MMSE: Design & Implementation of a Medical Meta Search Engine

Ali Rezaeian Joojadeh and Hamid Hassanpour

***Abstract*— We present a new next generation domain of meta search engine. A Medical Meta Search Engine (MMSE) was designed in this research for the users with no medical expertise. It is enhanced with the domain knowledge obtained from Unified Medical Language System (UMLS) to increase the effectiveness of the searches. The power of the system is based on the ability to understand the semantics of web pages and the user queries. This MMSE transforms a keyword search into a conceptual search.**

This medical meta search engine aims to generate maximum output with semantic value using minimum input from the user. Since this system is designed to help people seeking information about health on the web, our target users are not medical specialists who can effectively use the special jargon of medicine and access medical databases. Medical experts have the advantage of shrinking the answer set by expressing several terms using medical terminology. Medical meta search engine provides the same advantage to its users through the automated use of the medical domain knowledge in the background. The results of our experiments indicate that, expanding the queries with domain knowledge and increase dramatically the relevance of an answer set and the number of retrieved web pages that are relevant to the user request.

***Index Terms*—Medical information retrieval, UMLS, Medical, Meta search engine, Conceptual search.**

I. Introduction

Internet is a vast source of information from different domains. Among these different domains, health and medical information is one of the leading domains with huge amount of information. For example, MEDLINE reports containing approximately 13 million references to biomedicine articles. Having such massive amount of information at disposal, people become more conscious about their health problems. Being more aware of their health, patients wish to be armed with knowledge about their medical problems. Nowadays, patients tend to prefer to be more informed before, during, and after consulting to their physicians [1]. As opposed to the classical patient behavior, these patients wish to take part in the decision process for their health problems. For this purpose, such patients search the Internet, which is the easiest and fastest way to access information, and try to dig information out. This situation leads to a growing number of medical inquires on the Internet by the people with no medical training [1]. Therefore, there is a growing need for a medical information retrieval techniques and tools to help ordinary people in searching health information on the Internet.

Manuscript received April 18, 2012; revised June 13, 2012.

Ali Rezaeian Joojadeh is with the Sama technical and vocatinal training college, Islamic Azad University, Sari Branch, Sari, Iran (e-mail: Alirezaeian_j@yahoo.com).

II. Related Works

Currently, there are four groups of gateways available to a medical information seeker:

- Specialized guides in medicine
- Medical databases and search engines on these databases
- General purpose search engines
- Medical-domain web search engines

Specialized guides in medicine are a structured compilation of medical resources maintained by health care experts. Since these guides are maintained manually by experts, the contents of the documents (web pages) are reliable, accurate and relevant to the topics. On the other hand, since the evaluation of documents is a slow process, the maintenance of these guides cannot cope with the growth rate of health information on the Internet [2]. An extension to medical guides is a subject directory where the users add documents to the hierarchy to achieve amore extensive coverage. As the health information grows, users fail to select the correct category to add and the search results become unpredictable.

Medical databases and search engines working on these collections are built for the professionals armed with the terminology of medicine. To use such medical databases and collections for obtaining a satisfying answer set, professionals form a query detailed with medical terminology. Such queries are usually impossible for an ordinary person to construct. In addition, the answer set returned as a response for such complicated queries would have documents with heavy medical terminology that is hard for an ordinary person to understand. Therefore, medical databases are not suitable for an ordinary person to use [2].

An alternative to the use of the medical literature database search is the use of general purpose search engines. These engines have the ability to deal with exponential growth rate of Internet through the use of a robot called crawler. These robots periodically navigate the web to automatically discover and retrieve documents to be indexed for later use. Although these tools maximize the number of documents retrieved and respond to their users quite fast, obtaining the domain specific information, such as the health domain, is difficult. The information seeker is often presented an answer set with a large number of entries most of which are irrelevant to her interest. This is a classical example of low precision, which is the ability of retrieving only the related documents [3]. These search engines produce an answer set with

relatively low precision in the medical domain since these tools aim to serve searches without a domain restriction.

Having considered the different advantages of the above two information retrieval solutions, researches have been directed towards hybrid approaches. Medical search engines, which retrieve health information available on the Internet, combine the advantages of two approaches. The high precision advantage of specialized guides and subject directories suggests that we should restrict the search space to the medical domains. The high recall advantage of general search engines, which is the high proportion of relevant documents retrieved out of all relevant documents available, suggests that we should use special crawlers to discover and maintain a comprehensive collection. Medical search engines aim to improve the precision and recall through different techniques such as restricting their crawlers to harvest only medical documents, extending or modifying user queries, suggesting their users to select a search category [3], [4].

