Information Retrieval in Medicine: The SAPHIRE Experience

William R. Hersh
Biomedical Information Communication Center, Oregon Health Sciences University, 3181 S.W. Sam Jackson Park Rd., Portland, OR 97201. Phone: 503-494-4563; Fax: 503-494-4551; E-mail: hersh@ohsu.edu

David Hickam
VA Medical Center, Health Services Research and Development (152), P.O. Box 1034, Portland, OR 97207. Phone: 503-273-5305; Fax: 503-273-5367; E-mail: hickam&hsrd.gov

Information retrieval systems are being used increasingly in biomedical settings, but many problems still exist in indexing, retrieval, and evaluation. The SAPHIRE Project was undertaken to seek solutions for these problems. This article summarizes the evaluation studies that have been done with SAPHIRE, highlighting the lessons learned and laying out the challenges ahead to all medical information retrieval efforts.

Once confined mainly to medical libraries and computer pioneers with modems, medical information retrieval (IR) systems have become widespread. Few medical schools or large medical centers lack access to MEDLINE, which is often subsidized for students and staff. Other IR databases in areas such as nursing and drug information are also widely available, and there is increasing access to multimedia materials and the Internet.

Despite their prevalence, however, many impediments to effective use of medical IR systems remain. It is still not known, for example, how to index databases most effectively. Likewise, there are disagreements over the best way to phrase search statements, whether with traditional Boolean operators, "natural language" input, or more complex representations. Finally, there are problems with evaluation approaches that evolved in library settings but are less suitable for the end-user environment.

For the past 5 years, the SAPHIRE (Semantic and Probabilistic Heuristic Information Retrieval Environment) Project has been devoted to identifying the optimal approaches to indexing, retrieval, and evaluation of IR resources in the biomedical domain. The scope of the research project has gone beyond just the development and testing of the SAPHIRE software itself. The purpose of this article is to review the evaluation studies of the project, which provide insight into the problems that exist in all IR systems and how they may guide future research. We begin by describing the basic issues that motivate the SAPHIRE Project, followed by a summary of six evaluation studies. We then summarize the conclusions drawn from this work and lay out the challenges ahead.

IR Problems and SAPHIRE Solutions

IR databases are generally of two types: bibliographic and full-text. The former typically consist of references to the original medical literature, while the latter contain the complete text of documents from journals, textbooks, and other print sources. Bibliographic databases usually contain indexing terms assigned by a human indexer from a controlled vocabulary, although the other content words in the reference (e.g., those from the title and abstract) are usually searchable. In contrast, full-text databases are usually indexed based on the words present in the entire document.

Although human indexing of bibliographic databases is considered state-of-the-art, it is still unclear how much benefit this indexing provides. While humans are good at discerning the focus of a document, manual indexing is known to be inconsistent (Funk & Reid, 1983), and the vocabulary terms can be difficult for novices to master (Kirby & Miller, 1986). The usual alternative to human indexing is to index all words in the document. But this too presents problems. For example, medical language is known to have much synonymy (different words meaning the same thing) and polysemy (the same words meaning different things). Furthermore, extracting single words for indexing removes the context in which they occur (i.e., the phrase *high blood pressure* has a different meaning from any of the three words used individually). Computational linguistic approaches to discerning concepts and relationships have been advo-

cated, but constructing knowledge resources to recognize these elements has been difficult (Salton, Buckley, & Smith, 1990).

On the retrieval side, search terms combined by the Boolean operators AND and OR have been used in most systems, but they have been shown to be difficult for novices (Sewell & Teitelbaum, 1986) and their benefit is also unknown. Another retrieval problem is that most systems return documents in arbitrary order. A potential solution is to rank documents based on the frequency of terms similar to the query and document, called relevance ranking (Salton, 1991).

The only way to determine which methods of indexing and retrieval work best has been to evaluate them empirically. There is a large literature on IR system evaluation, and much has been learned about the performance of various techniques. However, there is also disagreement over the value of the measures used to assess performance and how to apply them in different settings (Hersh, 1994).

