MedSearch: A Specialized Search Engine for Medical Information

Gang Luo[1] Chunqiang Tang[1] Hao Yang[1] Xing Wei[2]
IBM T.J. Watson Research Center[1] University of Massachusetts – Amherst[2]
{luog, ctang, haoyang}@us.ibm.com xwei@cs.umass.edu

ABSTRACT

People are thirsty for medical information. Existing Web search engines cannot handle medical search well because they do not consider its special requirements. Often a medical information searcher is uncertain about his exact questions and unfamiliar with medical terminology. Therefore, he prefers to pose long queries, describing his symptoms and situation in plain English, and receive comprehensive, relevant information from search results. This paper presents MedSearch, a specialized medical Web search engine, to address these challenges. MedSearch can assist ordinary Internet users to search for medical information, by accepting queries of extended length, providing diversified search results, and suggesting related medical phrases. A full version of this paper is available in [1].

Categories and Subject Descriptors
H.3.3 [Information Search and Retrieval]: search process
General Terms: Algorithms, Experimentation
Keywords: medical query, medical Web search engine

1. INTRODUCTION

Health care is a major business in many countries and a large part of this business is related to the management and retrieval of medical information. To facilitate people to acquire medical information in the Web era, many medical Web search engines (e.g., Healthline and Google Health) have come into existence. While these systems have their own merits, they all treat medical search in much the same way as traditional web search.

Medical search has several unique requirements that distinguish itself from traditional Web search. A common scenario in which a person performs medical search is that he feels uncomfortable but is uncertain about his exact medical problems. In this case, the searcher usually prefers to learn all kinds of knowledge that is related to his situation. However, existing medical Web search engines are optimized for precision and concentrate their search results on a few topics. This lack-of-diversity problem is aggravated by the nature of medical web pages. When discussing a medical topic, many medical web sites use similar, but not identical, descriptions by paraphrasing contents in medical textbooks and research papers. Hence, search results provided by existing medical Web search engines often contain much semantic redundancy, which cannot be easily handled by existing methods for identifying near-duplicate documents or result diversification. To find useful medical information, the searcher often has to go through a large number of Web pages laboriously.

Another unique feature of medical search is due to the fact that most Internet users do not have much medical knowledge. A medical information searcher is often unclear about the problem that he is facing and unaware of the related medical terminology (e.g., panophthalmitis). As a result, it is difficult for him to choose a few accurate medical phrases as a starting point for his search. Instead, considering the importance of his health, the searcher is typically willing to take his time to describe his situation in detail (e.g., his medical history, where and how he feels uncomfortable, and what happened in the last several days) by posing long queries in plain English, much like the way he talks to a doctor. Actually, many medical questions that people posted on medical forums contain several hundred words, and a recent study on medical queries [2] has reported that medical information searchers prefer to pose detailed long questions to Web search engines. Figure 1 shows one typical example of such query.

> www.medhelp.org/forums/RespiratoryDisorders/messages/2584.html
> ... My 23 month old son has been coughing since 6 months old … Seems to be constantly on antibiotics for every kind of chest infection, on pulmicort, albuterol 2x's a day, constant ear infections (tubes, adnoids, and tonsils are scheduled), chronic loose stools. Seen an allergist, he has lots of environmental allergies, did all the mattres covers, rugs are gone, air purifier in. All this to no avail. Chest xray showed streaking in the main bronch tubes (?) perihilar stuff hazy areas, left lobe is alot grayer than the right. … Went to pedi pulmonologist in Boston, scheduled for sweat test on Friday, he doesnt think he has it, but wants to rule out CF. He wants to do CT and bronchoscope next week. Mentioned something about poss. deformed broch tubes, or weak lung walls, or even a cyst compressing his lungs causing this cough … what are the possibilities he has a verison of pulmonary micobacterial infection? ...

Figure 1. An exemplary medical question posted on the Med Help International Medical and Health Forum (www.medhelp.org/forums.htm).

Even after stopword removal, the above query still cannot be fed directly into existing medical Web search engines, because they all impose certain limits on query length for various reasons. For instance, Google truncates long queries into the length limit of 32 words. Such a low limit on query length is a serious obstacle for medical information searchers. Moreover, a medical information searcher often prefers the search engine to automatically suggest diversified, related medical phrases that can help him quickly digest search results and refine his query. However, this cannot be done with existing medical Web search engines when the query is written using plain English description and has a terminological discrepancy from medical phrases.

2. MEDSEARCH

In this paper, we present MedSearch, a prototype medical Web search engine that addresses the aforementioned limitations of existing systems. MedSearch uses several key techniques that greatly improve its usability and the quality of search results. First, MedSearch accepts queries of extended length and supports the use of plain English description. This is a great convenience for the majority of Internet users who do not have much medical knowledge. MedSearch automatically rewrites long queries into moderate-length queries by selectively dropping unimportant terms (i.e., words). Since unimportant terms not only appear in a large number of Web pages but also obscure the main theme of the query,

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

dropping them can both greatly increase the query processing speed and improve the quality of search results. Second, MedSearch returns diversified Web pages without significantly increasing query processing time or deteriorating the quality of the returned top Web pages, which allows the searcher to see various aspects related to his situation. Third, MedSearch automatically suggests diversified, related medical phrases to the searcher based on information from several sources: the standard MeSH medical ontology (www.nlm.nih.gov/mesh/meshhome.html), the collection of crawled Web pages, and the query itself.

