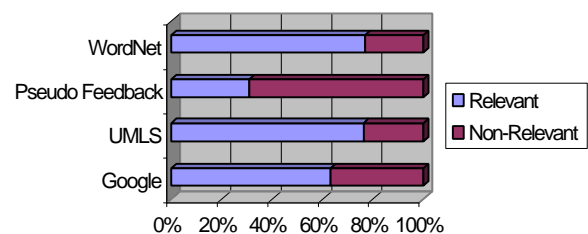
An important observation while using WordNet was that there are many medical terms for which no synonyms are retrieved using WordNet, but when synonyms are retrieved using WordNet, they represent the exact sense of given query term as against UMLS which gives a number of meanings of the given term, some which are more specific for medical domain but maybe irrelevant for normal user and hence may cause the focus on the query to be shifted thus giving poor performance.

For example, for query *dysentery*, WordNet does not give any synonyms, whereas for query *hypertension* the synonym retrieved using WordNet is *high blood pressure* which when ANDed with semantic type of *hypertension* yields good results.

5. COMPARISON OF RESULTS

The system was evaluated by four medical experts and three non-medical persons, by performing test runs over a set of 30 queries and comparing the top 10 results with Google. They were asked to rank these documents as *relevant* or *non-relevant* from the point of view of lay-person. Thus, for every query expansion technique, a set of 300 documents was examined. Following graph shows the comparison of results obtained.

Performance of Automatic Query Expansion Techniques



6. LIMITATIONS

- The system uses binary value of relevance/non relevance judgment while evaluating documents. There is not way of determining whether a document is more relevant than other.
- The system performance is not very effective for multiword queries since such queries are treated as potential phrase queries by Google and our system does not apply any heuristics to improve search for phrase queries.
- For pseudo feedback retrieval, the system assumes that the top 10 documents retrieved using Google are relevant which in many cases is not true.
- For pseudo feedback, the performance is not effective since it is not possible to assign weights to query terms in Google.
- Google has limitation of maximum 10 terms per query.

7. FUTURE WORK

- To apply a better heuristic than semantic type to refine the set of documents retrieved.

- To apply some heuristics for post processing of documents, refining the initial set of documents so that they are related to medical domain in case of pseudo feedback.
- To develop an efficient algorithm for multiword queries.

8. CONCLUSION

For web search for medical domain using Google for one word queries, automatic query expansion using UMLS and WordNet significantly improve the precision of top 10 results than those obtained using Google. However performance is poor in case of pseudo feedback retrieval since not all top k documents are relevant and it is not possible to assign weights to query terms using Google.

Also, *AND*ing the expanded query with its semantic type significantly improves the precision of topmost results. But this heuristic does not work in case of multiword queries.

9. ACKNOWLEDGEMENTS

We are thankful to Dr. Hinrich Schuetze, Dr. Chris Manning and Dr. Prabhakar Raghavan for their valuable guidance and support throughout the project, and Dr. Lawrence Fagan for his guidance regarding UMLS. We are also thankful to our evaluators Dr. Kambiz Merati, Dr. Neel Joshi, Dr. Aruna Khare, Dr. Ravindra Panse, Omkar Deshpande, Nipun Mehra and Paritosh Ambekar.

9. REFERENCES

- [1] Aronson , A. , Rindfleisch, T.
Query Expansion Using the UMLS Metathesaurus.
- [2] Buckley, C.
New Retrieval Approaches Using SMART: TREC 4.
- [3] Buckley, C., Allan,J., Salton, G.
Automatic Routing and Ad-hoc Retrieval Using SMART : TREC 2 (1994)
- [4] Buckley, C., Salton, G., Allan, J., & Singhal, A.
(1995) Automatic query expansion using SMART: TREC 3.
- [5] Cui.Hang, Wen.Ji-Rong, Nie.Jian-Yun , Wei-Ying Ma
Probabilistic Query Expansion Using Query Logs
- [6] Hisao Imai, Nigel Collier, Jun'ichi Tsujii
A Combined Query Expansion Approach for Information Retrieval
- [7] Mitra , M., Singhal , A. , Buckley, C.
Improving Automatic Query Expansion.
- [8] Page, L., Brin S., Motwani , R., Winograd ,T.
The PageRank Citation Ranking: Bringing Order to the Web (1998).
- [9] Qu Yan, Alla N. Eilerman, Hongming Jin, David A. Evans
The Effect of Pseudo Relevance Feedback on MT-Based CLIR
- [10] Richardson, R., Sineaton, A.,
Using WordNet in a Knowledge-Based Approach to Information Retrieval
- [11] Richardson , R., Smeaton , A., Murphy , J.
Using WordNet as a Knowledge Base for Measuring Semantic Similarity between Words (1994) .
- [12] Srinivasan , P.
Query expansion and MEDLINE.
- [13] W. Hersh and S. Price and L. Donohoe
Assessing Thesaurus-Based Query Expansion Using the UMLS Metathesaurus.
- [14] William T. Hole, Director UMLS R&D
Personal communication.

Tools/APIs used: Jakarta Lucene, Xerces XML parser, UMLS API, WordNet API, Google API.

Demo URL:

<http://www.stanford.edu/~vkakade/MedSem.html>

Note: Since there is a limit of 10 seconds for Stanford CGI account, pseudo feedback retrieval can give the error 'CPU time exceeded' while running from web UI. If you get such an error, please use the command line UI, using `~vkakade/public/cgi-bin/runPFB.sh` with command line parameter as the query term (multiple query terms can be separated by '+'). Also the program takes long time to run. The output of the program is an HTML document which can be redirected and when viewed in a browser will display the search results for expanded query. (Note: this program has to be run from `tree1` or `tree2`, otherwise UMLS will give an error 'Invalid client IP address')