Searching for Information on the Internet Using the UMLS and Medical World Search

Humbert H. Suarez, M.D Ph.D., Xiaolong Hao, Ph.D., Ifay F. Chang, Ph.D.,
Polytechnic University, Hawthorne, NY

Medical World Search is a search engine for medical information on the Internet that distinguishes itself from other search engines by its built-in knowledge of medical terminology through its use of the National Library of Medicine's UMLS and its carefully selected but large database of medical sites. After discussing some of the previous uses of the UMLS for medical information retrieval, we describe the Medical World Search system. In October 1996, Medical World Search became operational on the World Wide Web at http://www.mwsearch.poly.edu. It has been operating uninterrupted since then. We review our experiences with creating a search engine for medical information on the Internet and using the UMLS in this application. The UMLS has some clear advantages in this application. Some aspects of the UMLS also decrease its usefulness in information retrieval. Medical World Search's usage by medical information seekers is summarized. Future directions for research are outlined.

INTRODUCTION

The World Wide Web is triggering a large growth in the amount of medical information that is potentially freely available to practicing physicians quickly. This readily available information could be important to help physicians maintain the level of their clinical skills, which tend to decline with time[1,2], and learn more quickly about advances in medical diagnosis and therapy, reducing the long delay that is often observed [3,4,5]. The result could be an improvement in clinical decision-making, which has been found lacking in a number of studies[6,7,8,9,10].

However, medical information is often difficult to find on the World Wide Web, compared to MEDLINE (standing in contrast with the potentially larger usefulness of the Web, given that it presents full-text information instead of just abstracts). In MEDLINE, abstracts have been indexed manually with MeSH terms[11], bypassing some of the complexity and richness of medical terminology. A given medical term often has a number of synonyms, as well as potentially a large number of more specific terms, and locating the most relevant document to answer a specific information need should include searching for these related terms. A variety of search systems allow to search MEDLINE records using Boolean operators, explosion to search for broader concepts by retrieving documents that contain narrower terms, and other features. For instance, Haynes et al.[12] listed 27 MEDLINE systems.

Another major problem while searching for medical information on the World Wide Web is the presence of much noise, resulting from the varying values of the published information. Search engines have usually taken one of two approaches. They may index the full text of a large number of sites and rely on the user to filter out noisy results. Altavista, Lycos, and Infoseek are examples of such systems. The major disadvantage is the large result set returned for many queries and the variable quality of these results. Another approach is to choose the major sites manually and search or browse over this limited selection, and instead of searching the full text, index manually generated descriptions of the information resources. An example is Yahoo. For both approaches, clinically useful information is usually lost in these search engines among the consumer-oriented, medically irrelevant, or variable quality information. This observation led to the creation of sites such as Medical Matrix, CliniWeb[13] and MedWeb[14], which take a similar approach to Yahoo, but focus on medical information.

Medical World Search is a system that allows searching the full text of a large collection of Web pages for medical information and retrieve results that are ranked by relevance to the user's query. For indexing and searching, it uses a word-statistical approach[15], similar to Altavista, Lycos, and Infoseek, and to Knowledge Finder, a MEDLINE system[16]. However, in contrast to the Internet search engines, Medical World Search selectively indexes medical documents, carefully choosing sites that are of high quality and high clinical content to reduce noise in the search results.

In addition, a major feature of Medical World Search is its use of the Unified Medical Language System

1091—8280/97/$5.00

(UMLS) to allow including synonymous terms in searches and doing explosions. The UMLS is a large project of the U.S. National Library of Medicine (NLM)[17]. The UMLS Metathesaurus contains a representation of many different biomedical vocabularies (such as MeSH, CPT, ICD-9-CM, SNOMED-2, SNOMED International, just to name the major components) that are unified so that a UMLS concept might list the equivalent strings (synonyms) in the various vocabularies. As such, the UMLS is a large source of medical terminology information that can be used for information retrieval.

