Towards New Measures of Information Retrieval Evaluation

William R. Hersh
Diane L. Elliot
David H. Hickam
Stephanie L. Wolf
Oregon Health Sciences University
Portland, OR

Anna Molnar
Christine Leichtenstien
University of Ulm, Ulm, Germany

Abstract

All of the methods currently used to evaluate information retrieval (IR) systems have limitations in their ability to measure how well users are able to acquire information. We utilized an approach to assessing information obtained based on the user's ability to answer questions from a short-answer test. Senior medical students took the ten-question test and then searched one of two IR systems on the five questions for which they were least certain of their answer. Our results showed that pre-searching scores on the test were low but that searching yielded a high proportion of answers with both systems. These methods are able to measure information obtained, and will be used in subsequent studies to assess differences among IR systems.

Introduction

As information retrieval (IR) systems proliferate, it is necessary to assess their benefit to users. The most common approach for evaluating operational IR systems has been to measure usage frequency and/or user satisfaction. While usage frequency is easy to measure, it provides no insight into why the system was used or how successful the user was in finding information. Likewise, user satisfaction does not elucidate how users interact with or benefit from IR systems. Thus while systems installed in academic medical settings free to the user have generally been well-received (i.e., [1-3]), it has also been shown that over a third of community-based physician users stopped using the a microcomputer-based MEDLINE system during a several-year period [4], and that MEDLINE usage in a university hospital dropped by two-thirds when access fees were imposed [5].

The next level of retrieval evaluation has been to measure users' success at retrieving relevant documents using indices such as recall and precision. While these indices provide a starting point at determining the quantity of useful information obtained from an IR system, they say little about the quality of that information. It has never been proven, for example, that moderate differences in recall or precision (i.e., the 5-10% improvement seen in experiments such as TREC) have any effect on the overall success of a user's interaction with an IR system. Indeed, with ranked retrieval systems the differences may be solely due to ordering of the documents. Furthermore, when comparing two systems, while it may be possible to show statistical significance between the results (with a t-test or some other appropriate statistical measure), we have no idea whether the difference is "clinically" significant.

One of the reasons why recall and precision may not accurately reflect the quality of information obtained is that most technical literatures are both *redundant* and *contradictory*. The medical literature, for example, is redundant in that original clinical studies are often described in other documents, such as review articles or consensus reports. But the medical literature is also contradictory, particularly as new diagnostic and therapeutic approaches supersede old ones, such as in the case of treatment of hypercholesterolemia [6]. Thus on one hand it may only be necessary to retrieve one of many potentially relevant documents to obtain the right information, while on the other the user may be misled if the entire scientific picture over time is not retrieved.

A more fundamental problem with recall and precision is the subjective nature of relevance judgments. Not only is interobserver agreement in relevance judgments low [2, 7, 8], but judgments of relevance are influenced by factors such as document order and expertise of the judge [9, 10]. Meadow has argued that relevance is not fixed, but changes based on the users past and current knowledge as well as over time [11].

There are also some practical concerns in the use of recall and precision, especially in interactive settings. For example, what constitutes a retrieved document? While this is straight-forward in a batch-style retrieval evaluation, it can become problematic when a user is interacting with a system. The interactive experiments at TREC-3 showed that each of the four participating systems had different mechanisms for entering queries and displayed different portions of a document after a search [12-15]. Likewise, an earlier study of ours showed instances of users who started out with a poor search, retrieving a large number of

SIGIR'95 Seattle WA USA© 1995 ACM 0-89791-714-6/95/07.$3.50

nonrelevant documents, but later refined the search to retrieve many relevant documents [16]. In some cases, the poor search was just due to a typing error. Yet despite its ultimate success, the recall and precision values of the search were poor, since a document was considered retrieved if found by any query formulation during the search.

Based on these concerns with recall and precision, we explored in this study the feasibility of an alternative method to evaluate how well IR systems help users meet their information needs. Our approach was an adaptation of a method previously used to evaluate a hypertext statistical textbook [17], a historical encyclopedia [18], and a microbiology factual database [19]. The major difference in our study was the use of two different IR approaches, Boolean and natural language searching. The overall goal of this study was to assess how well medical students answered clinical questions with an IR system. The purpose was to determine whether this method could measure information acquisition and thus be used as a method to determine the effectiveness of user interaction with the system.

