Evaluation of SAPHIRE: An Automated Approach to Indexing and Retrieving Medical Literature

William Hersh, M.D.
David H. Hickam, M.D., M.P.H.
Oregon Health Sciences University
Portland, Oregon, USA

R. Brian Haynes, M.D., Ph.D.
K. Ann McKibbon, M.L.S.
McMaster University
Hamilton, Ontario, Canada

Abstract

An analysis of SAPHIRE, an experimental information retrieval system featuring automated indexing and natural language retrieval, was performed on MEDLINE references using data previously generated for a MEDLINE evaluation. Compared with searches performed by novice and expert physicians using MEDLINE, SAPHIRE achieved comparable recall and precision. While its combined recall and precision performance did not equal the level of librarians, SAPHIRE did achieve a significantly higher level of absolute recall. SAPHIRE has other potential advantages over existing MEDLINE systems. Its natural language interface does not require knowledge of MeSH, and it provides relevance ranking of retrieved references.

Introduction

MEDLINE is the one of the largest and most frequently used online databases in the world. In addition to approximately four million searches carried out annually on the networks of the National Library of Medicine (NLM) [1], MEDLINE can also be searched via several commercial online vendors, as well as in the form of many CD-ROM products. Despite the great commercial success of MEDLINE, it does have limitations in both indexing and retrieval. Human indexing is expensive; over two million dollars and 44 full-time equivalent indexers are used by the NLM each year to index MEDLINE [2]. This problem will become exacerbated as more medical text becomes available online. Human indexing is also inconsistent. Funk and Reid looked at MEDLINE references that were, for a variety of reasons, indexed in duplicate [3]. They found that the consistency of index term assignment for central concept headings (starred MeSH terms, the most important concepts in the article) was 61%, while for heading-subheading combinations it was only 38%.

On the retrieval side, most online implementations of MEDLINE have cryptic command-line interfaces. While recent microcomputer products provide a more user-friendly interface, searching is still limited to Boolean search statements phrased in either the Medical Subject Headings (MeSH) vocabulary or individual text words. Slingluff has found that MeSH terms are often difficult for users to find and apply in Boolean expressions [4]. Also, MeSH terms are assigned by indexers as specified by the MEDLARS Indexing Manual [5], with which few end-users have familiarity. Thus, end-users are often unable to apply the proper MeSH terms for searching.

These limitations of MEDLINE and other information retrieval systems are motivation for automated approaches to information retrieval, utilizing such features as automated indexing, natural language query input, and ranking retrieved citations [2]. Early systems that have experimented with these approaches include Salton's SMART system [6] and the NLM's IRX Project [7]. These automated systems use indexing and retrieval based only on individual words. SAPHIRE is a new system for automated indexing and retrieval which uses many of the features of these word-based programs, but it indexes and retrieves at the level of full medical concepts [8]. SAPHIRE indexes text by means of *concept-based automated indexing*, in which the assignment of indexing terms is based not on individual words, as is done in individual word indexing, but on actual medical concepts, similar to the MeSH terms assigned to articles in the MEDLINE database. Unlike MEDLINE, however, this is done by machine, avoiding the expense and inconsistency of human indexers. Retrieval is conducted by *natural language input, relevance-based retrieval.* The user can enter questions in plain English, have the computer identify the relevant concepts, and have the computer search for documents based on those concepts. Furthermore, the output of retrieved content material is ranked by its relevance.

Background

SAPHIRE

The core of the SAPHIRE program is the *concept-matching algorithm*, which recognizes concepts for both indexing and retrieval. This algorithm takes as its input any string of text, such as a document sentence or a user query, and returns a list of all concepts found, mapped to their *canonical* or preferred form. This is done by

0195-4210/91/$5.00

detecting the presence of *word-level synonyms* between words (e.g., *high* and *elevated*) and *concept-level synonyms* between concepts (e.g., *hypertension* and *high blood pressure*). The algorithm requires a vocabulary of concepts and their synonym forms, which are related by a common unique identifier. A word thesaurus, while not required, allows an even greater diversity of concepts to be recognized.