The SAPHIRE Project was implemented to address the problems in current IR systems. SAPHIRE provides potential solutions to the problems of inconsistent human indexing, clinician difficulty with controlled vocabularies and Boolean searching, and the rich synonymy of medical language. The details of the system have been described elsewhere (Hersh, 1991), but the basic approach is to extend word-based automated methods by indexing on concepts found in text instead of individual words. The vocabulary for identifying concepts and their synonyms is based on the Metathesaurus from the National Library of Medicine's (NLM) Unified Medical Language System (UMLS) Project (Lindberg, Humphreys, & McCray, 1993). In order to compare SAPHIRE with existing approaches to IR, two additional systems have been implemented: One combining word-based indexing with Boolean searching (BOOLEAN), and another combining word-based indexing with natural language searching (SWORD, Statistical Word-Oriented Retrieval from Databases) (Hersh & Hickam, 1995).

Evaluation Studies

Over the past 4 years, there have been six evaluation studies in the SAPHIRE Project. Each of these studies has provided incremental information about the benefits and limitations of SAPHIRE as well as adding to general knowledge about the indexing, retrieval, and evaluation of IR systems in the biomedical setting. All of these studies used the measures of *relative recall* (proportion of known relevant references retrieved from the database, hereafter referred to as *recall*) and *precision* (proportion of relevant references retrieved by the search). These measures are often considered the gold standard of retrieval system evaluation, yet have limitations in measuring the true effectiveness of retrieval systems.

SAPHIRE vs. Human Indexing of AIDSLINE Documents (Hersh & Hickam, 1992)

The first evaluation study of SAPHIRE used a 200-document subset from the AIDSLINE database, along with 12 queries generated by library users at the NLM and Oregon Health Sciences University (OHSU). The initial component of the study was to compare SAPHIRE indexing with both human indexing of MEDLINE and simple word-based indexing of the title and abstract fields in a command-line searching environment. SAPHIRE performed inferiorly as an indexing replacement in this environment. An unexpected finding was that the search results of librarians were as good using text words only as with the full MEDLINE feature set, while physicians actually had better results with text words.

Because SAPHIRE was not designed to run in a command-line Boolean searching environment, another component of the study looked at SAPHIRE's performance by entering the free-text query statement directly in the natural language interface. SAPHIRE performed better in this latter mode, and, in fact, achieved better recall and precision than physicians using Boolean search statements, though not as good as librarians.

This study also looked at the frequency and type of concept matching errors made by SAPHIRE. An average of 2.7 (out of an average of 18) inappropriate indexing assignments per abstract were made, with problems arising due to syntax (the verb *lead* matched as the chemical element, a noun), abbreviations (PCP meant *P. carinii pneumonia* in this domain but matched to the drug *phencyclidine*), and stemming (the state *Maine* was stemmed to the word *main*).

SAPHIRE vs. Conventional MEDLINE Searching of MEDLINE Documents (Hersh, Hickam, Haynes, & McKibbon, 1994c)

The second study of SAPHIRE used previously searched topics and judgments of relevance from a clinical evaluation of Grateful Med at McMaster University (Haynes et al., 1990). A test collection of 2,344 MEDLINE references was created, consisting of all references that were retrieved (and contained abstracts) for 75 queries generated by clinicians. In this study, SAPHIRE's recall and precision appeared intermediate between expert and novice clinician searchers, although none of the differences among the groups was statistically significant.

A failure analysis identified some recurring patterns for false-negative and false-positive retrievals. The most common causes of relevant documents failing to be retrieved were the presence of synonyms not recognized by SAPHIRE (i.e., the form in the document was not in the Metathesaurus) or terms being present but at a different level of granularity (i.e., the query might have had the term *antibiotic* while the document contained the actual antibiotic name). The most common reason for retrieval

of nonrelevant documents was the presence of most or all query terms in the document, but with a different focus or relationship between the terms.

SAPHIRE vs. Word-Based Boolean Searching of Yearbook Series *Extended Abstracts (Hersh & Hickam, 1993)*

One problem with the first two studies was the lack of interactive searching by real users. The next study corrected that problem by using a group of 16 senior medical students to search on 10 questions generated on medical rounds at the University of Pittsburgh. The database used was six volumes from the *Yearbook Series*, a publication which provides abstracts and commentaries for all the major articles in a given field published each year. For indexing and retreival purposes, each document consists of the title and text. Each student searched half of the questions with SAPHIRE and the other half with BOOLEAN. No statistically significant difference in recall or precision was found between the two systems.

