OHSUMED: An Interactive Retrieval Evaluation and New Large Test Collection for Research

William Hersh, Oregon Health Sciences University
Chris Buckley, Cornell University
TJ Leone, Oregon Health Sciences University
David Hickam, Oregon Health Sciences University

Abstract

A series of information retrieval experiments was carried out with a computer installed in a medical practice setting for relatively inexperienced physician end-users. Using a commercial MEDLINE product based on the vector space model, these physicians searched just as effectively as more experienced searchers using Boolean searching. The results of this experiment were subsequently used to create a new large medical test collection, which was used in experiments with the SMART retrieval system to obtain baseline performance data as well as compare SMART with the other searchers.

1 Introduction

Information retrieval systems, once the purview of search intermediaries in libraries, are increasingly becoming tools for professionals, such as physicians, lawyers, and educators. This is a timely advent for these tools, since many professional fields are facing unprecedented growth in quantity of information, resulting in significant information management problems. This has led to increased advocacy for end-user retrieval applications that are fast and easy-to-use.

We undertook a study to evaluate the use of one such system in a medical practice setting, the General Medicine Clinic at Oregon Health Sciences University. An Apple Macintosh IIsi computer was installed with several applications, one of which was Knowledge Finder (KF) (Aries Systems, North Bedford, MA), a CD-ROM retrieval system featuring a primary care subset of the MEDLINE database covering 270 journals for five years. It features a "natural language" searching interface and relevance ranking based on the vector space model [1]. The usage frequency and user satisfaction of this system have been reported elsewhere [2]. We subsequently created a new large test collection from the searches. This collection, called OHSUMED, was used to conduct further experiments with the SMART retrieval system.

The main objectives for the experiments described here were to:
1. Compare the use of a vector space-like system with novice end-users in an operational setting.
2. Build a large medical test collection for further medical information retrieval research
3. Obtain baseline results from a well-known retrieval system (SMART) and compare its results with that of human searchers.

Interactive comparison of advanced retrieval systems has rarely been done. There have been many batch-style evaluations of experimental retrieval systems [3, 4] as well as studies using user-generated search strategies [5-9]. However, with the exception of the CIRT experiments [10], there have been no "head-on" comparisons of interactive Boolean and natural language/weighting systems.

KF does not feature all of the advanced features seen in systems like SMART (such as relevance feedback and automatic phrase generation), but it does provide an easy-to-use natural language interface that allows searchers to begin with a minimal amount of training. KF also allows traditional Boolean searching with the MeSH vocabulary as well as searching on specific fields in the MEDLINE record, such as journal titles, publication dates, and author names. Of these, only author names were used by any of our searchers.

The second objective, building a large test collection for medical information retrieval research, was designed in the spirit of the TREC Project [11], aiming to bring real world-sized test collections to the experimental retrieval systems. Although TREC is already providing researchers with such a test collection, other large collections must be developed in different domains to determine which results can be generalized.

The final motivation was to perform runs with SMART, generating baseline performance levels with the OHSUMED collection and carrying out some experiments comparing SMART with human searchers

using other systems. In particular, we analyzed variation in recall and precision while simulating the advantages that human searchers have, such as ability to control retrieval set size and modify the search based on terms in the documents.

2 Interactive search study methods

The Macintosh workstation was placed in the workroom of the General Medicine Clinic at Oregon Health Sciences University. Users were given minimal training with the system, as little as 10-15 minutes, although they were offered more if they desired it. Each was also given an instruction packet with screen pictures and examples, which was also available, along with user manuals for each of the applications, beside the computer. Most users had previous experience searching the MEDLINE database in different implementations.

2.1 Query generation and replication

User queries were captured by requiring an on-line questionnaire to be completed before an application could be launched. Searchers were required to give a brief statement of their patient and their information need. In addition, the KF application logged the natural language search statement. KF does not, however, log the documents that the user actually views. After the ten month period of the experiment, all KF searches were reviewed, and a subset consisting of all searches where the user entered adequate information for replication (more than three words each of patient information and topic specification) was designated. After elimination of duplicate topics and author searches, the subset contained 106 searches.