For the indexing process, a document is defined to be any collection of text with a title, a unique identifier, and a body of text. The document title and its text are passed to the concept-matching algorithm. As concepts are returned, the frequency of each concept occurrence is summed. The result of the indexing process is that all concepts are recognized in each document and mapped to their canonical form. Each concept is then given a weight, based on the same type of statistical approach used by IRX, a word-based experimental information retrieval program developed at the NLM [7]. The weight contains two components, the first of which is the inverse document frequency:

$$IDFi = \log\left(\frac{\text{\# of documents in collection}}{\text{\# documents with concept i}}\right) + 1 \quad (1)$$

The second factor is the intradocument term frequency:

$$TFij = \log(\text{freq of concept i in document j}) + 1 \quad (2)$$

The weight of a concept i in document j is therefore:

$$WEIGHTij = IDFi * TFij \quad (3)$$

Natural language input, relevance-based retrieval is carried out by taking the user-entered natural language search statement and returning a list of matching documents in ranked order. After the user enters a natural language query, the text is passed to the concept-matching algorithm. A wild-card character can be used to have words completed for the user when, for example, they are unclear on the exact spelling. The algorithm extracts all concepts for searching and returns them in a list. This list is then scored in a fashion modeled after IRX, in which each document receives a score based on the sum of the weights of the terms in the query. The resulting list of matching documents is then sorted, with the weights normalized such that the highest ranking document is given a score of 100. The output can then be adjusted to emphasize recall or precision by removing all documents below a certain score.

Meta-1

One of the major factors in allowing SAPHIRE to perform concept-based indexing in a broad domain such as biomedicine is the existence of a large vocabulary with a great breadth of concepts as well as a great depth of synonyms. The vocabulary used by SAPHIRE is Meta-1 [9], which is one of the knowledge sources in the Unified Medical Language System Project of the NLM [10]. Meta-1 contains two types of records, concept records and synonym records. The former contain the main concepts that occur in the metathesaurus, while the latter link synonyms to the main concepts. The purpose of the synonym records is to store information about the synonyms that would be inappropriate in the main concept. Each type of record has lexical variants. In order for SAPHIRE to use Meta-1, it must be "collapsed" to a one-concept, one-record format, consisting of the main concept, its synonyms, its lexical variants, and the lexical variants of its synonyms. In our version of Meta-1 there are 28,423 concepts, 78,244 synonyms, and 28,603 word stems.

Retrieval Evaluation

There are many facets involved in the evaluation of information retrieval systems. In this study, we measured recall and precision for each search. Recall is defined as the proportion of relevant articles that are retrieved from the entire collection:

$$Recall = \frac{\text{articles retrieved and relevant}}{\text{total articles relevant in collection}} \quad (4)$$

Precision is defined as the proportion of articles relevant from a given search:

$$Precision = \frac{\text{articles retrieved and relevant}}{\text{total articles retrieved}} \quad (5)$$

Recall is difficult to measure in large databases such as MEDLINE due to the improbability of knowing the total number of articles relevant for a given search. This problem is overcome by either using estimates of total relevant articles (such as by having three searchers search on the same query) or by using a test collection where each article is known to be relevant or not for each query.

A number of recall and precision studies evaluating MEDLINE have been carried out over the last two decades. The first large-scale study of MEDLINE was first done in 1966-68 by Lancaster, who evaluated 302 searches submitted by librarians to the NLM [11]. The mean recall of these searches was 57.7% while precision was 50.4%. The most recent analysis of MEDLINE was performed by Haynes et al, who divided their subjects into three groups - librarians, expert physicians, and novice physicians [12]. In this study, physicians and medical students who were novices in using MEDLINE themselves originated the search request. Before searching online, they wrote a brief description of the search question; this was later used by physicians expert in the use of MEDLINE and librarians to conduct independent searches. All citations retrieved were judged for relevance by a clinician who was expert in

the area of the search topic and was unaware of which searcher had retrieved a given citation. Their results showed medical librarians to have a recall of 48.7% and a precision of 57.9%. Expert physician searchers had about the same recall (47.7%) but lower precision (47.1%), while inexperienced physician searchers had much poorer recall (27.0%) and precision (37.1%).

