Information Retrieval at the Millenium

William Hersh, M.D.
Division of Medical Informatics and Outcomes Research
Oregon Health Sciences University
Portland, OR

Information retrieval systems were among the first medical informatics applications, yet their use has changed substantially in this decade with the growth of end-user computers and the Internet. While early challenges revolved around how to increase the amount of information available in electronic form, more recent challenges center on how to manage the growing volume. Traditional information retrieval issues – such as how to organize and index information to make it more retrievable as well as how to evaluate the effectiveness of systems – are still as pertinent as ever.

Introduction

Information retrieval (IR) systems – that is, systems to catalog and provide information about documents – were among the first applications of computers. Their potential for organizing and allowing access to the medical literature was recognized by the National Library of Medicine (NLM), and by 1966, the MEDLINE database was launched [1]. At that time, only specially-trained intermediaries could access MEDLINE, and to do so required mailing a search statement that had a several week turnaround time.

As we enter the 21st century, the access to on-line medical information has changed substantially. A large spectrum of medical information is now available electronically – not only journal references, but full-text journal literature, textbooks, image collections, and all sorts of other resources. MEDLINE can be searched instantaneously and for free by anyone in the world with access to the Internet.

This paper will review the state of medical IR systems at the millenium. In particular, two questions will be addressed:

1. What is new in the IR marketplace, with a particular focus on how the fruits of research now benefit commercially available content and systems?
2. What is new in IR research, focusing on approaches that demonstrate real-world utility to users?

Background

This paper assumes a basic familiarity with the tenets of IR. A variety of resources exist for achieving such familiarity:

1. A textbook on medical IR [2].
2. Well-known general textbooks on IR [3-5].
3. How-to guides for accessing on-line medical information [6, 7].

This section, however, will provide an abbreviated overview to define the core principles and terminology.

The ultimate goal of using an IR system is to retrieve *documents* containing information. While documents at the millenium may be electronic and contain images, sounds, and other multimedia elements in addition to text, the focus of IR systems is still largely the retrieval of text. In order for the user to retrieve documents, he or she must enter *queries* requesting documents. Most IR system queries consist of typing text at a keyboard, but queries in the next millenium may be entered by voice or manipulation of graphical icons. In order for queries to be matched to documents, there must be an *indexing language*, which is a set of descriptors that describe the contents of documents and can be entered by users to retrieve them. A *search engine* is the computer program that uses the indexing language to match queries and documents for the user.

There are two intellectual processes in the IR process, indexing and retrieval. These processes require human intelligence, either to carry them out directly or to develop computer programs that perform them. *Indexing* is the process of assigning terms from the indexing language to the document, whereas *retrieval* is the procedure of entering terms to obtain documents or their surrogates. These processes will be discussed next, followed by an overview of the means by which IR systems are evaluated.

Indexing

There are two general approaches to indexing, which are often labeled human and automated. *Human*

indexing typically consists of the assignment of indexing terms from a controlled vocabulary by a specially-trained indexer. The indexing language consists of the terms from the controlled vocabulary. The largest human indexing operation in the medical domain is performed by the NLM for MEDLINE. Using terms from the Medical Subject Headings (MeSH) vocabulary [8], indexers follow a protocol to assign about 5-10 terms per article [9].

The second major approach to document indexing is *automated* or *word* indexing. This approach is done strictly by computer programs that identify and designate all words in the document as indexing terms. Some systems filter out common words that have little retrieval value from a *stop list* (e.g., the, of). Some systems also perform *stemming*, whereby plurals and common suffixes are removed from words (e.g., *coughs* and *coughing* are reduced to *cough*). After stop word removal and stemming, the remaining word stems comprise the indexing language in the automated indexing approach.