Meta-search engines (MSE) also known as multi-threaded engines, do not necessarily maintain their own listings/databases, but send the user's query simultaneously to other search engines, Web directories or to deep Web, collect the results, remove the duplicate links, merge and rank them according to their own algorithm in a single list and display it to the user. The structure of the Meta Search Engine is shown in "Fig. 1". Meta-search engines provide fast and easy access to the desired search, because they can search from multiple search engines simultaneously and save the precious time of the searcher. Meta-search engines provide a broader overview of a topic as compared to traditional search engines and increase the coverage of Web by combining the coverage of multiple search engines. Querying multiple search engines is more scalable then the centralized general purpose search engine [5].

Our context-based meta-search engine is based on the combination of context to normal search results from a search engine such as Google. By taking advantage of Google's Web Services and existing directory structure, it provides more accurate and relevant results as well as give users ease in the query formulations. Our goal is to design a medical meta-search engine that will help ease and guide the searching efforts of novice web users towards their desired objectives.

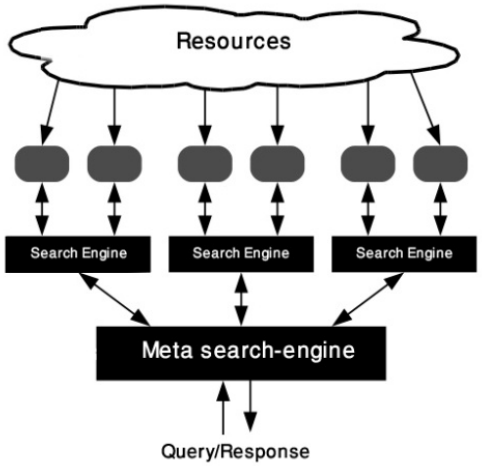

Fig. 1. The structure of the meta search engine

III. Using UMLS Library

Our approach is empowered with the Unified Medical Language System. UMLS is a knowledge repository of medical terms and their relationships. UMLS has three knowledge components:

- SPECIALIST Lexicon
- Semantic Network
- Metathesaurus

The SPECIALIST Lexicon is a collection of biomedical terminology and medical terms [6]. This component also offers Lexical Tools that help users to manage variation in biomedical terms [7]. The UMLS Semantic Network involves in the classification of Metathesaurus concepts. This component provides a consistent categorization of the concepts and defines a set of semantic relationships among the terms in SPECIALIST Lexicon [8], [9]. Medical MSE uses UMLS for different purposes. We use the lexicon of UMLS for indexing and generating concepts for the crawler.We use the Metathesaurus and Semantic Network for extending user queries and formulating query terms as well as ranking the query results and filtering irrelevant documents [10], [11].

IV. Medical Meta Search Engine System

The structure of Medical Meta Search Engine is shown in "Fig. 2". System got a simple query from end user and then sent to the UMLS Expand Query portion to get the special medical expression. The output was a high level medical concept. This Conceptual query was sent to the single search engines -Google, Yahoo, Bing- and got the relevant results of this query from each search engines.

Conceptual query and results of search engines were sent to Re-ranking portion. In this unit retrieved documents from different search engines were compared with the high level medical concepts and re-ranked from most related to least related one. Finally medical results with desired clustering were presented to end user.

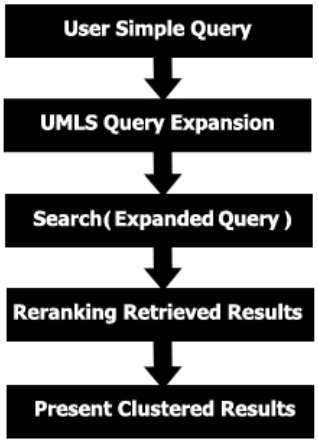

Fig. 2. The structure of the medical meta search engine

A. Concept Extractor

In the Second stage of system (UMLS Query Expansion) the input query was traversed at first and the keywords of query were extracted, then these keywords were sent to

UMLS Find Concepts unit that medical concepts in all levels were extracted and finally the longest and richest concepts were selected.

The significance of the conception defined by Is Significant function means that whether the length of the investigative conception includes noticeable contribution of all conceptions in initial query or not. The reason is the sensitivity of the query expansion process, because if an improper word is added to a query, it can impact on the retrieval result negatively, so we should be careful about the selection of conceptions in query expansion. "Fig. 3" shows these steps.