One problem in using recall and precision to evaluate literature searching is the varied approaches used for their calculation. Lancaster's study and our own initial study estimated overall recall and precision by calculating recall and precision for each individual search and then averaging values over all searches. Haynes, on the other hand, summed the number of relevant and retrieved articles over all searches and calculated recall and precision based on the sums. With the former approach, two methods of searching can be compared by using a paired t-test or Wilcoxon signed-rank test to evaluate the difference in mean values of recall and precision. With the latter method, two searching methods are compared using a chi-square test for equality of proportions. The two methods yielded similar estimates of recall and precision for Haynes' original MEDLINE data, except for novice physician recall, which is 37.2% with the former method and 26.8% with the latter method. This difference is due to the latter giving more weight to searches with a larger number of retrieved references. Because the former method may be less biased, we have chosen to continue to average recall and precision over all individual searches to calculate the overall values.

Comparison of retrieval performance is made more difficult when comparing traditional searching systems with those that use document ranking, such as SAPHIRE, because recall and precision will vary based on how far down the ranked retrieval list the user wishes to look. In general, as more documents are evaluated, recall will increase while precision will decrease. For this reason, many evaluations of systems using ranking techniques will consist of a recall-precision curve, whose points are the recall and precision values for a given cutoff of the ranked list. The optimal retrieval system will have its points closest to the upper right corner of the graph, which is where 100% recall and precision occur. In the real world, of course, there is a trade-off between the two values, so that most recall-precision curves are downsloping with high precision and low recall at one end and low precision and high recall at the other. Retrieval systems that do not rank documents only produce a single recall-precision point. In this study, we compared SAPHIRE with traditional MEDLINE searching with novice physicians by choosing the point on the recall-precision curve closest to the original physicians' values.

A previous evaluation of SAPHIRE showed it to have better retrieval performance than traditional MEDLINE searching for physicians, although the study was limited by several factors [13]. First, all MEDLINE-style Boolean search statements were prepared by users on paper, preventing them from deriving feedback typically present in an online session. Second, the content material was limited to conference proceedings abstracts exclusively on the topic of AIDS, which is not representative of MEDLINE in general. Nonetheless, this study did show that entering the text of the original query into SAPHIRE's natural language interface led to better recall and precision than Boolean search statements generated by physicians.

Methods

A test collection was constructed to serve as a gold standard for recall and precision calculations. We were fortunate to have access to a large collection of actual user queries with citations retrieved and relevance judgments for each. These data consisted of the citations and judgments from Haynes' MEDLINE evaluation described above [12]. The raw data consisted of 78 queries and 3,403 citations. Since SAPHIRE relies on indexing of the title and abstract, all citations in which an abstract was not present were eliminated, leaving a total of 2,344 citations. After elimination of citations without abstracts, there were three queries with no relevant citations, which were also eliminated from the test collection, leaving a total of 75 queries.

With the test collection, we first recalculated recall and precision for each novice, expert, and librarian searcher based on the new 2,344-reference test collection. SAPHIRE searching was performed by entering the user's initial free text statement of the search subject into SAPHIRE's natural language interface. The text was largely the same as it occurred in the raw data, with the exception of a few spelling corrections and acronym expansions (when the acronyms were not present in Meta-1). Recall and precision were calculated for SAPHIRE by eliminating all documents below weight cutoffs between 0% and 95% at 5% increments. Levels of recall and precision were compared between each group and SAPHIRE at 60% cutoff, which was the closest that the SAPHIRE recall-precision graph came to the results for MEDLINE searchers (see Figure 1). A Wilcoxon signed-rank test was used to compare the mean values. We also compared the absolute recall of SAPHIRE (with no weight cutoff) versus librarians by a Wilcoxon signed-rank test.