For replication of the searches, 11 medical librarians and 11 physicians experienced with MEDLINE were recruited. Each librarian had to be a reference librarian who used MEDLINE at least several times per week. Each physician had to use MEDLINE more than once a month for more than 2-3 years as well as be clinically active in an ambulatory setting. The replicated searches were randomized so that each was searched by two physicians and two librarians, with one of each pair searching with the full MEDLINE feature set and the other using just text words. Searching was done on the ELHILL computer of the NLM, a traditional command line-oriented Boolean system. Searchers were required to search the last six years of the database (current MEDLINE file and first back-file). A reference was considered retrieved if it was displayed to the screen or returned by their last search statement.

2.2 Relevance judgments

References retrieved in the MEDLINE searches that were not in the test database were discarded for performance assessment. Every reference in the test database retrieved for a given query was judged for relevance by physicians who were clinically active and were current fellows in general medicine or medical informatics or senior medical residents. The judges used the patient information and topic designation from the original searcher as the information need of the user, with part of the MEDLINE record (title, source, authors, abstract, and publication type) as the reference. They were encouraged to seek the original articles when relevance could not be assessed from the title and abstract. The reviewers were blinded as to which searcher retrieved which articles. Relevance was judged on a three-point scale: definitely relevant (article provided highly relevant information for clinician faced with the recorded patient data and information need), possibly relevant (article might provide useful information to the clinician), and not relevant (article did not provide any relevant information for this information need). Reviewers were asked to judge relevance from the standpoint of a clinician seeking an answer to the question posed. About 11% of the judgments were duplicated to assess interobserver reliability as measured by the kappa statistic.

3. Interactive searching results

Searching quality was assessed by the measures of relative recall (hereafter referred to as recall) and precision. The test database for the calculation of recall and precision consisted of all MEDLINE references from the 270 journals in the current KF subset spanning from 1987 through 1991, which contained 348,566 references.

The study design provided results for five searches on each query: the original KF search, two MEDLINE searches (librarian and clinician) using the full MEDLINE feature set (MeSH terms and text words), and two searches (librarian and clinician) using text words only. Searching performance was characterized for each of the five groups by the mean relative recall and precision for each of the queries.

Statistical analysis of the results was performed using repeated measures analysis of variance. Post-hoc t-tests for paired observations were performed using the Bonferroni correction.

3.1 Retrieval

Table 1 shows the total number of references and average number of references per query retrieved by the five searching groups for all of the queries, along with the proportion of those references that were in the test database. Only a quarter of all the references retrieved by replicated searches were in the test database. There were also a small number of references retrieved by KF (0.5%) which were not in the test database, due to journals that were added or dropped over time on the actual KF CD-ROM. A total of 15,859 references (149.6 per query) were retrieved by all searchers from the test database, 12,565 (118.5 per query) of which were unique query-reference pairs.

3.2 Relevance judgments

There were 8,714 (69.3%) references judged as not relevant, 2,053 (16.3%) judged as possibly relevant, and 1,798 (14.3%) judged as definitely relevant. The mean number of definitely relevant references per query was 17.0 (median 9, range 0-100). The mean number of possibly relevant references per query was 19.4 (median 13, range 0-81). There were 1,435 (11.4%) pairs judged in duplicate, with 1,003 agreements and 432 disagreements. Table 2 shows the interobserver agreement for duplicated relevance judgments. The kappa score of 0.41 was comparable with other experiments of this type [12-14]. The majority of the disagreements were relevance threshold disagreements; i.e., one judge marked relevant almost all of the other judge's relevant documents plus other additional documents.

There were five queries without definitely relevant references. Because recall and precision are meaningless when there are no relevant documents, and the SMART evaluation program used below ignores queries with no relevant documents, we eliminated these five queries from further analysis.