There are several key challenges in designing MedSearch. In order to rewrite long queries into moderate-length queries, we must aggressively drop unimportant terms yet avoid losing much useful information. For this purpose, we rank all the terms in the query according to the Okapi term weighting formula. Those terms with small weights are treated as unimportant ones and dropped from the query.

One major challenge in providing diversified search results is to efficiently handle the excessive redundancy among different medical Web pages. For this purpose, all the crawled Web pages are clustered into multiple clusters in a pre-processing step. Each of these clusters roughly corresponds to a different topic. When ranking Web pages, each cluster can contribute only a limited number of results to the returned top few Web pages. Then the searcher is likely to see different aspects in the top results.

The process of suggesting related medical phrases consists of two sub-steps. The first sub-step is to generate the candidate set *S* of related medical phrases in the MeSH ontology. The second sub-step is to rank the medical phrases in *S*. In the first sub-step, MedSearch selects *V=60* medical phrases from the returned top-*20* Web pages. The suggested medical phrases need to be both relevant and diverse in order to provide the greatest convenience to the searcher. Intuitively, to ensure that a medical phrase *M* is relevant, it is better for *M* to appear in one of the returned top Web pages with a large tf×idf value that is computed using the Okapi formula. To ensure enough diversity in the list of suggested medical phrases, a single Web page should not contribute too many medical phrases to that list. We use a continuous discounting method to achieve these two goals. Each time a medical phrase is chosen from a Web page *P*, a discount is given to the tf×idf values of the remaining medical phrases in *P*. As a result, the more medical phrases have already come out from *P*, the more difficult the remaining medical phrases in *P* will come out in the future. We select *V* medical phrases in *V* passes. In each pass, we select a medical phrase with the largest tf×idf value.

The main challenge in the second sub-step of ranking the suggested medical phrases is to resolve the terminological discrepancy between medical phrases and queries written in plain English. For this purpose, a set of representative Web pages are computed offline for each medical phrase *M*, by using *M* to retrieve the top-ranked Web pages. Since a large part of these high-quality representative Web pages are written in plain English, they provide good linkages between medical terminology and plain English words. The relevance between a query *Q* and a medical phrase *M* is computed as a function of the relevance scores between *Q* and *M*'s representative Web pages. Then all the suggested medical phrases are sorted in descending order of their relevance scores. A detailed description of our techniques is available in [1].

3. RESULTS

To demonstrate the effectiveness of our techniques, we conducted experiments using 30 representative medical questions that people posted on a popular medical forum, the Med Help International Medical and Health Forum (www.medhelp.org/forums.htm). One such query is shown in Figure 1. We crawled 6GB of Web pages from WebMD (www.webmd.com), one of the most popular medical web sites.

Both relevance and diversity are judged using a single metric: *usefulness*. A returned Web page *P* is useful if *P* is relevant to the query, and much of *P*'s relevant content has not been mentioned in the Web pages that are ranked higher. If *P* is useful, its usefulness score $score_u(P) = 1$; otherwise, $score_u(P) = 0$. A similar definition of usefulness holds for the suggested medical phrases. For the returned top-20 Web pages P_i ($1 \le i \le 20$), their weighted average usefulness score is defined as

$$avg\_score_u = \sum_{i=1}^{20} score_u(P_i) / \log(1+i).$$

For the suggested 60 medical phrases, their weighted average usefulness score is defined similarly. The mean of the weighted average usefulness scores over the 30 queries is the main quality metric for the returned pages and the suggested phrases. Five colleagues served as assessors and independently determined the usefulness scores of the returned Web pages and the suggested medical phrases. None of them has formal medical training.

To give the reader a feeling of the contents returned by MedSearch, we present detailed results of the returned Web pages and the suggested medical phrases for the query in Figure 1. Table 1 shows some of the returned relevant Web pages. The suggested relevant medical phrases include bronchoscopy (rank 1), bronchitis (rank 2), and sarcoidosis (rank 4). In general, for a medical query *Q*, MedSearch can find several relevant Web pages and medical phrases that cover multiple aspects of *Q*.

Table 1. Some of the returned relevant Web pages.

rank	URL	topic
1	www.webmd.com/content/chat_transcripts/1/108027.htm?printing=true	asthma
3	www.webmd.com/hw/ear_disorders/hw184529.asp@printing=true	ear infection
4	www.webmd.com/content/chat_transcripts/1/107597.htm?printing=true	spring allergies

The means of the weighted average usefulness scores over the 30 queries for the returned top-20 Web pages and the suggested 60 medical phrases are 7.9 and 6.1, respectively. We present a simple calculation below to give the reader some intuition on these numbers. Let ws_i denote the weighted average usefulness score when the returned top-*i* Web pages (or medical phrases) are useful while the others are not useful. Then $ws_1 = 3.3$, $ws_2 = 5.4$, $ws_3 = 7.1$, and $ws_4 = 8.5$.

Our results show that MedSearch can process long queries efficiently, at a speed roughly comparable to that of existing medical Web search engines in processing short queries. Our experiments also show that users' satisfaction is crucially tied to MedSearch's capability of returning diversified Web pages and suggesting diversified, related medical phrases that can help users quickly understand the returned pages and refine their queries. Detailed results are available in [1].

4. REFERENCES

[1] Full version of this paper is available at http://www.cs.wisc.edu/~gangluo/medsearch.pdf.

[2] A. Spink, Y. Yang, and J. Jansen et al. A Study of Medical and Health Queries to Web Search Engines. Health Information and Libraries Journal 21(1): 44–51, 2004.