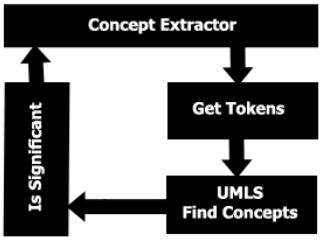

Fig. 3. Converting keyword search to coceptual search

When the user submits a query text as a search phrase or keywords through the system's user interface, the text is processed by the Concept Extractor to extract the terms. During this extractions of terms,

- The search phrase specified by the user is skimmed in order to discover whether there are any UMLS SPECIALIST Lexicon terms and phrases or not,
- If any UMLS SPECIALIST Lexicon terms or phrases exists, they are kept as they appear in the search text,
- The keywords that are not lexicon terms are stemmed and the duplications, if any, are removed.

With respect to these extracted user query terms, the Concept Extractor module runs terminology search, context search and relations search queries on the UMLS. Details of this unit are described in next section.

B. Get Keywords Concepts

Each query includes N keywords; some of these keywords can combine with each other to make an expression, so by these combinations, 2^N-1 expressions can be formed. The best and longest expression consists of all N keywords and the weakest one is the one-word expression.

In this part all of the combinations made by keywords were obtained by Get Keyword Combination function, and then medical expressions by different length were extracted by Concept Candidate Extractor function. In other word, the outputs of this section are high level expressions of primary query that include the medical concepts. The structure of the UMLS Find Concepts is shown in "Fig. 4".

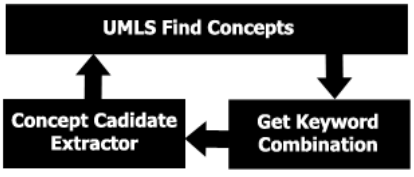

Fig. 4. Get keywords concepts with UMLS database

C. Re-Ranking Retrieved Results

In this stage we were sure that all of the extracted concepts retrieved from previous stage are purified and there's not any unused expression. These expressions were sent to the UMLS Find Concept Definitions unit that uses UMLS Semantic Network to get full and comprehensive definition of query. For example, the concept set generated for the query term "breast cancer" is shown in Table I [12].

TABLE I: CONCEPT RELATION LIST FOR QUERY TERM "BREAST CANCER"

Concept list obtained from UMLS	Relation Type
Breast carcinoma	Synonym
Cancer of the breast	Synonym
Mammary carcinoma	Synonym
Carcinoma of breast	Synonym
Malignant neoplasm of breast	Partial Relevance
Malignant tumor of breast	Contextual Relevance

Finally as shown in "Fig. 5" complete and organized definitions of user query gotten by UMLS database are compared with retrieval documents by different search engines using Apache Lucene library that proper and more relevant results to refined query were retrieved and ranked.

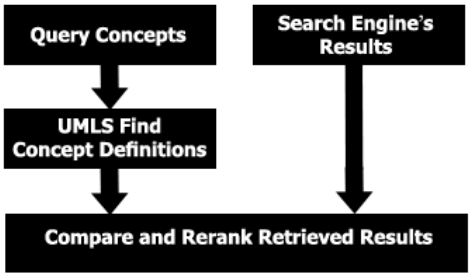

Fig. 5. Rarank results of different search engines

Lucene is an open source, highly scalable text search-engine library available from the Apache Software Foundation. Steps in building applications using Apache Lucene are shown in "Fig. 6". You can use Lucene in commercial and open source applications. Lucene's powerful APIs focus mainly on text indexing and searching. It can be used to build search capabilities for applications such as e-mail clients, mailing lists, Web searches, database search, etc. Web sites like Wikipedia, TheServerSide, jGuru, and LinkedIn have been powered by Lucene. Lucene has many features [13] , [14]. It:

- Has powerful, accurate, and efficient search algorithms.
- Calculates a score for each document that matches a given query and returns the most relevant documents ranked by the scores.
- Supports many powerful query types, such as PhraseQuery, WildcardQuery, RangeQuery, FuzzyQuery, BooleanQuery, and more.
- Supports parsing of human-entered rich query expressions.
- Allows users to extend the searching behavior using custom sorting, filtering, and query expression parsing.
- Uses a file-based locking mechanism to prevent concurrent index modifications.
- Allows searching and indexing simultaneously.