3.3 Recall and precision results

Recall and precision values for each group were calculated in two ways: using only definitely relevant references (DR) and using definitely plus possibly relevant references (D+PR) (Table 3). For both levels of relevance, the KF searchers obtained recall that was significantly ($p < .0001$) higher than each of the other search groups. In turn, precision was significantly higher ($p < .0001$) for each of the non-KF groups. Librarians obtained better recall and precision using the full MEDLINE feature set than with text words, but the differences between the two were not statistically significant. Physicians obtained better recall with text words and better precision with the full feature set, but these differences were also not statistically significant. Librarians using the full feature set did have significantly ($p < .02$) higher recall than physicians using the full feature set (but not text words) for D+PR articles.

Since the searchers using KF retrieved so many more references per search, an additional searching group was added to the table, which was the KF searches cut off at a maximum of 15 references per search (as opposed to the program's default operation of 100 references per search), which was the average number of references retrieved per query in the test database by all the non-KF searchers. With this definition of retrieval, recall was not significantly different from either librarian or expert physician searchers. However, the precision of the KF searches based only on the top 15 articles was significantly worse than searches using the full MEDLINE feature set conducted by librarians ($p < .01$ for DR articles and $p < .0001$ for D+PR articles) and by expert physicians ($p < .01$ for DR articles and $p < .02$ for D+PR articles).

3.4 Retrieval overlap

Since a number of studies have shown that searchers tend to find non-overlapping sets of relevant references, we also looked at the proportions of relevant references retrieved by one searcher, two searchers, and so on. Over half of the 1,798 DR references were retrieved by only one of the five searchers (Table 4). Another quarter were retrieved only by two searchers. Less than ten percent of all DR references were retrieved by four or five searchers. When the KF retrieval set was reduced to 15 references per search, the number of DR references not retrieved by any searcher was 23.9%. Table 5 shows the proportions of singly retrieved relevant references by search group which retrieved them. For the full KF retrieval set, over one-quarter of definitely relevant references were retrieved by KF alone. With the reduced KF set, librarians ranked highest for unique retrievals.

Table 1 -- References retrieved in test database.

Group	References retrieved	References retrieved per search	Percent of references in test database
Clinic Physicians	9470	89.3	99.5
Librarians-Full MEDLINE	7032	66.3	27.1
Librarians-Text words only	7027	66.3	25.7
Physicians-Full MEDLINE	4550	42.9	25.3
Physicians-Text words only	5897	55.6	26.6

Table 2 -- Interobserver agreement of definitely relevant (DR), possibly relevant (PR), and not relevant (NR) relevance judgments.

		Rater 1 DR	Rater 1 PR	Rater 1 NR
	DR	127	112	96
Rater 2	PR		97	224
	NR			779

Table 3 -- Mean recall (R) and precision (P) results for interactive searching.

Group	Mean number of documents retrieved	Definitely relevant only R	Definitely relevant only P	Definitely and possibly relevant R	Definitely and possibly relevant P
Clinic Physicians-using KF	88.9	68.2	14.7	72.5	30.8
Clinic Physicians-KF top 15	14.6	31.2	24.8	25.5	43.8
Librarians-Full MEDLINE	18.0	37.1	36.1	30.8	59.4
Librarians-Text words only	17.1	31.5	31.9	27.0	50.3
Physicians-Full MEDLINE	10.9	26.6	34.9	19.8	55.2
Physicians-Text words only	14.8	30.6	31.4	24.1	48.4

Table 4 -- Number of searchers who retrieved relevant references.

Relevant references retrieved by:	Full KF retrieval		Reduced KF retrieval	
0 searchers	n/a		429	(23.9%)
1 searcher	957	(53.2%)	788	(43.8%)
2 searchers	474	(26.4%)	355	(19.7%)
3 searchers	190	(10.6%)	121	(6.7%)
4 searchers	99	(5.5%)	73	(4.1%)
5 searchers	42	(2.3%)	32	(1.8%)

Table 5 -- Sources of retrieval among references retrieved only by one searcher.