An increasing number of IR systems employ an additional step beyond simple word recognition, which is term weighting. This process, originated by Salton in the 1960s but not adopted widely until the 1990s, is usually coupled with natural language queries and relevance ranking, which will be described below [10]. A variety of weighting measures have been implemented over the years, but the one which has shown the greatest general performance is the IDF*TF scheme. In this approach, terms in documents are given weight based on how frequently they occur in the document (term frequency or TF) and how infrequently they occur in others (inverse document frequency or IDF).

Retrieval

There are likewise two major approaches to retrieval, Boolean and natural language. Each type of retrieval can be used with either type of indexing term, controlled vocabulary or word. In *Boolean* searching, the user enters terms and connects them via the Boolean operators AND and OR. The AND operator returns documents that contain all of the specified terms, whereas the OR operator returns those that contain any of them.

Natural language retrieval does not use Boolean operators. Instead, the user enters a natural language query statement. This approach is typically coupled with term weighting, which through the techniques pioneered by Salton, allows documents not only be to matched, but also ranked for "relevance" [10]. In the process of *relevance ranking*, retrieved documents are sorted by presence and frequency of query terms, or some variation thereof.

Evaluation

Understanding the effectiveness of IR systems is important not only for researchers but also for those who use and purchase them. All of these individuals must have effective means for knowing how well such systems are suited for their task and how they can be improved. The most commonly used evaluation measures are recall and precision.

Recall and precision are based on the notion of documents being relevant to an information need [11]. *Recall* is the proportion of relevant documents in a collection that are retrieved (sensitivity). *Precision* is the proportion of retrieved documents in a search that are relevant (positive predictive value). Recall and precision are often measured using a test collection of known queries, documents, and relevance judgments.

For systems that perform relevance ranking, a table or graph combining recall and precision can be derived. In this approach, precision is measured at fixed intervals of recall, e.g., 0, 0.1, 0.2, and so forth up to 1.0. This allows aggregate measures combining recall and precision to be developed. The most common aggregate measures used are average precision of the fixed points of recall, or precision at some number fixed of documents retrieved, e.g., 20.

These approaches are very easy to use with "batch" studies that require no user. Some authors have criticized over-reliance on recall and precision as well as the use of test collections in general [12, 13]. It is unclear, for example, whether small differences in recall and precision have any effect on a user's ultimate success at searching.

What is new in the IR marketplace?

Now that a basic overview of IR has been given, various trends in the IR marketplace can be described. Four trends can be gleaned from an overview of the marketplace: the growth of "free" resources, the emergence of aggregated and/or synthesized resources, the development and evolution of Web search engines, and the adaptation of techniques from research systems into commercial ones.

Free resources

A variety of Web-based medical resources are now available without charge. The most notable among these is MEDLINE. In June, 1997, the NLM announced that Web-based MEDLINE would be available for free on its Web site. As two Web-based MEDLINE systems had been developed – PubMed (www.ncbi.nlm.nih.gov/PubMed/) and Internet Grateful Med (igm.nlm.nih.gov) – both were made available without charge.

Both systems provide a number of innovative searching features. PubMed provides a very simple, easy-to-use interface. It allows relevance feedback, whereby users can select a retrieved reference and obtain more references with similar MeSH terms. PubMed also implements queries that provide the "best evidence" for certain types of clinical questions, based on research by Haynes et al. [14]. It also establishes a mechanism to provide direct links to the full text of references when available elsewhere on the Internet. Internet Grateful Med provides more databases than PubMed and implements user feedback via the COACH system [15].

The NLM is not the only health-related government agency to use the Web for dissemination of free information. Others, including the Centers for Disease Control and Prevention (www.cdc.gov), the Food and Drug Administration (www.fda.gov), and the National Cancer Institute (www.nci.nih.gov) have also made databases and other information available. A forthcoming clearinghouse of clinical practice guidelines is being developed by the Agency for Health Care Policy Research (www.guideline.gov). There is also a government resource for consumer health information, HealthFinder (www.healthfinder.gov).

