PERGAMON

Information Processing and Management 37 (2001) 383–402

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Challenging conventional assumptions of automated information retrieval with real users: Boolean searching and batch retrieval evaluations

William Hersh *, Andrew Turpin, Susan Price, Dale Kraemer, Daniel Olson, Benjamin Chan, Lynetta Sacherek

Division of Medical Informatics and Outcomes Research, Oregon Health Sciences University, 3181 SW Sam Jackson Park Road, Portland, OR 97201, USA

Accepted 19 September 2000

Abstract

Two common assumptions held by information retrieval researchers are that searching using Boolean operators is inferior to natural language searching and that results from batch-style retrieval evaluations are generalizable to the real-world searching. We challenged these assumptions in the Text Retrieval Conference (TREC) interactive track, with real users following a consensus protocol to search for an instance recall task. Our results showed that Boolean and natural language searching achieved comparable results and that the results from batch evaluations were not comparable to those obtained in experiments with real users.

Keywords: Information retrieval (IR); Evaluation; User studies; Text Retrieval Conference (TREC)

1. Introduction

A common assumption held by many researchers in the information retrieval (IR) field is that "natural language" searching (i.e., the entry of search terms without Boolean operators and the relevance ranking of results) is superior to searching using Boolean operators (Salton, 1991). Some research to support this notion comes from "batch" retrieval evaluations, in which a test collection of fixed queries, documents, and relevance judgments is used in the absence of real searchers to determine the efficacy of one retrieval system versus another. It has also been

* Corresponding author. Tel.: +1-503-494-4563; fax: +1-503-494-4551.
E-mail address: hersh@ohsu.edu (W. Hersh).

0306-4573/01/$ - see front matter
PII: S0306-4573(00)00054-6

advocated that this approach to evaluation can be generalized to real world searches (Salton & Buckley, 1988).

Previous research comparing Boolean and natural language systems has yielded conflicting results. The first study to compare Boolean and natural language searching with real searchers was the CIRT study, which found roughly comparable performance between the two when utilized by search intermediaries (Robertson & Thompson, 1990). Turtle found, however, that expert searchers using a large legal database obtained better results with natural language searching (Turtle, 1994). We have performed several studies of medical end-user searching comparing Boolean and natural language approaches. Whether using recall-precision metrics in bibliographic (Hersh, Buckley, Leone, & Hickam, 1994) or full-text databases (Hersh & Hickam, 1995), or using task-completion studies in bibliographic (Hersh, Pentecost, & Hickam, 1996) or full-text databases (Hersh et al., 1995), the results have been comparable for both types of systems.

Likewise, there is also debate as to whether the results obtained by batch evaluations, consisting of measuring recall and precision in the non-interactive laboratory setting, can be generalized to real searchers. Much evaluation research dating back to the Cranfield studies (Cleverdon & Keen, 1966) and continuing through the Text Retrieval Conference (TREC) (Harman, 1993) has been based on entering fixed query statements from a test collection into an IR system in batch mode with measurement of recall and precision of the output. It is assumed that this is an effective and realistic approach to determining the system's performance (Sparck Jones, 1981). Some have argued against this view, maintaining that the real world of searching is more complex than can be captured with such studies. These authors point out that relevance is not a fixed notion (Meadow, 1985), interaction is the key element of successful retrieval system use (Swanson, 1977), and relevance-based measures do not capture the complete picture of user performance (Hersh, 1994). If batch searching results cannot be generalized, then system design decisions based on them are potentially misleading.

We used the TREC interactive track to test the validity of these assumptions. The TREC-7 and TREC-8 interactive tracks use the task of instance recall to measure success of searching. Instance recall is defined as the number of instances of a topic retrieved (Hersh & Over, 2000). For example, a searcher might be asked to identify all the discoveries made by the Hubble telescope; in this case each discovery is an instance and the proportion of instances correctly listed is instance recall. This is in contrast to document recall, which is measured by the proportion of relevant documents retrieved. Instance recall is a more pertinent measure of user success at an IR task, since users are less likely to want to retrieve multiple documents covering the same instances. This paper reviews the results of our experiments in the TREC-7 and TREC-8 interactive tracks, where we assessed: (a) Boolean versus natural language searching (Hersh et al., 1998) and (b) batch versus actual searching evaluation results, respectively (Hersh et al., 1999).