Relevant references retrieved by only:	Full KF retrieval		Reduced KF retrieval	
Clinic Physicians-using KF	503	(28.0%)	110	(6.1%)
Librarians-Full MEDLINE	195	(10.9%)	282	(15.7%)
Librarians-Text words only	128	(7.1%)	173	(9.6%)
Physicians--Full MEDLINE	63	(3.5%)	93	(5.2%)
Physicians-Text words only	68	(3.8%)	130	(7.2%)

3.5 Interpretation of interactive searching results

This study looked at the quality of searching from two perspectives that had not been taken before: the comparison of a system like KF in the hands of novice searchers to traditional tools used by more experienced searchers, and the comparison of the full MEDLINE feature set to searches based only on title and abstract words. The significantly higher recall obtained by the clinic physicians, even at the cost of diminished precision, shows that end-users can effectively access the medical literature with tools such as KF.

There are some aspects of the study which could explain the significantly better recall by novice searchers besides just the value of KF's innovative approach. To begin with, both the experienced clinician and librarian searchers faced the handicap of having limited information, usually a brief sentence on the patient and information need. It has been shown that librarians perform better searches when they can interview the information seeker, getting more detail on the exact information request [15]. Second, the experienced searchers may have searched differently if their database only contained the references on the KF CD-ROM. While the 42.9 to 66.3 references retrieved per search indicated appropriate strategies for the database they were searching (six years of full MEDLINE), the experienced searchers may have broadened their searches if the initial retrievals led to just the 10.9 to 18.0 references per search present in the test database. Broadening the search could have led to higher recall. Indeed, the higher recall of the novices may have been a function of their much larger retrieval sets as shown by their more modest recall levels when their retrieval set was limited in quantity to the average size of the retrieval sets of the experienced searchers (Table 5).

The other previously unexplored perspective was the comparison of using the full MEDLINE feature set versus just the use of text words by the experienced searchers. The differences between the two types of searching were small and statistically insignificant, indicating that at least for the types of searches done in this study, the full MEDLINE feature set did not confer any strong advantage in retrieving relevant references. When comparing physicians with librarians in using the full feature set, there was a statistically significant advantage in recall for librarians, suggesting that advanced MEDLINE features are a tool of most benefit to librarians.

4 The test collection

The test collection is a subset of the MEDLINE database, which is a bibliographic database of important, peer-reviewed medical literature maintained by the National Library of Medicine (NLM). There are currently over seven million references in MEDLINE dating back to 1966, with about 250,000 added yearly. While the majority of references are to journal articles, there are also a small number of references to letters to the editor, conference proceedings, and other reports. About 75% of the references contain abstracts, while the remainder (including all letters to the editor) have only titles. Each reference also contains human-assigned subject headings from the 17,000-term Medical Subject Headings (MeSH) vocabulary.

4.1 Current collection

The OHSUMED collection contains 348,566 references, which are derived from the subset of 270 journals covered in the KF MEDLINE Primary Care product covering the years 1987 to 1991. The test collection contains the 101 queries that were generated by actual physicians in the course of patient care and had at least one DR document. Each query contains a brief statement about the patient, followed by the information need. The queries are generally terse, more like the older test collections and unlike the long queries from the TREC collection. The collection also contains the relevance judgments for two levels of relevance, as described above: definitely relevant (DR) and definitely or possibly relevant (D+PR).

The NLM has agreed to make this test collection available to information retrieval researchers provided they do not make it available in actual (clinical or library) settings, use it only for research, and inform the users that the database itself is out of date. The files are available via FTP by contacting the first author of this paper.

4.2 Future plans

There are future plans to enhance this collection in two directions. The first is to do at least one additional round of relevance judgments. The SMART experiments described below led to retrieval of documents

not already judged for relevance for a given query. Continued experimentation with the collection by other researchers will undoubtedly identify even more unjudged documents. In addition, there will likely be disagreements with the existing relevance judgments.