Most consumer-oriented Web sites have also adopted the approach of providing information for free and supporting themselves by advertising. Most information for health care providers, on the other hand, still requires payment. This is not surprising, since the cost of producing this information is high and the market for it is limited, at least compared with consumer-oriented information. Only a handful of clinical sites have adopted the free-with-advertising model, one of which is Medscape (www.medscape.com).

Some have expressed concern about the quality of free information on the Web. Silberg et al. have suggested standards for Web-based health information [16], including the presence on all health-related Web sites of each page's:

- Authorship – names, affiliations, and credentials
- Attribution – references, sources, and (where appropriate) copyright
- Disclosure – potential and real conflicts of interest
- Currency – dates content posted and updated

Aggregation and synthesis of content

Many medical publishers have focused on aggregating content. Most of these products began as CD-ROMs which are now being adapted to the Web. Two well-known products that aggregate medical textbooks are *Stat!-Ref* (Teton Data, Jackson, WY) and *Harrison's Plus* (McGraw-Hill, New York, NY). Some newer products that are only Web-based include *Primary Care Online* (Lippincott, Williams, and Wilkins, Philadelphia, PA) and *MDConsult* (St. Louis, MO). Another form of aggregated on-line information increasingly available is the linkage of bibliographic references with full text journal articles.

Clinicians have always preferred not only aggregated information, but synthesized sources as well. One of the oldest paper-based products, now available on the Web, is the *Yearbook Series* (Mosby, St. Louis, MO). In recent years, new content has appeared based on the evolution of evidence-based medicine (EBM) [17]. While the initial approach to EBM focused on literature retrieval and its critical appraisal by the clinician, advocates have found that routine searching of the literature by clinicians is impractical and that clinicians have limited skills in these areas. As a result, publishers and others have developed synthesized content based on EBM approaches.

One of the first forms of evidence-based synthesized content is the extended structured abstract. The first product to do this was *ACP Journal Club* (American College of Physicians, Philadelphia, PA), which focused on internal medicine. Another journal, *Evidence-Based Medicine* (American College of Physicians and British Medical Journal, London, UK), extends this approach to all of health care. A similar product, *Evidence-Based Practice*, has been developed by Appleton & Lange (Stamford, CT) which features shorter synopses but emphasis on patient-oriented (e.g., mortality, symptom reduction) as opposed to disease-oriented (e.g., test result improvement) evidence.

Another emerging form of evidence-based synthesized content is the *systematic review*. These reviews are different from ordinary review articles,

which are not often as comprehensive and methodologically sound as they could be. Systematic reviews are instead based on exhaustive literature review and advanced statistical methods, including meta-analysis [18]. The best known producers of systematic reviews are the Cochrane Collaboration [19] and the Agency for Health Care Policy Research Evidence-Based Practice Centers. An on-line version of EBM resources, *Evidence-Based Medicine Reviews*, has become available from Ovid Technologies (New York, NY).

The goal of the Cochrane Collaboration is to produce a database of systematic reviews on all interventions in health care (www.cochrane.org). Their work is predicated on the principle that the best means to assess the efficacy of a health care intervention is the randomized controlled trial. Their approach includes an exhaustive search for all trials, including those published in foreign language journals or not published at all. Meta-analysis is used where appropriate to aggregate results from like trials. The reviews are updated as results of new trials become available.

There are limitations of the Cochrane approach. From an economic standpoint, it is not clear whether there is a sustainable market for reviews that can fund the infrastructure needed to develop and maintain reviews. There is also a "chicken and egg" problem in that the Cochrane database does not currently have enough reviews to be clinically useful, hence users will not purchase it. Another limitation is the reliance on volunteer authors of reviews, whom have other constraints on their time competing with the ongoing maintenance of their reviews. Finally, randomized controlled trials and meta-analysis have limitations [20].

Web search engines

Unlike MEDLINE and CD-ROM textbooks, the Web is not a single database. Rather, it is a dynamic mass of information that changes every time someone adds, deletes, or changes a page. Therefore there can be no search system that can search the entire Web. Search systems for the Web have taken two forms: Web crawlers and filtering-classifying systems.