A number of researchers have used the UMLS for information retrieval, by matching medical concepts in user queries and database documents so as to abstract away from the variability of medical language (normalization). The most comprehensive approach has been the SAPHIRE system. Two different concept-matching algorithms were used by SAPHIRE, an exact one[18,19] and an approximate system[20]. SAPHIRE has undergone substantial evaluation, in the AIDSLINE[21], MEDLINE databases[22], and a textbook, *Scientific American Medicine*[23]. These studies compared SAPHIRE with word-based natural language and word-based Boolean searching of the databases, and no differences in recall or precision were found among these systems.

Despite these inconclusive results, it is conceivable that an information retrieval system that allows close interaction with its users during the process of thesaurus term selection will perform substantially better; also, the concept matching algorithm can be improved substantially. Medical World Search was conceived under these premises.

METHODOLOGY

System Design

Medical World Search consists of five modules: a Web crawler, a normalizer, a statistical information retrieval system, and a user interface.

Web Crawler. The Web crawler uses a combination of automated retrieval of Web sites and manual selection to retrieve and store only Web pages that have valuable clinical information. The Web crawler uses the Harvest system (http://harvest.transarc.com) together with additional modules. See also below under "Database selection".

Normalizer. The normalizer uses an optimized algorithm to recognize concepts from the UMLS in free text. It is used for creating normalized text from the full text of Web pages, and normalizing queries to be submitted to the statistical information retrieval system. The end result of concept matching is rather similar to that obtained by the exact match version of SAPHIRE[18], but the algorithm is much more efficient to allow quick indexing of the large database.

Statistical Information Retrieval System. This is a traditional word-statistical system that allows searching for every word in a document and returns results that are ranked by relevance. It consists of an indexing module, that builds an index of words in documents, and a retrieval module, that performs queries against the index. A variety of systems were used at different points in Medical World Search's history, including freeWAIS, Isearch, which uses some of SMART's algorithms, PLWeb (Personal Library Software, Rockville, MD; http://www.pls.com/), and the TOPIC system (Verity, Mountain View, CA; http://www.verity.com/).

User Interface. The user interface allows users to easily specify the desired query, as a combination of medical concepts and words. The user first enters a plain English query, resulting in a list of concepts matched. Depending on the results of the match, the user may select to delete concepts, words, or search the UMLS for concepts that better match his information needs. Boolean queries can be formulated. By default, explosion is performed, but can also be turned off selectively for individual terms. Searching by using Medical World Search is thus highly interactive, which should allow optimal results.

Database Selection

An important component of Medical World Search is the process of selecting and retrieving Web pages with valuable clinical information, to reduce the noise in searches. Sites and Web pages were selected for inclusion in Medical World Search by staring from pages that are listed in the main Internet medicine directories, such as Medical Matrix, CliniWeb, and MedWeb, and retrieving other pages referenced by these pages. In general, for sites listed in these directories, the full contents of the sites were not downloaded, because it was found that doing so would result in the addition of too much noise to the database. Also, some sites were selected as especially

relevant and their full content was downloaded. A good example of such sites is the Virtual Hospital[24].

Discussion Forum

In order to easily obtain feedback from Internet users and facilitate discussions that might arise, a Web-based discussion forum was created on the site. Users may post questions or comments in any of several categories. The discussion forum is regularly monitored.

OBSERVATIONS

Database

Using the Web crawling procedure outlined above, 29,769 Web pages were retrieved and incorporated into the database as of February 1997.

Searches with Medical World Search

The database size obtained seemed sufficient for many queries. For instance, a search on "prostate cancer" results in hits on 286 documents. A search on "heart attack" returns 424 documents. However, for more specific queries, the retrieval sets are relatively small. For "teratocarcinoma", 5 documents are found. For "Christmas disease", 7 documents are found.

Medical World Search Usage

Initially, Medical World Search was publicized on a number of Internet newsgroups. Following this initial publicity Medical World Search has been steadily used by Internet users since October 1996. Until December 31, there has been approximately 1,000 hits a day, with a peak at 1651 hits on December 11. Since then, the usage has been steadily increasing.

Use of the UMLS

Overall, the UMLS provided significant benefits in terms of being able to account for many more synonyms and generate more explosions than its constituent vocabularies.