2. Commonalities across studies

There were a number of common methods in both experiments, which we present in this section. Both studies used instance recall as the outcome (or dependent) variable. The study

consisted of a searcher who belonged to a group (librarian type in the TREC-7 experiment and librarian vs. graduate student in the TREC-8 experiment) and had a measurement of instance recall for each question (a total of eight in the TREC-7 experiment and six in the TREC-8 experiment).

All other data collected were predictor (or independent) variables. These variables can be grouped into five categories, each of which is listed with its individual data items in Table 1:

1. Demographic.
2. Experience.
3. Cognitive traits.
4. User satisfaction.
5. Search mechanics.

Demographic and experience attributes were collected via a common questionnaire used by all TREC-7 and TREC-8 interactive sites. This instrument queried age, gender, and experience in a variety of computer tasks. It also asked how frequently the subject searched as well as how much he or she enjoyed it.

Cognitive abilities were assessed using standardized instruments. Many studies have examined the association of these abilities with computer skills. The results have been decidedly mixed, precluding generalization. However, some have been shown in some studies to be associated with successful use of computer systems in general or retrieval systems specifically. These include:

1. *Spatial visualization*. The ability to visualize spatial relationships among objects has been associated with retrieval system performance by nurses (Staggers & Mills, 1994), ability to locate text in a general retrieval system (Gomez, Egan, & Bowers, 1986), and ability to use a direct-manipulation (3-D) retrieval system user interface (Swan & Allan, 1998).
2. *Logical reasoning*. The ability to reason from premise to conclusion has been shown to improve selectivity in assessing relevant and non-relevant citations in a retrieval system (Allen, 1992).
3. *Verbal reasoning*. The ability to understand vocabulary has been shown to be associated with the use of a larger number of search expressions and high-frequency search terms in a retrieval systems (Allen, 1992).
4. *Associational fluency*. The ability to associate words in meaning or context has been shown to be associated with effectiveness in using retrieval systems (Dumais & Schmitt, 1991).

User satisfaction attributes were measured from common instruments used by all TREC-7 and TREC-8 interactive sites. Instruments were developed for post-topic, post-system, and post-experiment administration. We added to the post-system assessment the Questionnaire for User Interface Satisfaction (QUIS) 5.0 instrument, which measures user satisfaction with a computer system (Chin, Diehl, & Norman, 1988). QUIS provides a score from 0 (poor) to 9 (excellent) on a variety of user factors, with the overall score determined by averaging responses to each item.

Searching mechanics attributes were determined by analysis of searching logs collected by the system. These attributes were defined as follows:

1. Number of search cycles – number of times a query was submitted to the system.
2. Number of total search terms used – total unique number of search terms for a topic.
3. Number of documents viewed – total number of unique documents viewed in all of the post-query summary lists.
4. Number of documents seen – total number of documents selected by user for reading.
5. Number of search terms per cycle – average number of search terms per search cycle.

Table 1
Common data collected during TREC-7 and TREC-8 interactive searching experiments