Web crawlers

Web crawlers index the words on Web pages. Starting at a *seed* site, pages are identified by following the links from these to other pages. There is no discrimination of the information that is indexed; everything encountered is added to the database. Most Web crawlers are queried by natural language searching with relevance ranking of the output. The Search buttons on Netscape and Internet Explorer take users to a page that provides access to most of the Web search engines.

Some well-known Web crawlers include

- AltaVista – altavista.digital.com
- Infoseek – www.infoseek.com
- Excite – www.excite.com
- HotBot – www.hotbot.com

Filtering and classifying systems

Some Web search engines take a different approach. They filter information, based on certain criteria such as "quality" or "clinical relevance," and/or use classification schemes, such as MeSH, to provide better indexing.

Some systems provide elaborate filtering:

- Medical Matrix – www.medmatrix.org
- Yahoo Health – www.yahoo.com

Others provide both filtering and classifying:

- CliniWeb – www.ohsu.edu/cliniweb/
- Medical World Search – www.mwsearch.com

In CliniWeb, for example, pages that are written to the level of health care students and above are indexed with terms from a subset of the MeSH vocabulary [21].

Adaptation of research techniques

A variety of IR techniques formerly viewed as "research" are now used widely in commercial systems. For example, most Web search engines use natural language searching with relevance ranking that was pioneered by Salton in the 1960s [10]. Other research techniques that have been adopted include relevance feedback and query augmentation.

In *relevance feedback*, the system adds new search terms to the query based on those present in documents that the user has designated as relevant [22]. The most common approach used on the Web (e.g., PubMed, Excite) is to allow the user to select "more documents like this one." These systems add words (Excite) or MeSH terms (PubMed) from relevant documents to the query.

Query augmentation is the process of presenting the user with suggested words or phrases from retrieved documents that are not in the original query that the user may add (e.g., [23]). In Excite, the list of suggested terms appears across the top of the page.

This approach can also be limited to terms from documents that are designated relevant.

What is new in IR research?

Despite the growth of commercial IR applications, research into new methods is thriving as well. One of the major events fueling research is the Text Retrieval Conference (TREC). A variety of other research methods are being developed outside of TREC, as are new approaches to evaluating how well systems perform.

TREC experiments

Organized by the National Institutes for Standards and Technology (NIST), the TREC conference is designed to allow different research groups to work on a common large test collection [24]. While not designed to be a "competition," it does allow various groups to compare their techniques with others. The test collection consists of several gigabytes of newswire, computer publications, and government reports, along with a set of real-world queries and relevance judgements regarding which documents are relevant to the queries. A Web site provides more details, including proceedings from past conferences (trec.nist.gov).

Logistics of TREC

TREC was initiated in 1992 and has been held annually since then. The first six conferences were organized around two major tasks, ad hoc retrieval and routing. The ad hoc task was the typical retrieval task, with queries searching against a database of unknown documents. The goal of the routing task, on the other hand, was to identify new documents based not only on queries but also on documents previously identified as relevant. This was a variation of the relevance feedback task, where systems could use terms from known relevant documents to improve their queries. With the TREC-7 conference in 1998, routing will be eliminated as a major task.

For each task, there is a database of content consisting of about two gigabytes of text (one-half to one million documents) and 50 queries that contain a statement of information need and a definition of what constitutes a relevant document. Each participating group submits two runs consisting of ranked documents for each query. All documents in any group's top 100 documents are submitted to relevance judges, with the remainder assumed not relevant. NIST then prepares recall-precision results for each group.

In addition to the major tasks, a number of "tracks" have developed with focus on specific areas. These include:

- Interactive – Since most of the TREC experiments are system-oriented, this track aims to focus more on the user.
- Natural language processing (NLP) – For groups using NLP techniques, this track provides an opportunity to focus on such approaches.
- Very large corpus – Some groups are interested in issues related to extremely large collections, to which this group has access.
- Filtering – A variant of routing that requires a binary instead of ranked decision on document selection.