SAPHIRE vs. Word-Based Natural Language Searching of AIDSLINE, MEDLINE, and Yearbook Series *Documents (Hersh, Hickam, & Leone, 1992)*

The next study took advantage of the existence of the test collections from the previous three studies to assess various approaches to word-based and concept-based automated systems. Like the first two studies, this study was conducted in a non-interactive setting, with queries entered in batch mode. This study compared SAPHIRE with SWORD alone, SWORD and SAPHIRE combined, and a version of SAPHIRE with a different concept-matching algorithm. The latter eliminated the exact word order requirement of SAPHIRE's original concept-matching algorithm and instead only required that words in a concept be adjacent. It also allowed partial matching as long as more than half of the words were present, aiming to overcome the problem of synonyms not matching the exact form in the Metathesaurus. The results showed that SWORD had the best overall performance. The combination of SWORD and SAPHIRE (weighting both individual words from SWORD and concepts from SAPHIRE) performed intermediately between the two programs alone, while the version of SAPHIRE with the new concept-matching algorithm performed worst, due to excess inappropriate concept matching.

This study also assessed a simple form of relevance feedback for SAPHIRE and SWORD, using the entire top-ranking relevant document to replace the original query. This enhanced performance for both systems, although SWORD still outperformed SAPHIRE.

SAPHIRE vs. Word-Based Boolean Searching and Word-Based Natural Language Searching of Scientific American Medicine *(Hersh & Hickam, in press)*

This study compared SAPHIRE, SWORD, and BOOLEAN using a different type of database, which was the internal medicine textbook, *Scientific American Medicine* (SAM). The textbook was subdivided into 6,623 "documents." In the study, 21 senior medical students searched on 18 queries each, half with one of the three systems and half with another. These queries were generated by the internal medicine faculty and the house staff in the OHSU General Medicine Clinic. Each query was searched using each of the three systems. As with the *Yearbook Series* study above, there was no statistically significant difference in recall or precision among the three systems.

Boolean vs. Free-Text Searching in MEDLINE (Hersh & Hickam, 1994)

A final study did not involve the SAPHIRE software, but rather compared the commercial product Knowledge Finder (KF) (Aries Systems, Inc., North Andover, MA), which uses word-based natural language searching similar to SWORD, with conventional command-line use of MEDLINE on the NLM's ELHILL system. In this study, KF was placed on a Macintosh workstation in the OHSU General Medicine Clinic. Before each search, users entered a brief statement about their patient and information need. These statements were used by librarians and experienced clinician searchers to replicate the searches. Each search was repeated by two librarians and two clinicians, with one librarian and one clinician using the full MEDLINE feature set, and the other librarian and clinician using Boolean combinations of text words. (On ELHILL, text words are defined as all words that appear in the title, abstract, and MeSH fields.)

The results of this study showed that the KF searchers had significantly higher recall and lower precision than all of the other searchers. Thus they found many more relevant references but also many more nonrelevant ones. This was due to the much larger retrieval sets that they obtained, an average of 88 references for the KF group and 15 for the others. KF (and word-based natural language systems in general) tend to have larger retrieval sets, since they retrieve and rank all of the documents that contain as little as one word from the query. However, their relevance ranking techniques lead to relevant documents tending to be ranked nearer the top of the retrieval set. In an attempt to control for the larger retrieval set size, additional recall and precision values were calculated for KF with the default retrieval size set at 15 (the average size of the non-KF retrieval set). With the reduced KF set, the KF searching results were very close to those of the other searchers.

This study showed that word-based natural language

searching in the hands of clinicians was as effective as searching by index term. This study also verified the observation from the AIDSLINE study above that simple text word searching is just as effective as using all of the advanced MEDLINE techniques, especially for non-librarians. In fact, while librarians obtained statistically significant improvement in recall over clinicians using the full MEDLINE feature set, they did not obtain significant improvement over clinicians using just text words, suggesting that advanced MEDLINE features are beneficial mainly to librarians.

Conclusions

The various experiments performed in the SAPHIRE Project have provided insight into the performance of a number of different indexing and retrieval techniques with a wide variety of resources (bibliographic databases, extended abstract collections, and textbooks). A number of conclusions can be drawn from these studies, although the results also serve to show the limitations of current evaluation methods and the need for better ones.