Methods

For this study we used two IR programs developed at Oregon Health Sciences University (OHSU). The first of these was SWORD, which features a natural language searching interface with relevance ranking [20]. With SWORD, the user enters a free-text query and retrieved documents are ranked based on the IDF*TF formula (see Figure 1). The second program was BOOLEAN, which utilizes a Boolean interface modeled after the NLM's Grateful Med system, where the words within each line are connected by logical OR, followed by the connection of each line with logical AND [7] (see Figure 2). Both programs eliminate stop words and use a simple stemming algorithm for indexing and user queries. They also log every interaction with the user, including submitting a query, selecting a document to view, and browsing other documents. The database searched by both programs was an electronic version of the textbook, *Scientific American Medicine* [21], divided into over 6,600 "documents" based upon the hierarchical structure of the print version.

Figure 1 -- The SWORD interface. After a query is entered and the **Find** button clicked, the words found in the database, not found in the database, and in the stop list are listed, along with the top 10 matching documents, weighted by IDF*TF. Additional documents are added to the matching documents list by clicking **More Documents**. The documents themselves are viewed by double-clicking on their titles.

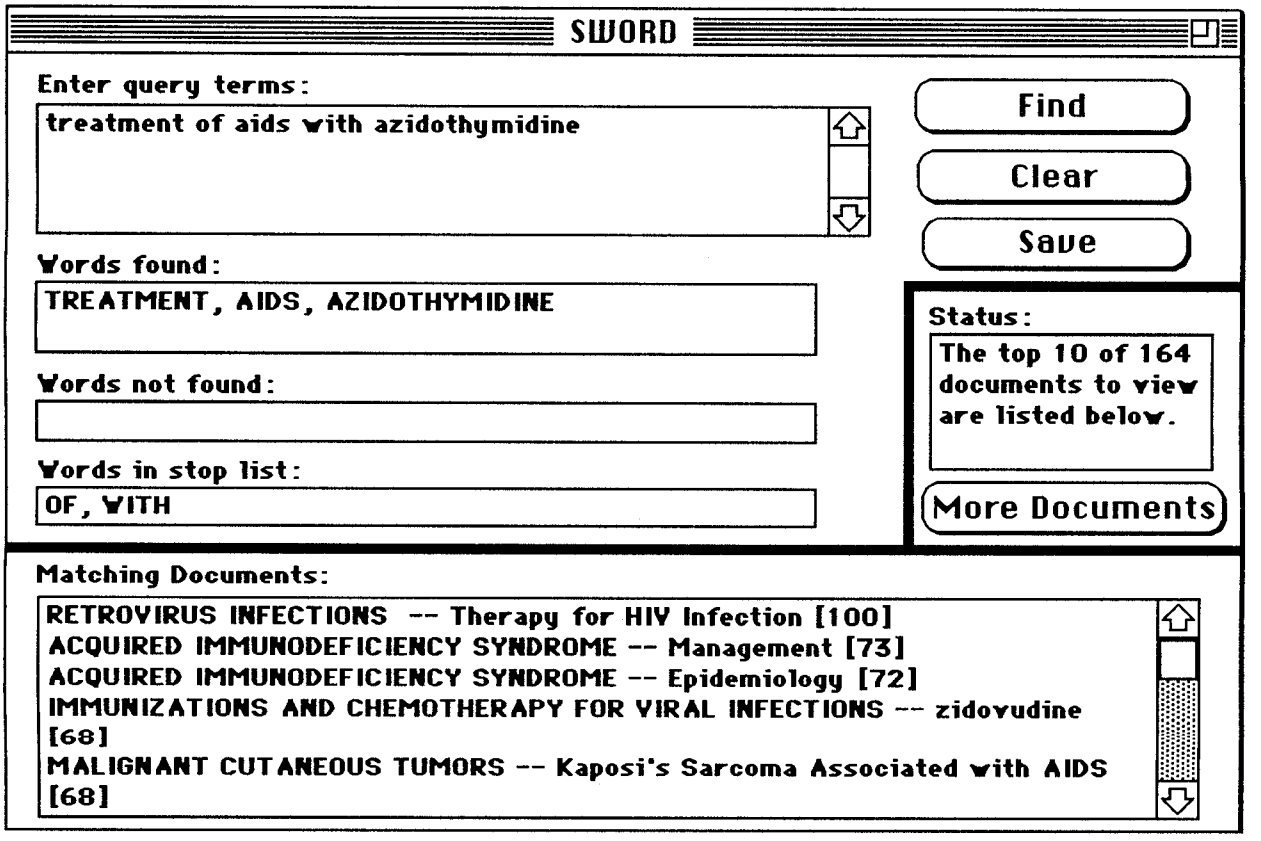
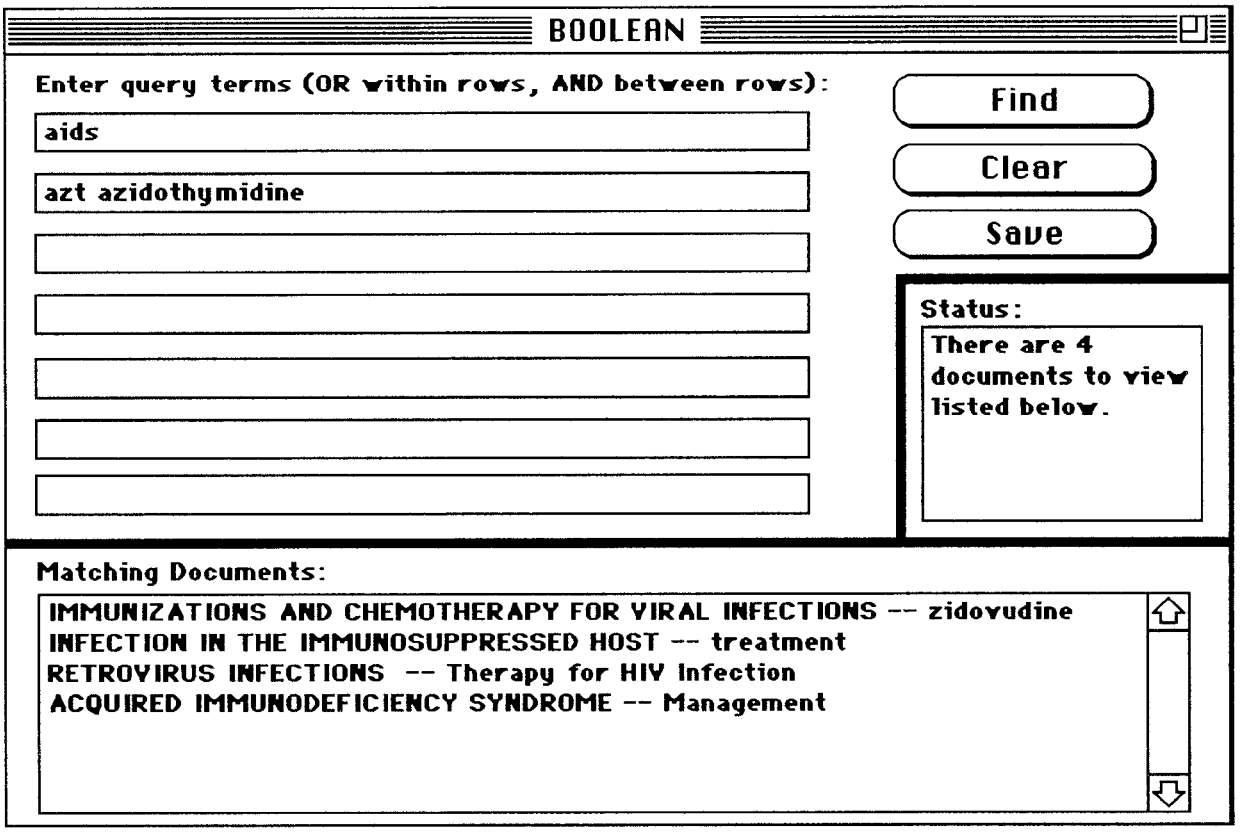


Figure 2 -- The BOOLEAN interface. Query terms are entered on each row, with OR performed on terms in the same row and AND performed between rows. After the **Find** button clicked, the matching documents are displayed. The documents themselves are viewed by double-clicking on their titles.