What has been learned in TREC?

While there is some concern that the batch-style experiments of TREC and its focused subject domain limit the generalizability of the results obtained, it does provide a platform for assessing diverse approaches to IR. Furthermore, results from similar approaches implemented across different systems have shown consistency. The approaches from TREC that consistently improve results have been new weighting algorithms, passage retrieval, and query expansion. Each of these techniques improve average precision in batch runs by about 10-20%. They have been implemented in a variety of systems whose underlying approach is different (e.g. [23, 25]).

The new weighting techniques showing the most benefit include *2-Poission term weighting* [26] and *pivoted normalization* [27]. Both of these measures "normalize" document length so that longer documents do not get higher relevance scores solely due to their weight.

Passage retrieval was pioneered by Salton about ten years ago [28]. It is based on the premise that for full-text documents, information sought in a query is likely to be in one or more particular parts of a document. Passage retrieval is claimed to reduce linguistic ambiguity. In passage retrieval, full-text documents are broken into *passages*. Each passage is assigned a weight, with the highest-weighting passage(s) contributing to the documents overall weight. TREC results show that the best passages are overlapping fixed-length windows, not those based on any of the usual document structure (e.g., paragraphs) [29].

Query expansion has been tried for many years (e.g., [30]), but it did not appear to show benefit until the TREC experiments [23, 25], perhaps due to individual documents and the collection as a whole being much larger than those used in previous experiments. In query expansion, the top ranking documents are all assumed relevant and used in a relevance feedback process. Unlike relevance feedback, the initial query can be used, with the process requiring no action by the user.

The TREC experiments have also identified approaches that do not improve results, at least in the context of these experiments. One of these approaches is NLP. Due to the large database, the traditional complete NLP approach is impractical in this setting. A number of groups have adopted "partial" NLP approaches that focus on identification of noun phrases and use of thesauri [23, 31]. None of these techniques, however, provide improved performance over the beneficial techniques described above, and they often do worse.

Other medically-oriented IR research of note

A variety of other important medical IR research findings have emerged in recent years. These studies have demonstrated the benefit of MeSH terms for searching MEDLINE, the value of query expansion in the medical domain, and the demonstration that conventional systems can be enhanced by new techniques.

Hersh and Hickam found that users in an internal medicine clinic showed little benefit from some of the advanced retrieval features associated with MeSH indexing (e.g., subheadings, explosions), [32]. However, they did show that having the words from the MeSH terms available in the MEDLINE record to search against improved average precision in batch-type studies over just the title and abstract [33]. Srinivasan has also demonstrated that MeSH terms in MEDLINE records improves searching performance, which can be enhanced even more by query expansion using those terms [34].

A variety of research projects at the NLM have identified promise for improving MEDLINE retrieval. For example, an expert system to improve indexing may assist indexers, making their assignments more consistent [35]. On the retrieval side, the COACH expert system has been implemented within Grateful Med to help users diagnose faulty searches [15]. And of course the UMLS Project has the potential to offer a wider coverage of vocabulary terms for indexing and retrieval [36].

New approaches to evaluation

Another important area of IR research attracting increased interest is in evaluation methodology itself. As noted above, a number of investigators have questioned the traditional approach of recall and precision [12, 13]. They have noted that relevance judgements are often inconsistent. It has also been asserted that recall and precision were developed with a library orientation but may not be applicable to other areas where IR systems are used, such as the busy clinical setting. Finally, the most important question concerning IR systems is whether information leads the user to make better decisions or have better outcomes of care.

It should be noted that evaluating IR systems, especially in operational settings, can be a very difficult task. Unlike other informatics applications, such as knowledge-based decision support systems, the questions users pose may be diverse and/or vague. Health care decisions are also very complex and the IR system may only play a small or peripheral role in answering the question. Finally, as in all research studies, laboratory approaches control extraneous variables but introduce elements of unreality.