Variable	Definition
Demographic	
Gender	Male vs. female
Age	In years
Experience	
Years	Years experience of on-line searching (1-least, 5-most)
Point	Experience with point and click interface (1-least, 5-most)
Catalogs	Experience using on-line library catalogs (1-least, 5-most)
CDROM	Experience using CD-roms (1-least, 5-most)
Online	Experience searching commercial on-line systems (1-least, 5-most)
WWW	Experience searching Web (1-least, 5-most)
Frequency	How often searching done (1-least, 5-most)
Enjoy	How enjoyable searching is (1-least, 5-most)
Cognitive traits	
VZ2	Paper folding test to assess spatial visualization
RL1	Nonsense syllogisms test to assess logical reasoning
V4	Advanced vocabulary test I to assess verbal reasoning
FA1	Controlled associations test to assess associational fluency (TREC-7 only)
Satisfaction post-topic	
Familiar	User familiar with topic (1-least, 5-most)
EasyStart	Search was easy to get started (1-least, 5-most)
EasyUse	Search was easy to do (1-least, 5-most)
Satisfied	User was satisfied with results (1-least, 5-most)
Confident	User had confidence that all instances were identified (1-least, 5-most)
TimeAdequate	Search time was adequate (1-least, 5-most)
Satisfaction post-system	
SysEasyLearn	System was easy to learn to use (1-least, 5-most)
SysEasyUse	System was easy to use (1-least, 5-most)
SysUnderstand	User understand how to use system (1-least, 5-most)
QUIS	Average of all QUIS items (TREC-7 only)
Satisfaction post-experiment	
Understand	User understand nature of experimental task (1-least, 5-most)
TaskSim	Task had similarity to other searching tasks (1-least, 5-most)
TaskDiff	Systems were different from each other (1-least, 5-most)
QUIS	Average of all QUIS items (TREC-8 only)
Search mechanics	
Saved	Documents saved by user
DocRec	Document recall (relevance defined as having one or more instance)
Time	Time in seconds for search
Terms	Number of unique terms used for topic
Viewed	Number of documents viewed for topic
Seen	Number of documents seen for topic
Cycles	Number of search cycles for topic

3. Comparing Boolean versus natural language searching

The main goal of our TREC-7 interactive experiment was to compare searching performance with Boolean and natural language interfaces in a specific population of searchers, namely experienced information professionals. A secondary goal of the experiment was to identify attributes associated with successful searching in this population.

3.1. Methods

Users were randomly assigned to use one of two retrieval system interfaces – Boolean or natural language – for one block of four topics and then use the other interface for a second block of four topics per the general interactive track protocol. Both systems were accessed via Web-based interfaces (Fig. 1 shows the Boolean interface and Fig. 2 shows the natural language interface). There was a common retrieval system behind both interfaces, MG, a publicly available system with Boolean and natural language features (Witten, Moffat, & Bell, 1994). MG was run on a Sun Ultrasparc 140 with 256 megabytes of RAM. Each interface accessed MG via CGI scripts which contained JavaScript code for logging search strategies, documents viewed (titles displayed to user after search), and documents seen (all of document displayed after selection for viewing by user). Searchers accessed each system with either a Windows 95 personal computer or an Apple PowerMac, with each running Netscape Navigator 3.0.

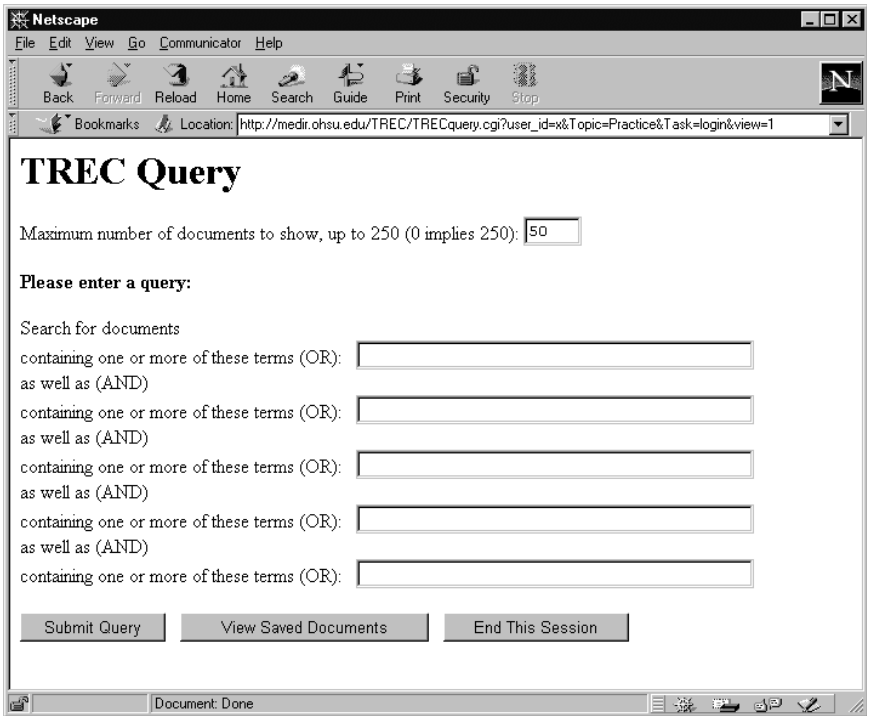

Fig. 1. Boolean interface for the TREC-7 interactive experiment.