To measure information acquisition, a ten-question short-answer test at the senior medical student level of difficulty was developed (Table 1). The test questions were designed to have specific answers in the database, so that at least one document that provided the "answer" to each question. The test was given before and after searching, with the measurements of difference assessed by correctness of answers.

All medical students from the senior class at OHSU were sent a letter recruiting them to participate, of which 13 volunteered. Each student completed a brief questionnaire asking about prior computer experience, and we also obtained each student's class rank from the OHSU Dean's office. Both factors were used to stratify randomization of students.

The subjects spent a total of two hours in the experiment. After a brief introduction explaining the purpose of the experiment, they were given one-half hour to complete the ten-question test. At the completion of the test, they designated the five questions for which they had the least certainty about their answer. After a short break, they were oriented for 15 minutes to their computer and IR system, SWORD or BOOLEAN. Students then had up to 30 minutes to search for answers to the five questions for which they had greatest uncertainty about their original answers. They were required not only to answer each question, but also to give one or more document references that supported their answer.

The searching logs captured data about each query, including number of searches, total documents retrieved and viewed, and time taken. A *query* was defined as all of the interactions in attempting to find the answer to a question. A *search* was the entering of a search statement and retrieval of matching document titles. A document was considered *retrieved* if its title was in the list of document titles displayed after a search. A document was considered *viewed* if the user displayed the full text on the screen. For each user's query, we determined the number of searches, number of documents retrieved, and number of documents viewed. In addition to total number of searches, retrieved documents, and viewed documents for each query, we also calculated the number of each of these parameters required to reach an answer document.

Table 1: Ten questions for searching - answers in *italics*

1. A 60-year-old man from a poor socioeconomic environment is admitted with an acute illness characterized by mental disturbances, a sixth nerve palsy, and ataxia of gait. What specific emergency treatment is needed? *Thianine.*
2. What percent of patients with Type II diabetes respond to oral hypoglycemic agents as their initial drug treatment? *60-70%.*
3. Mr. Rogers is seen in the Bend, OR Emergency Room. He states that he was bitten by a 'spider.' He is relatively certain that it was a black widow. What are the expected initial symptoms of the bite? *Muscular pain and rigidity.*
4. What organism is most commonly found in anaerobic osteomyelitis? *Bacteroides.*
5. You are seeing a diabetic man with severe gastropariesis. He has not improved on oral metoclopramide (Reglan) and was sent to you for additional treatment. What would you recommend? *Suppository form of metoclopramide.*
6. What electrocardiographic feature distinguishes Prinzmetal's angina from more typical angina pectoris? *ST elevation.*
7. Mrs. Towel, an 80-year-old woman on no medication, is seen for light-headedness and found to have a heart rate of 36 and third degree heart block. What is the most likely etiology of her heart block? *Lenegre's Disease* or *age-related changes in A-V conduction system.*
8. A strongly positive antibody test to which antigen is most typical of Mixed Connective Tissue Disease? *Anti-RNP antibody.*
9. What is the most common cause of sudden death among young athletes? *Hypertrophic cardiomyopathy.*
10. How is the organism which causes Rocky Mountain Spotted Fever transmitted? *Tick bite.*

The tests were scored independently by two members of the study team (WRH and SLW), whose interobserver agreement was good (kappa = 0.71). To assess information acquisition, a pre-test/post-test analysis was used. A McNemar's Test was performed for each test question, using data from those subjects who answered that question on the post-test.

Results

A total of 13 subjects participated, six of whom used BOOLEAN and seven of whom used SWORD. There were no significant differences between the BOOLEAN and SWORD groups in computer experience or class rank. The average number correct on the initial ten-question test was 1.2, with no statistically significant difference between groups. The average number correct for the five questions searched upon was 4.0, again with no significant differences between groups (Table 2). Because there were no differences in general user characteristics or answers between the programs, the data were then pooled to determine information acquisition. Four of the ten questions showed a statistically significant difference in information found when using a searching program, while four others had a trend towards significance (Table 3).

Table 4 compares all of the questions in terms of searches done, documents retrieved, and documents viewed for each question, both in total as well as number required to retrieve an answer document. The majority of answer documents were found on the first search, within the top ten documents retrieved, and on the first document viewed.