The second future plan is to build a parallel collection of the full text documents from as many as the references in the collection as possible. Full text is desirable both as a delivery vehicle to the end-user (who does not subsequently need to go to the library to obtain the article in the reference) and as a source for more text to process in indexing. Over a dozen major journals have been available in full-text electronic form since the beginning of the time interval of this test collection (1987), and the publishers are being contacted about their use. The publisher of one of these journals, Annals of Internal Medicine, has already agreed and delivered text.

5 Experiments with SMART

Additional experiments were run using SMART version 11 to obtain baseline performance levels with a well-known experimental retrieval system as well as to compare that system's performance with the human searchers in the original experiment. Five sets of experiments were carried out:

1. Comparison of SMART weighting approaches
2. Effect of documents retrieved by SMART which were not judged for relevance
3. Alternative query and document formulations
4. Relevance feedback
5. Comparisons of SMART and human searchers

Except for as noted in specific experiments, the queries for the SMART experiments were the statements of information need (IN) by the original searcher. Likewise, except as noted, the documents for the SMART experiments consisted of the title, abstract (when present), and MeSH heading fields. All of the experiments reported in this section were done using DR documents only. Experiments were also done with D+PR documents, but while the magnitude of recall and precision levels changed, their relative values did not. For all experiments, recall and precision is shown at levels of 5, 15, and 100 documents retrieved. The latter two levels represent the average level of non-KF and KF human retrieval in the original experiments.

5.1 Comparison of SMART weighting approaches

A comparison of 16 different document weighting methods by 8 query weighting methods (using the IN as the query) yielded a wide range of performance. (The usual SMART weighting triples were used, as shown in Table 6.) A sampling of these results is shown in Table 7. The best document-query weighting was ann.atn, indicating that the only beneficial weighting component for documents was the normalized non-logarithmic term frequency. The cosine normalization did not prove beneficial for documents. Given no normalization, the IDF factor could appear in the document or query without affecting results. However, including it in both the document and query led to mildly decreased effectiveness.

The lack of benefit for the IDF with documents parallels the results of SMART in TREC 2, where the IDF also was not of benefit. The IDF is, however, known to be beneficial with smaller document collections [3, 13], but its benefit seems to wane with larger collection size. The lack of benefit for the cosine normalization was clearly due to a peculiarity in MEDLINE, which in addition to journal citations also contains references to letters to the editor, which have no associated text beyond the title and for this collection were almost always deemed nonrelevant. Normalization leads to higher ranking of these "shorter" documents, markedly lowering recall and precision values. The non-letter documents are also all approximately the same length, as MEDLINE (until recently) imposed a limit of 250 words per abstract and most articles pushed this limit as much as possible. Thus there was no advantage to normalization even for the longer documents.

5.2 Effect of non-judged documents

One problem with the above experiment is one that commonly occurs when new retrieval systems are used with a test collection, in that many documents that have not been judged for relevance are retrieved. Table 8 shows the same runs as in Table 7, but with non-judged documents eliminated from recall-precision calculations. There are substantial improvements in both recall and precision, indicating the maximum possible detrimental effect that non-judged documents have on results. In SMART's performance with TREC (which also has this problem), the maximum detrimental effect after 100 potentially unjudged documents was between 2% and 5%. For the rest of the experiments (with

exceptions described below), we included non-judged (assumed nonrelevant) documents in recall-precision calculations.

5.3 Alternative query and document formulations

We also tested variations in both the query and document representations. The original queries contained statements about the patient (BIO) and IN. The results of this baseline query, using the best weighting from Table 7, are given in the first row of Table 9. The results in the next row show the value of adding the BIO statement to the query. This improved the results, but showed maximum benefit when the words in the patient information statement were given only 0.25 of the weight given to words in the IN statement. This is most likely due to the fact that patient information is helpful but not central to the information need.

For the next analysis, we used the original user's natural language query entered into the KF system itself (hereafter referred to as the KF statement). This also provided a modest performance benefit over the baseline at higher document cut-off levels. The major difference between the IN and KF statements was that the latter tended to be more terse. That is, the user typed more information into the pre-search questionnaire than the KF query box. This may be an artifact of the users' past experience with Boolean searching, where one must be careful in entering search terms to avoid output overload.