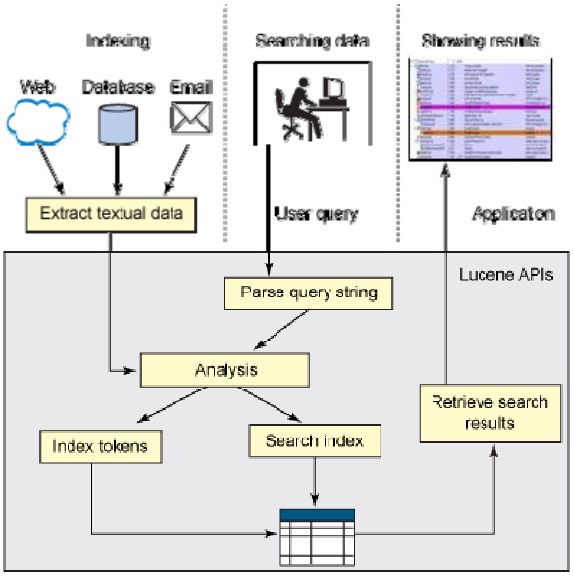

Fig. 6. Steps in building applications using Lucene[14]

V. System Evaluation

In information retrieval contexts, precision and recall are defined in terms of a set of retrieved documents (e.g. the list of documents produced by a web search engine for a query) and a set of relevant documents (e.g. the list of all documents on the internet that are relevant for a certain topic) [15].

Precision (P) is the fraction of retrieved documents that are relevant to the search.

$$P = \frac{|\{relevant\ \ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|}$$

Precision takes all retrieved documents into account, but it can also be evaluated at a given cut-off rank, considering only the top most results returned by the system. This measure is called precision at n or P@n.

For example for a text search on a set of documents precision is the number of correct results divided by the number of all returned results.

Precision is also used with recall, the percent of *all* relevant documents that is returned by the search. The two measures are sometimes used together to provide a single measurement for a system.

Recall in Information Retrieval is the fraction of the documents that are relevant to the query that are successfully retrieved [15].

$$R = \frac{|\{relevant\ \ documents\} \cap \{retrieved\ documents\}|}{|\{relevant\ documents\}|}$$

For example for text search on a set of documents recall is the number of correct results divided by the number of results that should have been returned.

It is trivial to achieve recall of 100% by returning all documents in response to any query [15]. Therefore recall alone is not enough but one needs to measure the number of non-relevant documents also, for example by computing the precision.

To evaluate the implemented system 10 queries were given to each of Yahoo, Bing and our designed meta search engine, and 100 retrieved results were judged, means that 300 judgments were done for each query.

A. Query Expansion Evaluation

Here, the effect of query expansion was investigated, as said before; the length of a query was determined by Significance parameter. As it is shown in "Fig. 7" increase in this parameter caused decrease in number of concepts in expanded query and decrease in this parameter caused increase in number of irrelative concepts to user requirement in query expansion leading to reduction of performance of information retrieval.

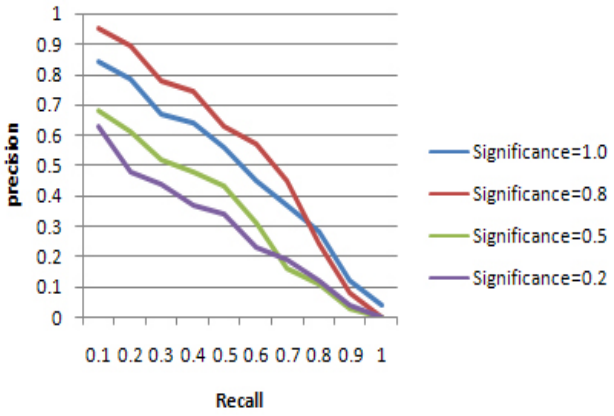

Fig. 7. Effects of Significance Parameter

B. Results Re-Ranking Evaluation

Here, performance of re-ranking of results was studied. "Fig. 8" demonstrates that our method has high ability in identification of related documents to user requirements and increase in their ranks, using implicit definition in query words. So the performance precision in high ranked documents linked to low recall is much more than other search engines. It's obvious that improving the rank of more relevant documents leads to decrease in precision of them. Precision of our proposed method for lower ranked documents (higher recalls) is less than other engines and vice versa.

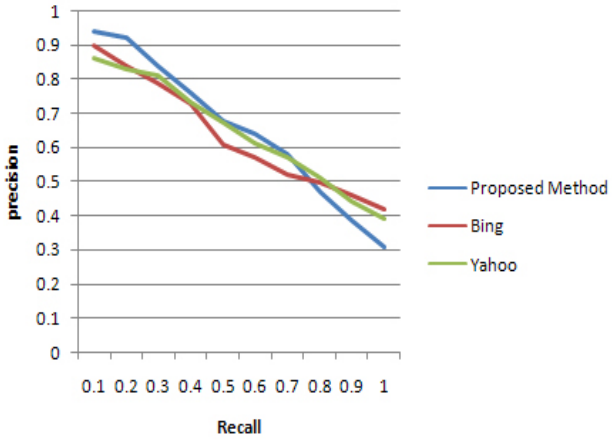

Fig . 8. Precision at Recalls for average 10 queries

VI. Conclusion

In this paper, we have proposed a medical meta search engine for anyone who seeks health information on the Internet. The uniqueness of our meta search engine lies in its

ability to effectively provide relevant and understandable medical information to the information seeker with no medical training. Our experimental results are promising. Currently, medical meta search engine stands as a powerful alternative for health information seekers with no medical expertise.