We also performed a failure analysis of questions where the wrong answer was obtained, or where there was an unsuccessful retrieval or viewing (Table 5). Only four of the ten questions had any incorrect answers at all. The majority of these came from question 8, although almost all of those who got this question wrong retrieved the answer document, and over half viewed that document, indicating that perhaps it was a poorly worded question.

Discussion

The purpose of this study was to explore alternative methods of evaluating the performance of IR systems, based on ability to acquire information. Our results indicate that this approach is a viable alternative to measuring recall and precision, and may even be preferable, in that it indicates whether the searcher was able to use the system to find needed information. We discovered in this study that medical students were successful in using an on-line textbook via natural language or Boolean searching.

There were some limitations to both this study and the assessment of this methodology for IR system evaluation. First, we only looked at one type of query in one domain, which was the factual question in the medical domain. While the questions used were quite similar to the types of questions that typically arise in clinical practice [22], there are other types of information needs besides the factual question. In particular, medical practitioners sometimes have questions that are broader, have no specific answer, or have no answer at all.

Another limitation was that each question had only a single relevant document. While this is typical for a single volume textbook, such as the one used in this study, other electronic databases have the redundancy and inconsistency mentioned previously. Future studies using this approach will have to handle issues such as retrieval of documents with partial or conflicting answers.

Table 2: Test results for the search groups

	BOOLEAN	SWORD	Both
Number	6	7	13
Pre-Test Score (correct of 10)	1.8	1.6	1.7
Post-Test Score (correct of 5)	4.2	3.9	4.0

Table 3: Pre-Test/Post-Test results for each query

	Pre-Test		Post-Test		
Question	No. responses	% correct	No. responses	% correct	p
1	13	30.8	3	100	.08
2	13	23.1	6	83.3	.08
3	13	0	8	100	.005
4	13	23.1	9	100	.01
5	13	0	8	87.5	.008
6	13	0	12	100	.0005
7	13	0	4	25	.3
8	13	0	11	27.3	.08
9	13	15.4	1	100	.3
10	13	76.9	3	100	.08

Table 4: Searching results for all queries with both programs

Total searches done	
1	48
>1	17
Searches to find answer	
1st	51
After 1st	5
Not found	9
Total documents retrieved	
<=10	46
>10	19
Documents retrieved to find answer	
<=10	49
>10	7
Not found	9
Total documents viewed	
<=10	60
>10	5
Documents viewed to find answer*	
1	41
2-5	13
>6	5
Not found	6
Time per query (min.)	5.40

* There were three queries with answer documents viewed but not retrieved by searching due to answers being found by browsing through the database.

Table 5: Failure analysis. The number of incorrect answers for each question are listed (those with no incorrect answers are not shown), with a tabulation of whether the answer document was retrieved or viewed.

Question	Incorrect	Retrieved		Viewed	
		Yes	No	Yes	No
2	1	1	0	1	0
5	1	1	0	1	0
7	3	0	3	0	3
8	8	7	1	5	3
Total	13	9	4	7	6

One procedural limitation of the study was allowing subjects to choose only five questions to search. Not only did this make the statistical analysis more difficult, but it also made assessing the adequacy of some questions difficult, as only a few users searched on them. In our next study, we will have users search on all questions in order to better assess the value of all questions searched by the IR system.

In summary, as IR systems achieve more widespread use, it will be increasingly important to characterize all aspects of systems, from numbers of relevant documents retrieved to user satisfaction and, ultimately, how the system impacts the tasks it is being used to assist, such as the delivery of health care, the practice of law, or scientific research. Many parameters will require assessment to determine the appropriate systems for specific settings. The technique used in this paper shows promise in this regard.

Our next step will be to utilize this approach with different types of questions and databases. We are currently comparing two commercial MEDLINE systems that are used in the OHSU library, one of which features Boolean searching (*CD Plus*, CD Plus, Inc., New York, NY) and the other natural language searching (*Knowledge Finder*, Aries Systems, Inc., North Andover, MA) using this type of approach. In this study, we will also attempt to correlate results with conventional recall-precision analysis.

Acknowledgments

This work was supported by Grant LM 05307 from the National Library of Medicine. The authors also thank Scientific American (New York, NY) for providing the text of *Scientific American Medicine* for this study.