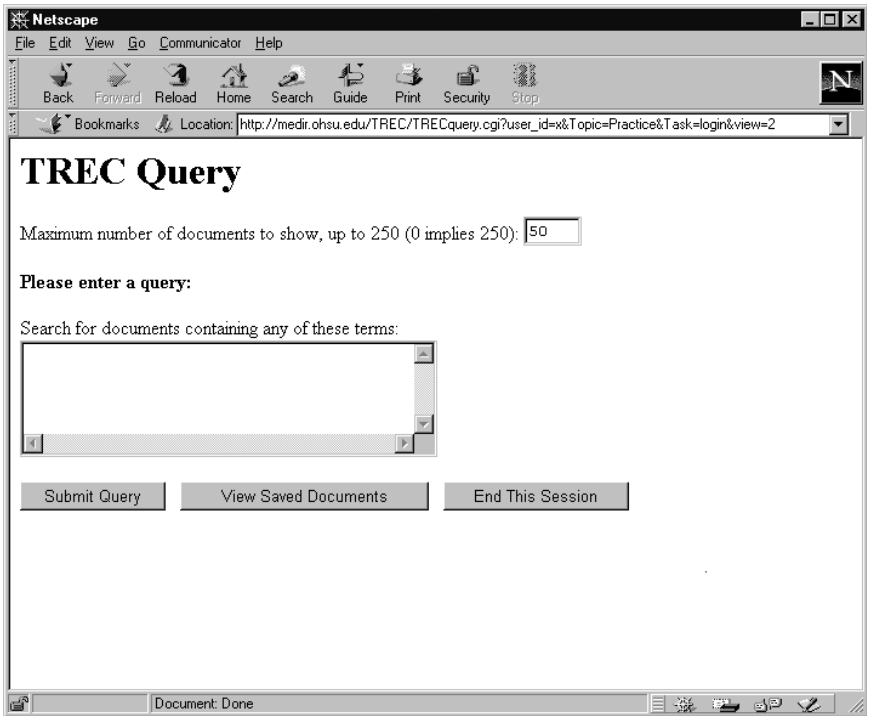

Fig. 2. Natural language searching interface for the TREC-7 interactive experiment.

Experimental subjects were recruited by advertising over three regional email listservs for information professionals and librarians (e.g., the American Society for Information Society Pacific Northwest Chapter). The advertisement explicitly stated that participation would be limited to experienced information professionals and that participants would be paid a modest remuneration for their participation. As subjects confirmed their participation, they were classified by type of information setting in which they worked: special (e.g., corporate, professional, or scientific), academic, and public. The experiments took place in a computer lab at Oregon Health Sciences University. The entire experimental session took 4 h, with the first half used for personal data and attributes collection and the second half used for searching.

After an introduction to the experiment, subjects were given the pre-search questionnaire to collect demographic and experience data. This was followed by administration of four cognitive tests described in Table 1 (VZ2, RL1, V4, and FA1). After this was an orientation to the searching part of the study. Subjects were then introduced to both retrieval systems, and performed a practice search on each. This pre-searching process took nearly 2 h.

The searching portion of the experiment began after a 10–15 min break. Each search was followed with measurement of per-search user satisfaction data. After the four searches in a given block, per-system user satisfaction data was measured using both the common instrument and QUIS. At the end of the experiment, overall user satisfaction was assessed.

Per the interactive track protocol, each subject was allowed 15 min per query. Subjects were instructed to identify as many instances as they could for each query. They were also instructed

for each query to write down each instance and save any document associated with an instance (either by using the "save" function of the system or writing its document identifier down on their worksheet).