A number of new approaches to evaluation have been undertaken. One approach for operational system observations has been the use of surrogate measures for clinical outcomes. Wyatt, for example, advocates using the clinician decision and not the patient outcome, which requires a smaller sample size to detect an impact, since the right decision does not improve every outcome [37]. Another technique, employed in laboratory evaluation, has been to assess a user's ability to complete a task. Hersh has focused on the ability of users to answer clinical questions [38, 39]. Further work is looking at factors associated with successful use of the system.

Future directions for IR

This decade has seen an explosion in growth of commercial IR systems and research on their use. While the Web provides a common platform to facilitate the use of IR systems, there are a number of issues that still hamper their effective use. Further improvements in content, systems, and their evaluation are essential. There must be continued commercial development as well as "no strings

attached" research to identify best approaches for systems and users.

Acknowledgements

The author would like to thank librarians Lynetta Sacherek and Andrea Ball as well as graduate student Susan Price, all members of his research staff, for their valuable comments and suggestions.

References

1. Miles, W., *A History of the National Library of Medicine: The Nation's Treasury of Medical Knowledge*. 1982, Bethesda, MD: U.S. Dept. of Health and Human Services.
2. Hersh, W., *Information Retrieval: A Health Care Perspective*. 1996, New York: Springer-Verlag.
3. Salton, G., *Introduction to Modern Information Retrieval*. 1983, New York: McGraw-Hill.
4. Pao, M., *Concepts of Information Retrieval*. 1989, Englewood, CO: Libraries Unlimited.
5. Meadow, C., *Text Information Retrieval Systems*. 1992, San Diego: Academic Press.
6. Feinglos, S., *MEDLINE: A Basic Guide to Searching*. 1985, Chicago: Medical Library Association.
7. Albright, R., *A Basic Guide to Online Information Systems for Health Care Professionals*. 1988, Arlington, VA: Information Resources Press.
8. Lowe, H. and G. Barnett, *Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches.* Journal of the American Medical Association, 1994. 271: 1103-1108.
9. Bachrach, C. and T. Charen, *Selection of MEDLINE contents, the development of its thesaurus, and the indexing process.* Medical Informatics, 1978. 3(3): 237-254.
10. Salton, G., *Developments in automatic text retrieval.* Science, 1991. 253: 974-980.
11. Hersh, W. and D. Hickam, *How well do physicians use electronic information retrieval systems? A framework for investigation and review of the literature.* Journal of the American Medical Association, 1998: in press.
12. Swanson, D., *Historical note: Information retrieval and the future of an illusion.* Journal of the American Society for Information Science, 1988. 39: 92-98.
13. Hersh, W., *Relevance and retrieval evaluation: perspectives from medicine.* Journal of the American Society for Information Science, 1994. 45: 201-206.
14. Haynes, R., *et al., Developing optimal search strategies for detecting clinically sound studies in MEDLINE.* Journal of the American Medical Informatics Association, 1994. 1: 447-458.
15. Kingsland, L., *et al., COACH: applying UMLS knowledge sources in an expert searcher environment.* Bulletin of the Medical Library Association, 1993. 81: 178-183.
16. Silberg, W., G. Lundberg, and R. Musacchio, *Assessing, controlling, and assuring the quality of medical information on the Internet: caveat lector et viewor - let the reader and viewer beware.* Journal of the American Medical Association, 1997. 277: 1244-1245.
17. Sackett, D., *et al., Evidence-Based Medicine: How to Practice and Teach EBM*. 1997, New York: Churchhill Livingstone.
18. Mulrow, C. and D. Cook, eds. *Systematic Reviews: Synthesis of Best Evidence for Health Care Decisions*. 1998, American College of Physicians: Philadelphia.
19. Bero, L. and D. Rennie, *The Cochrane Collaboration: preparing, maintaining, and disseminating systematic reviews of the effects of health care.* Journal of the American Medical Association, 1996. 274(1935-1938).
20. Feinstein, A., *Meta-analysis: statistical alchemy for the 21st century.* Journal of Clinical Epidemiology, 1995. 48: 71-79.
21. Hersh, W., *et al., CliniWeb: managing clinical information on the World Wide Web.* Journal of the American Medical Informatics Association, 1996. 3: 273-280.
22. Salton, G. and C. Buckley, *Improving retrieval performance by relevance feedback.* Journal of the American Society for Information Science, 1990. 41: 288-297.
23. Evans, D. and R. Lefferts. *Design and evaluation of the CLARIT TREC-2 system*. in *The Second Text REtrieval Conference (TREC-2)*. 1993. Gaithersburg, MD: NIST: 137-150.
24. Harman, D., *Overview of the second Text REtrieval Conference (TREC-2).* Information Processing and Management, 1995. 31: 271-289.
25. Buckley, C., *et al. Automatic query expansion using SMART: TREC 3*. in *Overview of the Third Text REtrieval*