The one alternate document formulation appears in the last row of Table 9. In this experiment, we compared the results with and without indexing of the human-assigned MeSH terms. In the original experiments, we found that experienced searchers, especially physicians, did not achieve substantial benefit from using MeSH terms in their queries. But this experiment did not address a more fundamental question, which is whether there was benefit to MeSH terms in documents. Our runs indicated that the MeSH terms did confer some benefit, most likely by providing words that were in user queries but not in document titles or abstracts.

5.4 Relevance feedback

We also assessed the performance of relevance feedback by using the Ide method [4]. For each feedback run, the query was expanded by adding 30 terms that occurred in the top five or 10 relevant documents. In order to allow comparison with the other results generated, we did not use residual sets or frozen ranks. The results (see Table 10), indicate better results at the level of fewer documents retrieved (i.e., 15 but not 100).

5.5 Comparisons of SMART and human searchers

The final experiments were an attempt to compare SMART directly with human searchers. This was done by comparing the recall and precision of SMART and the human searcher for each query by having SMART retrieve the exact number of documents retrieved by the human, somewhat analogous to the original SMART-MEDLARS study [7]. For example, if a human searcher retrieved 100 documents for a given query, then we calculated recall and precision for SMART at 100 documents, whereas if a human searcher retrieved only 15 documents, we calculated results for SMART at 15 documents.

The paired comparisons for each of the original searcher groups is shown in Table 11. The first column contains the original searching results. The second column contains the recall and precision values for SMART at the same number of retrieved documents using the IN statement. The third column contains values at the same number of documents using the KF statement, while the last column uses the IN statement with judged documents only. Statistical analysis was done for each searcher group, utilizing a repeated measures analysis of variance with post hoc comparisons done using a t-test with Bonferroni correction.

In general, the R and P values for the original searchers were intermediate between SMART with all documents and with judged documents only. There were, however, some exceptions. For the two original searcher groups that used the MEDLINE feature set, the original values were higher than any of the SMART groups, including those using judged documents only, although the differences from the latter did not achieve statistical significance. This could be interpreted as a tendency for the full MEDLINE feature set to confer some benefit over SMART. However, this group also represents the most experienced searchers, and the results could also be reflecting their advanced skill.

Table 6 -- SMART weighting nomenclature.

triple <term_freq x idf x normalization>

term_freq:
- b - binary (always 1)
- a - term_freq normalized between 0.5 and 1.0 (i.e., 0.5 + 0.5 * tf/max_tf_in_doc)
- l - 1 + ln(term_freq)
- n - term_freq (i.e., number of times term occurs in doc)

idf:
- t - ln (N/n) where N = no. docs in collection and n = no. docs in which term occurs
- n - no idf factor

normalization:
- c - cosine normalization
- n - no normalization

Table 7 -- SMART comparison of weighting approaches based on information need statements and DR documents. Mean percent recall (R) and precision (P) are shown at 5, 15, and 100 documents. Average P represents the 11-point average of precision at each recall point on the recall-precision table.

	5 documents		15 documents		100 documents		Average
	R	P	R	P	R	P	P
anc.atn	8.9	14.3	17.3	11.6	41.8	5.9	12.7
ann.atn	15.2	29.1	29.8	23.4	59.8	9.5	25.3
atn.atn	14.6	27.3	27.2	21.5	58.5	9.2	23.9
bnn.btn	13.9	21.8	23.8	17.0	52.2	8.0	19.7
nnn.ntn	5.0	12.3	10.6	9.7	31.8	5.1	9.6

Table 8 - SMART comparison of weighting approaches based on information needs statements and DR documents, with unjudged documents removed.

	5 documents		15 documents		100 documents	
	R	P	R	P	R	P
anc.atn	16.7	28.4	32.0	23.2	69.8	10.9
ann.atn	19.4	35.8	36.2	28.3	79.2	13.1
atn.atn	17.9	32.5	35.3	27.2	76.6	12.9
bnn.btn	16.1	28.7	31.9	24.1	68.6	11.4
nnn.ntn	11.5	24.7	25.2	21.4	53.5	9.1

Table 9 - Alternative formulations in queries and documents.