Indexing

The project has addressed two questions in indexing. First, for bibliographic databases, do human-assigned indexing terms offer benefit over machine-assigned words or concepts? Our studies suggest that the incremental benefit of human indexing as measured by retrieval performance is small. A follow-up on the KF study above using the SMART system found that the presence of the words in the MeSH term field conferred about a 10% performance benefit (Hersh, Buckley, Leone, & Hickam, 1994a).

The second question is whether concept-based automated indexing offers any benefit over the use of single words. In the aggregate, it appears it does not. Salton, the foremost advocate of word-based automated indexing, has argued that no methods of automated indexing have improved upon the use of words alone (Salton, 1991). Nonetheless, in the failure analysis of SAPHIRE, instances occurred when the synonyms present in the Metathesaurus led to superior retrieval performance (Hersh et al., 1994c). It is possible that more complex systems that utilize computational linguistic approaches (i.e., CLARIT [Evans, Hersh, Monarch, Lefferts, & Handerson, 1991]) may show a benefit for concept-based indexing, but such studies have not yet been published.

Retrieval

Three issues have been assessed in retrieval. The first is whether searching with MeSH terms offers a benefit over the use of text words alone, which may include the words of those terms. In the KF study, the benefit of using MeSH terms was seen mainly for librarians, who are well-trained in the use of those terms. Clinicians are less experienced in using MeSH terms, and as such do not show improved searching performance.

The second question is the comparison of natural language vs. Boolean searching. In studies using both types of approaches, comparable results were achieved. In the SAM and *Yearbook Series* studies using SAPHIRE, SWORD, and/or BOOLEAN, there were minimal differences in recall or precision. Likewise, for physicians in the KF study, there was little difference between KF and Boolean searching with either the full MEDLINE feature set or text words alone.

The third issue is the benefit of relevance feedback. With small databases, relevance feedback was definitely seen to offer benefit, while in a follow-on to the KF study, the gains were minimal (Hersh et al., 1994a). As with SAPHIRE's situation-specific benefit of synonyms, there may be only intermittent benefit for relevance feedback as well.

Evaluation

Our research has also provided insight into IR evaluation measures themselves. Although the measures of recall and precision have enhanced our understanding of IR systems in general and allowed assessment of individual features within and across different systems, they do not provide all of the insight we might like to have in assessing the use of these systems. While few would argue against retrieving more relevant and fewer nonrelevant documents, it has not been shown that the quantity of relevant documents necessarily correlates with the overall quality of a search.

Even if recall and precision did correlate with value of information obtained, there is another problem that arises when comparing systems, which is: What constitutes a significant difference? With a large enough sample size we can, of course, show that a difference in recall or precision is statistically significant. But it is less clear what a "clinically" significant difference would be. For example, while we know that a drop in diastolic blood pressure from 110 to 90 mm Hg would lead to significantly different medical outcomes in a patient population, it is less certain what level of difference in recall and/or precision is necessary to show that a given indexing or retrieval method is superior to another (i.e., would achieve a better quality search).

An additional problem in the use of recall and precision is how to define a relevant document. It has been argued that the relevance of a document to an information need cannot be assigned objectively, especially by a third party, and that the user and his/her situation must be taken into account (Schamber, Eisenberg, & Nilan, 1990). There is no data to support or refute that statement. It has also been argued that relevance judgments are unreliable. The level of interobserver vari-

ability in our relevance judgments for the test collections we built were moderate, with kappa scores on duplicated relevance judgments ranging from 0.35 to 0.59.

Future Challenges

Like many research efforts, the SAPHIRE Project has answered some questions but uncovered many new ones. This work has shown clearly that easy-to-use systems featuring automated indexing, natural language queries, and relevance ranking perform comparably to traditional Boolean systems. Whether more sophisticated indexing procedures—such as the concept mapping and synonym substitution used in SAPHIRE—are of benefit is less clear.

It is certain, however, that better measures of evaluating systems are needed. Recall and precision may not be adequate for comparing the benefit of systems. Not only must better measures of evaluation be developed, but they must also be applied in realistic settings. System assessment via batch input of queries may provide useful preliminary information, but evaluation of searches by real users with realistic databases (even if in a simulated setting) is necessary.