Results

The recalculated values of recall and precision for the

original searchers, expert physicians, librarians, and SAPHIRE at 60% weight cutoff are summarized in Table 1. The mean recall and precision values for SAPHIRE at different weight cutoffs are shown in Table 2. These values are plotted in a recall-precision graph in Figure 1. At the 60% weight cut-off using the individual search method, SAPHIRE's improvement over original physicians in recall (P = .5) and precision (P = .7) was not statistically significant by Wilcoxon signed-rank analysis. Similarly, the advantage in recall (P = .3) and precision (P = .3) for expert physician searchers over SAPHIRE was not statistically significant.

Search Group	Recall	Precision
Novice	42.3	39.7
Expert	51.3	46.6
Librarian	52.6	59.8
SAPHIRE at 60%	45.5	41.8

Table 1 -- Recall and precision for each search group.

SAPHIRE achieved the largest absolute recall of all groups, although for SAPHIRE to achieve equal recall with librarians (55.1% vs. 52.6% respectively at weight cutoff 50%), there was a significant cost in terms of precision (33.7% vs. 59.8%, P < .02). Nonetheless, SAPHIRE's recall went as high as 82.8%, a figure unmatched by librarians using regular MEDLINE (77.4% vs. 52.6%, P < .001).

Weight	Recall	Precision
0	77.4	11.4
5	77.4	11.4
10	77.1	11.4
15	76.4	11.6
20	75.5	12.6
25	74.9	14.1
30	73.1	17.2
35	70.5	21.5
40	65.7	26.5
45	60.1	30.7
50	55.1	33.7
55	48.0	38.1
60	45.5	41.8
65	37.1	44.6
70	32.6	47.3
75	28.5	51.4
80	24.3	52.2
85	17.9	51.9
90	14.8	52.8
95	12.1	55.0

Table 2 -- Recall and precision values for SAPHIRE searching at different weight cutoffs.

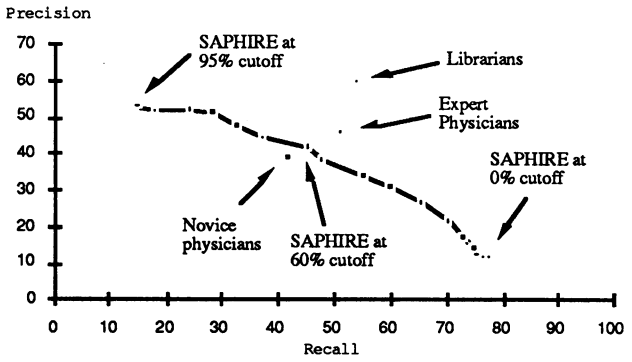

Figure 1 -- Recall-precision graph for SAPHIRE searching. The individual points for original, expert, and librarian searchers are also shown.

Discussion

This study demonstrated the equivalence of SAPHIRE's approach to indexing and retrieval of MEDLINE references for novice and expert physicians. SAPHIRE still does not match the performance of librarians using MEDLINE, although it does achieve higher absolute recall. There were some aspects to this study that potentially hindered SAPHIRE's results. First, the articles in the test collection created for this study were only those that could be retrieved from MEDLINE using MeSH and text word queries. Given the previous work by McKibbon et al [14] which showed that MEDLINE queries done by three different searchers often led to retrieval of disparate sets of references, it is likely that there are additional relevant MEDLINE references that were not part of this test set. Certainly some of those references could be retrieved by SAPHIRE only. A second limitation is the fact that SAPHIRE was not used interactively for these experiments. Since some of the queries in the test set do not contain important MeSH terms that are likely to occur in the documents, and thus make them retrievable by SAPHIRE, recall would most likely be improved by an interactive searcher using SAPHIRE's interface to add important MeSH terms not in the original query. In an interactive searching situation, a user would be likely to use SAPHIRE to find appropriate MeSH terms.