Although our system is developed as a medical domain search engine, it can be transformed into a powerful domain search engine that could serve for any domain. Due to its modular structure, UMLS can be replaced with any other knowledge source to make this system a meta search engine for another domain.

REFERENCES

[1] G. Eysenbach and C. Koehler, "How do consumers search for and appraise health information on the World Wide Web? Qualitative study using focus groups, usability tests, and in-depth interviews," *BMJ 324* (2008) 573‑577.

[2] Manoj M and Elizabeth Jacob, "Information retrieval on Internet using meta-search engines: A review*," Journal of Scientific & Industrial Research* vol. 67, October 2008, pp.739-746.

[3] Daniele Braga, Alessandro Campi, Stefano Ceri, and Alessandro Raffio, "Joining the results of heterogeneous search engines," Dipartimento di Elettronica e Informazione, Politecnico di Milano, via Ponzio 34/5, Milan, Italy.

[4] A. B. Can and N. Baykal, "MedicoPort: A medical search engine for all," *Computer methods and programs in biomedicine* 86 (2007) 73–86.

[5] Giansalvatore Mecca, Salvatore Raunich, and Alessandro Pappalardo, "A new algorithm for clustering search results," Dipartimento di Matematica e Informatica, Universita della Basilicata, viale dell'Ateneo Lucano, 10, 85100 Potenza, Italy.

[6] Unified medical language system. [Online]. Available: http://www.nlm.nih.gov/ research/umls/umlsdoc intro.html.

[7] Unified medical language system, specialist lexical tools. http://specialist.nlm.nih.gov/LexTools.html.

[8] O. Baujard, V. Baujard, S. Aurel, C. Boyerb, R.D. Appela, "Trends in medical information retrieval on Internet," Comput. Biol. Med. 28 (5) (2006) 589–601.

[9] National library of medicine (nlm), PubMed. [Online]. Available: http://www. ncbi.nlm.nih.gov/entrez/query.fcgi.

[10] J. M. Abasolo and M. Gomez, "Melisa an ontology-based agent for information retrieval in medicine," *in: Proceedings of the ECDL 2000 Workshop on the Semantic Web, Lisbon,* 2000.

[11] Y. Kagolovsky, D. Freese, M. Miller, T. Walrod, and J. Moehr, "Towards improved information retrieval from medical sources," *Int. J. Med. Inf.* 51 (1998) 181–195.

[12] C. Boyer, O. Baujard, V. Baujard, S. Aurel, M. Selby, and R.D. Appel, "Health on the net automated database of health and medical information," *Int. J. Med. Inf.* 47 (2008) 27–29.

[13] A. Gaudinat, P. Ruch, M. Joubert, P. Uziel, A. Strauss, M. Thonnet, R. Baud, S. Spahni, P. Weber, J. Bonal, C. Boyer, M. Fieschi, A. Geissbuhler, "Health search engine with e-document analysis for reliable search results," *Int. J. Med. Inf.* 75 (2006) 73–85.

[14] Daniel Naber and Mindquarry GmbH, "Apache Lucene:Searching the Web and Everything Else," *The international conference on java technonogy*, June 24-28, 2007, Zurich.

[15] Y. Zhou, J. Qin, and H. Chen, "CMedPort: an integrated approach to facilitating Chinese medical information seeking," *Decis. Support Syst.* 42 (3) (2006) 1431–1448.

Ali Rezaeian Joojadeh was born at sari, Mazandaran, Iran on 5th August 1984. He received the Bachelor's degree in Computer Engineering from Iran University of Science and Technology, Behshahr, Iran, in 2005, the Master's degree Computer Engineering from Islamic Azad University, Arak, Iran, in 2011.

His research interests include data mining, web mining, information retrieval algorithms designing and computer network security. He is working in Sama technical and vocatinal training college, Islamic Azad University, Sari Branch, Sari, Iran.

Hamid Hassanpour received the B.S. degree in computer engineering from Iran University of Science and Technology, Tehran, Iran, in 1993, the M.S. degree in computer engineering from Amirkabir University of Technology, Tehran, Iran, in 1996, and the Ph.D. from the Queensland University of Technology, Brisbane, Australia, in 2004.

His research interests include biomedical signal processing, time-frequency signal processing and analysis, new architectures in computer design, text syntax analyzing and image processing.