References

1. Horowitz GL, Jackson JD, Bleich HL. PaperChase: self-service bibliographic retrieval. Journal of the American Medical Association 1983;328:2495-2500.
2. Haynes RB, McKibbon KA, Walker CJ, et al. Online access to MEDLINE in clinical settings. Annals of Internal Medicine 1990;112(1):78-84.
3. Hersh WR, Hickam DH. The use of a multi-application computer workstation in a clinical setting. Bulletin of the Medical Library Association 1994:382-389.
4. Marshall JG. The continuation of end-user online searching by health professionals: preliminary survey results. Proceedings of the Medical Library Association Annual Meeting, 1990.
5. Haynes RB, Ramsden MF, McKibbon KA, Walker CJ. Online access to MEDLINE in clinical settings: impact of user fees. Bulletin of the Medical Library Association 1991;79:377-381.
6. Littenberg B. Technology assessment in medicine. Academic Medicine 1992;67:424-428.
7. Hersh WR, Hickam DH. A comparison of two methods for indexing and retrieval from a full-text medical database. Medical Decision Making 1993;13:220-226.
8. Hersh WR, Hickam DH. A performance and failure analysis of SAPHIRE with a MEDLINE test collection. Journal of the American Medical Informatics Association 1994:1:51-60.
9. Eisenberg M, Barry C. Order effects: a study of the possible influence of presentation order on user judgments of document relevance. Journal of the American Society for Information Science 1988;39:293-300.
10. Schamber L, Eisenberg MB, Nilan MS. A re-examination of relevance: toward a dynamic, situational definition. Information Processing and Management 1990;26:755-776.
11. Meadow CT. Text Information Retrieval Systems. Academic Press: San Diego, 1992.
12. Charoenkitkarn N, Chignell M, Golovchinsky G. Interactive exploration as a formal text retrieval method: how well can interactivity compensate for unsophisticated retrieval algorithms. In: Harman D, ed. The Third Text REtrieval Conference (TREC-3). Gaithersburg, MD: NIST, 1994:in press.
13. Koenemann J, Quatrain R, Cool C, Belkin NJ. New tools and old habits: the interactive searching behavior of expert online searchers using INQUERY. In: Harman D, ed. The Third Text REtrieval Conference (TREC-3). Gaithersburg, MD: NIST. 1994:in press.
14. Robertson SE, Walker S, Jones S, Hancock-Beaulieu MM, Gatford M. Okapi at TREC-3. In: Harman D, ed. The Third Text REtreival Conference (TREC-3). Gaithersburg, MD: NIST, 1994:in press.

15. Tong RM. Interactive document retrieval using TOPIC. In: Harman D, ed. The Third Text REtrieval Conference (TREC-3). Gaithersburg, MD: NIST, 1994:in press.
16. Hersh WR, Hickam DH. An evaluation of interactive Boolean and natural language searching with an on-line medical textbook. Journal of the American Society for Information Science 1994:in press.
17. Egan DE, Remde JR, Gomez LM, et al. Formative design-evaluation of Superbook. ACM Transactions on Information Systems 1989:7:30-57.
18. Mynatt BT, Leventhal LM, Instone K, Farhat J, Rohlman DS. Hypertext or book: which is better for answering questions? Proceedings of Computer-Human Interface 92, 1992:19-25.
19. deBliek R, Friedman CP, Wildemuth BM, et al. Database access and problem solving in the basic sciences. In: Safran C, ed. Proceedings of the 17th Annual Symposium on Computer Applications in Medical Care. Washington, DC: McGraw-Hill, 1993:678-682.
20. Hersh WR, Hickam DH, Leone TJ. Word, concepts, or both: optimal indexing units for automated information retrieval. In: Frisse M, ed. Proceedings of the 16th Annual Symposium on Computer Applications in Medical Care. Baltimore: McGraw-Hill, 1992:644-648.
21. Rubenstein R, Federman DD. Scientific American Medicine. New York: Scientific American, 1990.
22. Gorman PN, Ash J, Helfand M, Beck JR. Assessment of information needs of primary care physicians. Proceedings of the Third Annual American Medical Informatics Association Spring Congress, 1992:26.