Since a limitation of both initial evaluations of SAPHIRE has been the lack of interactive use of SAPHIRE, we are presently carrying out experiments with searchers using both SAPHIRE and traditional Boolean interfaces. These experiments will not only provide more useful performance data, but will also help to evaluate other features not assessed by recall and precision, such as ease of use and system response time with large databases. Furthermore, these experiments are being performed with additional types of databases, in particular those providing full-text.

More evaluation of SAPHIRE is needed before it can be advocated for usage with very large databases the size of all of MEDLINE. To begin with, SAPHIRE is currently a research program and cannot maintain the sheer quantity of references residing in the database of the ELHILL computer at the NLM. Furthermore, its weighting algorithms could not, in their present form and with current computer hardware, operate on collections with several million records. For very large databases, some modifications of the algorithms, such as document clustering techniques [6], would be necessary. Finally, it remains to be seen whether the lower precision found with SAPHIRE is problematic with very large databases.

While some might argue that the best approach to enhancing search results is to elevate the searching skills of physicians up to that of librarians, SAPHIRE provides a feasible alternative. Furthermore, it is not a static program, and we are aiming to improve the current level of performance by implementing and evaluating new approaches based on selective use of natural language processing techniques that allow improved concept recognition [15] and use of article references of recent papers to find relevant articles more precisely. In the long run, we hope to identify and validate a set of techniques for optimal information retrieval in the biomedical domain.

Bibliography

1. Siegel E, Cummings M, Woodsmall R. "Bibliogrpahic Retrieval Systems." Medical Informatics: Computer Applications in Health Care. Shortliffe EH,Perreault LE ed. 1990 Addison-Wesley. Reading, MA.

2. Hersh W, Greenes R. Information retrieval in medicine: State of the art. MD Comp. 7: 302-311, 1990.

3. Funk M, Reid C. Indexing consistency in MEDLINE. Bull Med Lib Assoc. 71: 176-183, 1983.

4. Slingluff D, Lev Y, Eisan A. An end-user search service in an academic health sciences library. Med Ref Serv Q. 4: 11-21, 1985.

5. Charen T. "MEDLARS Indexing Manual, Parts I and II." 1983 National Technical Information Service. Springfield, VA.

6. Salton G. "Introduction to Modern Information Retrieval." 1983 McGraw-Hill. New York.

7. Harman D, Benson D, Fitzpatrick L, Huntzinger R, Goldstein C. IRX: An information retrieval system for experimentation and user applications. SIGIR Forum. 22: 2-10, 1988.

8. Hersh W, Greenes R. SAPHIRE: An information retrieval environment featuring concept-matching, automatic indexing, and probabilistic retrieval. Comput Biomed Res. 23: 405-420, 1990.

9. Hersh W. Evaluation of Meta-1 for a concept-based approach to the automated indexing and retrieval of bibliographic and full-text databases. Med Decision Making. in press. 1991.

10. Lindberg D, Humphreys B. The UMLS knowledge sources: Tools for building better user interfaces. SCAMC 14. 14: 121-125, 1990.

11. Lancaster F. Evaluation of the MEDLARS Demand Search Service. National Library of Medicine. 1968.

12. Haynes R, McKibbon K, Walker C, Ryan N, Fitzgerald D, Ramsden M. Online access to MEDLINE in clinical settings. Ann. Int. Med. 112: 78-84, 1990.

13. Hersh W, Hickam D. A comparison of retrieval effectiveness for three methods of indexing AIDS-related abstracts. Proceedings of Amer Soc for Information Science, in press. 1991.

14. McKibbon K, Haynes R, Dilks CW, Ramsden M, Ryan N, Baker L, Flemming T, Fitzgerald D. How good are clinical MEDLINE searches? A comparative study of clinical end-user and librarian searches. Comp. Biomed. Res. 23: 583-593, 1990.

15. Evans D, Hersh W, Monarch I, Lefferts R, Handerson S. Automatic indexing of abstracts via natural language processing using a simple thesaurs. Med Decision Making. in press, 1991.