Our research has begun to address these issues. We are currently porting our systems to run as client-server applications on the World Wide Web, making them available in numerous clinical sites (as well as a laboratory for simulation). We are also assembling a suite of realistic databases that would likely benefit clinicians. Finally, we have also begun to experiment with new measures of performance, including those that measure information obtained and not just number of relevant documents (Hersh et al., 1994b).

Acknowledgment

The majority of the SAPHIRE Project has been funded by grant LM 05307 of the NLM. Additional funding has been provided by Contract 467-MZ-001022 of the NLM and Grant 9040 of the Medical Research Foundation of Oregon.

References

Evans, D., Hersh, W., Monarch, I., Lefferts, R., & Handerson, S. (1991). Automatic indexing of abstracts via natural language processing using a simple thesaurus. *Medical Decision Making, 11,* S108–S115.

Funk, M., & Reid, C. (1983). Indexing consistency in MEDLINE. *Bulletin of the Medical Library Association, 71,* 176–183.

Haynes, R., McKibbon, K., Walker, C., Ryan, N., Fitzgerald, D., & Ramsden, M. (1990). Online access to MEDLINE in clinical settings. *Annals of Internal Medicine, 112,* 78–84.

Hersh, W. (1991). Evaluation of Meta-1 for a concept-based approach to the automated indexing and retrieval of bibliographic and full-text databases. *Medical Decision Making, 11,* S120–S124.

Hersh, W. (1994). Relevance and retrieval evaluation: Perspectives from medicine. *Journal of the American Society for Information Science, 45,* 201–206.

Hersh, W., Buckley, C., Leone, T., & Hickam, D. (1994a). OHSUMED: An interactive retrieval evaluation and new large test collection for research. *Proceedings of the 17th Annual International ACM Special Interest Group in Information Retrieval* (pp. 192–201).

Hersh, W., Elliot, D., Hickam, D., Wolf, S., Molnar, A., & Leichtenstein, C. (1994b). Towards new measures of information retrieval evaluation. *Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care* (pp. 895–899).

Hersh, W., & Hickam, D. (1992). A comparison of retrieval effectiveness for three methods of indexing medical literature. *American Journal of the Medical Sciences, 303,* 292–300.

Hersh, W., & Hickam, D. (1993). A comparison of two methods for indexing and retrieval from a full-text medical database. *Medical Decision Making, 13,* 220–226.

Hersh, W., & Hickam, D. (1994). The use of a multi-application computer workstation in a clinical setting. *Bulletin of the Medical Library Association, 82,* 382–389.

Hersh, W., & Hickam, D. (1995). An evaluation of interactive Boolean and natural language searching with an on-line medical textbook. *Journal of the American Society for Information Science, 46,* 478–489.

Hersh, W., Hickam, D., Haynes, R., & McKibbon, K. (1994c). A performance and failure analysis of SAPHIRE with a MEDLINE test collection. *Journal of the American Medical Informatics Association, 1,* 51–60.

Hersh, W., Hickam, D., & Leone, T. (1992). Word, concepts, or both: Optimal indexing units for automated information retrieval. *Proceedings of the 16th Annual Symposium on Computer Applications in Medical Care* (pp. 644–648). Baltimore: McGraw-Hill.

Kirby, M., & Miller, N. (1986). MEDLINE searching on Colleague: Reasons for failure or success of untrained users. *Medical Reference Services Quarterly, 5,* 17–34.

Lindberg, D., Humphreys, B., & McCray, A. (1993). The unified medical language system project. *Methods of Information in Medicine, 32,* 281–291.

Salton, G. (1991). Developments in automatic text retrieval. *Science, 253,* 974–980.

Salton, G., Buckley, C., & Smith, M. (1990). On the application of syntactic methodologies in automatic text analysis. *Information Processing and Management, 26,* 73–92.

Schamber, L., Eisenberg, M., & Nilan, M. (1990). A re-examination of relevance: Toward a dynamic, situational definition. *Information Processing and Management, 26,* 755–776.

Sewell, W., & Teitelbaum, S. (1986). Observations of end-user online searching behavior over eleven years. *Journal of the American Society for Information Science, 37,* 234–245.