Mapping of User Queries. A total of 36,716 user queries were examined. 9,479 queries were submitted that consisted of a single word (after stopword removal). Of those queries, 5,396 (56.9%) were successfully mapped to at least one UMLS concept. A significant number of those unmapped queries were misspellings of UMLS concepts, such as "spinalcord", "catecholomines", "lymphangio", and "taxomifen". More unmapped queries were parts of UMLS concepts, but not in the UMLS themselves, such as "temporomandibular", "hypothyroid", or "squamous". Other queries were proper nouns of people or organizations, such as "Wenkebach", "Taekman", "Corbamed", or "Jamison". Other words were apparently not medically related, such as "scouting", "impingement", "consumer", and "jobs". However, a significant number of unmapped queries appeared to be medically relevant, but absent from the UMLS, such as "bioinformatics", "microalbuminemia", and "thymocyte".

Explosion. Explosion can be problematic in the UMLS. Given the presence of vocabulary "contexts", some concepts will have more specific terms that are not always appropriate for explosion. For instance, "myocardial infarction" is a more specific term than "heart" in one of AI/RHEUM's findings contexts, although it is of questionable value for explosion purposes. In one of CPT's contexts, "cardiovascular system" is a more specific term than "surgery". Another potential problem when exploding is the presence of loops within the UMLS hierarchies. For instance, if the hierarchy is followed blindly, the concept "diseases of the skin and subcutaneous tissue" appears to be a more specific term than itself. Of course, these problems originate from some of the UMLS component vocabularies, which are often not strict hierarchies of concepts.

User Feedback

A substantial amount of user feedback was received through the discussion forum. The discussion forum has thus proved its usefulness.

DISCUSSION

Medical World Search's usefulness has been demonstrated by its significant use by the medical community during its short life. We have shown that a majority of single word queries are mapped to UMLS concepts, despite the presence of many terms that cannot be expected to be in the UMLS. Medical World Search thus appears to be an example of successful use of the UMLS for medical information retrieval. We have also noted that there are significant challenges in using the UMLS for information retrieval purposes given the particularities of its structure.

There is a need for evaluating rigorously the performance of Medical World Search in an interactive but well-defined setting, similarly to the way SAPHIRE and MEDLINE systems were evaluated. Such evaluations should be facilitated by the features of Medical World Search that allow monitoring its usage. For instance, there is a record of URLs that were accessed by registered users from the search results page.

There are a number of possibilities for further enhancements and improvements of Medical World Search. Medical World Search can clearly be improved by increasing its database size, resulting in a larger retrieval set on very specific queries. The challenge with this endeavor is to increase the database size without adding noise to the database in the shape of clinically irrelevant documents.

Another avenue is to improve mapping of free text to UMLS concepts, using approximate matching and ambiguity resolution techniques that use the UMLS' SPECIALIST lexicon, shallow parsing, and the semantic type information available in the UMLS, exemplified in the MetaMap system[25,26,27].

Acknowledgements

We thank Dr. Ted Shortliffe for his suggestions of potential medical professional user groups, and Dr. Malet for his suggestions about improving Medical World Search.