	5 documents		15 documents		100 documents		Average
	R	P	R	P	R	P	P
IN query statements	15.2	29.1	29.8	23.4	59.8	9.5	25.3
0.25 * BIO + IN	15.3	29.7	31.4	23.8	62.2	9.8	26.1
KF query statements	15.0	27.9	31.1	23.4	64.1	9.9	26.3
IN on docs without MeSH	15.0	29.1	27.0	22.1	54.5	8.5	23.1

Table 10 -- Relevance feedback.

	5 documents		15 documents		100 documents		Average
	R	P	R	P	R	P	P
After 5 docs	19.3	39.6	33.4	26.2	60.9	10.0	33.0
After 10 docs	25.6	49.1	35.3	28.8	63.7	10.5	39.9

5.6 Assessment of human searcher features

The next set of experiments attempted to simulate the features that could assist the human searcher, in particular the skilled medical librarian. Table 12 starts with the best run from Table 11, librarians using full MEDLINE, and examines differences in the environment caused by the searcher for that run and the SMART IN run. The original searcher and SMART runs start out with the same information need statement. The factors differing in the two environments are:
1. Retrieval set size determined by searcher.
2. Searcher added new query terms.
3. Searcher performed intermediate runs and refined original query.
4. Searcher's retrieved set was fully evaluated for relevance.

Runs 2-4 of Table 12 illustrate the difficulty of effectively comparing retrieval runs when the size of the retrieved set varies on a query-by-query basis, even if the average number of retrieved documents is kept constant. These runs are exactly the same SMART run with the same ranking of documents and each retrieving the same number of total documents. The only difference is in the number of retrieved documents for individual queries. Run 2 retrieves the same number for each query as Run 1, while run 3 always retrieves 18 documents, the average of librarians using full MEDLINE in the original experiments. Run 4 is a manually optimized run, where the number of documents for each query is chosen to maximize the overall recall-precision. (In a real ranked output environment, a user would be able to stop retrieval at any desired point, but undoubtedly would not be able to achieve the recall and precision of run 4.)

For recall and precision, run 1 is significantly better than run 2, while run 4 has a tendency (though not statistically significant) towards improvement over run 1. The power of a user (or system evaluator) to set the retrieval set size has an enormous impact on evaluation. If at all possible, a comparative evaluation of two methods needs to keep the retrieval set size constant at a value neutral to the two methods.

Run 5 is an attempt to emulate the multiple iterations and increased vocabulary of the librarian searchers. A standard automatic feedback run was done, looking at the top 5 documents. Up to 30 terms from the relevant documents were added to the query. Unlike most feedback evaluations, the seen documents were retrievable on the feedback iteration (i.e., no frozen ranking or residual collection). This was done to more closely approximate the multiple iterations of the Boolean searchers. Run 5 is much closer to run 1, but still less effective. Run 6, evaluating using only the documents that had been judged, is again much better than the original run, but still less effective than run 1.

Table 12 demonstrates the difficulty of comparing Boolean searches done by trained intermediaries with ranked output searches. Each of the latter three runs is an attempt to equalize one factor in the differences between the two environments. Together they would achieve much better performance than the MEDLINE searchers. However, each of these runs is an upper bound on the effect of the factor. All that can really be said is that the "true" performance of the ranked system is somewhere between this upper bound and run 2.

Table 11 -- Comparison of original searchers with SMART.

Searcher Group	Original		SMART IN		SMART KF		SMART-IN-judged	
	R	P	R	P	R	P	R	P
Original KF	68.2	14.7	57.5	11.8	61.7	13.7	75.6	19.2
Librarians - MEDLINE	37.1	36.1	24.4	22.7	26.3	24.8	32.2	29.1
Librarians - Text-words	31.5	31.9	27.9	25.3	25.4	26.1	34.1	31.6
Physicians - MEDLINE	26.5	34.9	19.9	23.5	18.9	25.4	24.1	30.6
Physicians - Text-words	30.5	31.4	24.0	25.7	26.0	25.5	30.0	31.9

Table 12 -- Comparisons of Librarians - MEDLINE search group with SMART features.