Conference (TREC-3). 1994. Gaithersburg, MD: NIST: 69-80.

26. Robertson, S. and S. Walker. *Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval*. in *Proceedings of the 17th Annual International ACM Special Interest Group in Information Retrieval*. 1994. Dublin: Springer-Verlag: 232-241.
27. Singhal, A., C. Buckley, and M. Mitra. *Pivoted document length normalization*. in *Proceedings of the 19th Annual International ACM Special Interest Group in Information Retrieval*. 1996. Zurich, Switzerland: ACM Press: 21-29.
28. Salton, G. and C. Buckley, *Global text matching for information retrieval.* Science, 1991. 253: 1012-1015.
29. Callan, J. *Passage level evidence in document retrieval*. in *Proceedings of the 17th Annual International ACM Special Interest Group in Information Retrieval*. 1994. Dublin: Springer-Verlag: 302-310.
30. Smeaton, A., *The retrieval effects of query expansion on a feedback document retrieval system.* The Computer Journal, 1983. 26: 239-246.
31. Strzalkowski, T., J. Carballo, and M. Marinescu. *Natural language information retrieval: TREC-3 report*. in *Overview of the Third Text REtrieval Conference (TREC-3)*. 1994. Gaithersburg, MD: NIST: 39-53.
32. Hersh, W. and D. Hickam, *The use of a multi-application computer workstation in a clinical setting.* Bulletin of the Medical Library Association, 1994. 82: 382-389.
33. Hersh, W., *et al. OHSUMED: an interactive retrieval evaluation and new large test collection for research*. in *Proceedings of the 17th Annual International ACM Special Interest Group in Information Retrieval*. 1994. Dublin: Springer-Verlag: 192-201.
34. Srinivasan, P., *Retrieval Feedback in MEDLINE.* Journal of the American Medical Informatics Association, 1996. 3: 157-168.
35. Humphrey, S., *Medindex system: medical indexing expert system.* Information Processing and Management, 1988. 25: 73-88.
36. Humphreys, B., *et al., The Unified Medical Language System: an informatics research collaboration.* Journal of the American Medical Informatics Association, 1998. 5: 1-11.
37. Wyatt, J. and D. Spiegelhalter. *Field trials of medical decision-aids: potential problems and solutions*. in *Proceedings of the 15th Annual Symposium on Computer Applications in Medical Care*. 1991. Washington, DC: McGraw-Hill: 3-7.
38. Hersh, W., *et al. Towards new measures of information retrieval evaluation*. in *Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care*. 1994. Washington, DC: Hanley-Belfus: 895-899.
39. Hersh, W., J. Pentecost, and D. Hickam, *A task-oriented approach to information retrieval evaluation.* Journal of the American Society for Information Science, 1996. 47: 50-56.