References

1. Ramsey PG, Carline JD, Inui TS, et al. Changes over time in the knowledge base of practicing internists. JAMA, 1991; 266:1103-8.
2. Leigh TM, Young PR, Haley JV. Performances of family practice diplomates on successive mandatory recertification examinations. Acad Med, 1993; 68:912-21.
3. Stross JK, Harlan WR. The dissemination of new medical information. JAMA, 1979; 241:2622-4.
4. Stross JK, Harlan WR. Dissemination of relevant information on hypertension. JAMA, 1981; 246:360-2.
5. Williamson JW, German PS, Weiss R, Skinner EA, Bowes F. Health science information management and continuing education of physicians. Ann Intern Med 1989; 110(2):151-160
6. Kunin CM, Tupasi T, Craig WA. Use of antibiotics - a brief exposition of the problem and some tentative solutions. Ann Intern Med, 1973; 79:555-60.
7. Simmons HE, Stolley PD. This is medical progress ? Trends and consequences of antibiotic use in the United States. JAMA, 1974; 227:1023-8.
8. Bernstein LR, Barriere SL, Conte JE. Utilization of antibiotics: analysis of appropriateness of use. Ann Emerg Med, 1982; 11:21-4.
9. Weiner JP, Parente ST, Garnick DW, et al. Variation in office-based quality - a claims-based profile of care provided to Medicare patients with diabetes. JAMA, 1995; 273:1503-8.
10. Ellerbeck EF, Jencks SF, Radford MJ, et al. Quality of care for Medicare patients with acute myocardial infarction: a four-state pilot study from the Cooperative Cardiovascular Project. JAMA, 1995; 273:1509-14.
11. Bachrach CA, Charen T. Selection of MEDLINE contents, the development of its thesaurus, and the indexing process. Med Inf, 1978; 3(3):237-54.
12. Haynes RB, Walker CJ, McKibbon KA, Johnston ME, Willan AR. Performance of 27 MEDLINE systems tested by searches with clinical questions. J Am Med Informatics Assoc, 1994;1:285-95.
13. Hersh W, Brown K, Donohoe L, Campbell E. Managing distributed information on the WWW: CliniWeb. Abstract Book of the 1996 AMIA Spring Congress, 1996: 91.
14. Kogelnik AM, Foote S. MedWeb: biomedical Internet resources. Proceedings of the 1996 AMIA Annual Fall Symposium (Formerly SCAMC), Hanley & Belfus, 1996:969.
15. Hersh WR. Information retrieval: a health care perspective. Springer-Verlag; 1996.
16. Hersh WR, Hickam D. Use of a multi-application computer workstation in a clinical setting. Bull Med Libr Assoc 82(4), October 1994.
17. Lindberg DAB, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med,1993; 32: 281-91.

18. Hersh WR, Greenes RA. SAPHIRE--an information retrieval system featuring concept matching, automatic indexing, probabilistic retrieval, and hierarchical relationships. Comput Biomed Res, 23: 5,1990 Oct, 410-25.

19. Hersh WR. Evaluation of Meta-1 for a concept-based approach to the automated indexing and retrieval of bibliographic and full-text databases. Med Decis Making, 11: 4 Suppl,1991 Oct-Dec, S120-4.

20. Hersh WR, Leone TJ. The SAPHIRE Server: a new algorithm and implementation. Proc Annu Symp Comput Appl Med Care, 1995:858-862.

21. Hersh WR, Hickam DH. A comparison of retrieval effectiveness for three methods of indexing medical literature. Am J Med Sci, 303: 5,1992 May, 292-300.

22. Hersh WR, Hickam DH, Haynes RB, McKibbon KA. A performance and failure analysis of SAPHIRE with a MEDLINE test collection. J Am Med Inform Assoc, 1: 1,1994 Jan-Feb, 51-60.

23. Hersh WR, Hickam DH. An evaluation of interactive Boolean and natural language searching with an on-line medical textbook. J Am Soc Info Sci, 1995; 46:478-9.

24. Galvin JR, D'Alessandro MP, Erkonen WE, Lacey DL, Feld RD, Schwabbauer M, Choi TA, Hatz DA, Finney DM. The Virtual Hospital: Internet publishing as an efficient and effective means for delivering continuing medical education. Proceedings of the 1996 AMIA Annual Fall Symposium (Formerly SCAMC), Hanley & Belfus, 1996: 986.

25. Rindflesch TC, Aronson AR. Semantic processing in information retrieval. Proc Annu Symp Comput Appl Med Care, 1993:611-5.

26. Rindflesch TC, Aronson AR. Ambiguity resolution while mapping free text to the UMLS metathesaurus. Proc Annu Symp Comput Appl Med Care, 1994:240-4.

27. Aronson AR. The effect of textual variation on concept based information retrieval. Proceedings of the 1996 AMIA Annual Fall Symposium (Formerly SCAMC), Hanley & Belfus, 1996:373.