Run	Retrieval Set Size	R	P
1. Librarians - MEDLINE	searcher set (avg. 18)	37.1	36.1
2. SMART IN	searcher set	24.4	22.7
3. SMART IN	always 18	32.7	21.9
4. SMART IN	optimized (avg. 18)	41.1	38.4
5. SMART IN feedback 5	always 18	35.6	24.3
6. SMART IN (remove non-judged)	searcher set	32.2	29.1

6 Conclusions

This first part of this study showed that novice physicians could effectively use a vector space-like retrieval system for searching MEDLINE in a clinical setting. This provides evidence against the notions that this approach is not effective in operational settings and cannot scale to large document collections. Further interactive experiments in different domains are necessary to see whether this success can be generalized.

This study also resulted in a new large test collection for medical information retrieval research. From both the interactive and SMART experiments, baseline performance data is available, which will allow experimenters and system builders to evaluate new approaches with a real world-sized collection from the medical domain. Experiments with SMART also showed the difficulty of comparing Boolean and ranked searches, especially in the interactive searching environment.

Acknowledgments

This work was supported in part by Grant LM05307 of the U.S. National Library of Medicine.

References

1. Knowledge Finder Reference Manual.North Andover, MA: Aries Systems Corp., 1988
2. Hersh W, Hickam D. Impact of a computerized information system in a university general medicine clinic. Clinical Research 1992;40:567A.
3. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Info Proc Mgmt 1988; 24(5):513-523.
4. Salton G, Buckley C. Improving retrieval performance by relevance feedback. J Am Soc Info Sci 1990; 41:288-297.
5. Belkin N, Cool C, Croft W, Callan J. Effect of multiple query representations on information retrieval system performance. In: Korfhage R, Rasmussen E, Willett P, ed. Proceedings of the 16th Annual International ACM Special Interest Group in Information Retrieval. Pittsbrugh, PA: ACM Press, 1993: 339-346.
6. Fuhr N, Knorz G. Retrieval test evaluation of a rule-based automatic indexing (AIR/PHYS). In: vanRijsbergen C, ed. Research and Development in Information Retrieval. Cambridge: Cambridge University Press, 1984: 391-408.
7. Salton G. A new comparison between conventional indexing (MEDLARS) and automatic text processing (SMART). J Am Soc Info Sci 1972; 23(2):75-84.
8. Salton G, Fox E, Wu H. Extended boolean information retrieval. Comm Assoc Comp Mach 1983; 26:1022-1036.
9. Turtle H, Croft W. Evaluation of an inference network-based retrieval model. ACM Transactions on Information Systems 1991; 9:187-222.
10. Robertson S, Thompson C. Weighted searching: The CIRT experiment. Informatics 10: Prospects for Intelligent Retrieval. York: ASLIB, 1990: 153-166.
11. Harman D. Overview of the first Text REtrieval Conference. In: Korfhage R, ed. Proceedings of the 16th Annual International ACM Special Interest Group in Information Retrieval. Pittsburgh: ACM Press, 1993: 36-47.
12. Hersh W, Hickam D. A comparison of two methods for indexing and retrieval from a full-text medical database. Proceedings of the 55th Annual Meeting of the American Society for Information Science. 1992: 221-230.
13. Hersh W, Hickam D. A performance and failure analysis of SAPHIRE with a MEDLINE test collection. Journal of the American Medical Informatics Association 1994; 1:51-60.
14. Haynes R, McKibbon K, Walker C, Ryan N, Fitzgerald D, Ramsden M. Online access to MEDLINE in clinical settings. Ann Int Med 1990; 112(1):78-84.
15. Saracevic T, Kantor P. A study of information seeking and retrieving. III. Searchers, searches, and overlap. J Am Soc Info Sci 1988; 39:197-216.