Statistical analysis was done by constructing a general linear model to evaluate the effect of search system that incorporated all of the study design effects and their interactions, regardless of the statistical significance of individual effects. Attributes that were significantly associated with instance recall ($P = 0.1$) were included as candidates for a multiple attribute model. The list of attributes was further shortened by fitting the model including all of these candidate attributes, in addition to the design effects. Significant attributes ($P = 0.1$) from this multiple attribute model were included in a final model.

3.2. Results

A total of 24 subjects participated in the study – eight each of special, professional, and academic librarians. All subjects were information professionals, and all but two had a library degree. All completed the protocol as described above.

The gender breakdown of the 24 subjects analyzed was 16 women and 8 men. The average age of all subjects was 41.1 years. All subjects were highly experienced searchers. The average duration that they had been doing on-line searching was 7.8 years. In addition, they all stated that they searched once or twice daily in their jobs. They also either agreed (41.7%) or strongly agreed (58.3%) with the statement, "I enjoy carrying out information searches".

Table 2 summarizes the least square means comparing the systems for search outcomes, search mechanics, and system-specific user satisfaction variables. There were significant differences for several system-related factors. In particular, many fewer documents were viewed (had their titles shown on the screen after a search) with the Boolean interface, although that interface had more documents seen (had their full text shown after clicking on the title) and saved as instances. The Boolean interface was also deemed easier to learn and use, and had a higher QUIS user satisfaction score.

Table 2
Least square means comparing factors related to Boolean and natural language searching

Factor	Least squares mean		*P*-value
	Boolean	Natural language	
Instance recall	0.346	0.342	0.8854
Instance precision	0.688	0.698	0.7822
Total search terms	6.59	6.15	0.2815
Documents viewed	141.1	241.4	0.0004
Documents seen	14.68	13.58	0.0641
Documents saved	5.32	4.65	0.0543
Post-system easy to learn (1–5)	3.45	2.92	0.0850
Post-system easy to use (1–5)	2.95	2.13	0.0073
Post-system easy understand (1–5)	3.42	3.04	0.1310
Post-system QUIS average	4.61	4.09	0.0007

Table 3 shows responses from the exit survey. Users were asked their system preference in terms of ease of learning, ease of use, and overall preference. The Boolean system was clearly preferred on all three measures.

The full ANOVA model for instance recall is shown in Table 4. The model results show that there was no significant difference in instance recall between the Boolean and natural language search systems, with least squares means of 0.346 and 0.342, respectively. Librarian type was a marginally significant effect in this analysis. Pairwise comparisons revealed that special librarians were not significantly different from academic librarians (0.387 vs. 0.344, respectively), and academic librarians were not significantly different than public librarians (0.344 versus 0.302, respectively). However, there were differences between special and public librarians. Topic (nested within block and sequence) was also a significant effect, indicating variation in instance recall across different topics.

The final model included, in addition to the design effects, the following user attributes: librarian type, post-search satisfied with results, point-and-click experience, number of search cycles, number of documents seen, and number of search terms per cycle. All attributes but librarian type were modeled as a linear trend. Positive values indicate a positive association with instance recall, while negative values indicate a negative association. The results in Table 4 also show positive and negative associations of search success with various user attributes. There was a

Table 3
Exit survey responses comparing Boolean and natural language systems

Factor	Boolean	Natural language	*P*-value[a]
Users said easier to learn	17	6	0.0347
Users said easier to use	19	4	0.0026
Users said liked better	19	4	0.0026

[a] One-sample test for binomial proportion = 1/2.

Table 4
Analysis of variance model for user attributes comparing Boolean and natural language searching

	Factor	Model coefficients	*P*-value
Design effects	Sequence	–	0.9232
	Block	–	0.5554
	Sequence × Block	–	0.9205
	Topic (Sequence × Block)	–	0.0001
User attributes	Librarian type		0.0679
	Special	0[a]	
	Academic	−0.0193	0.5348
	Public	−0.0708	0.0249
	Point-and-click experience	−0.0638	0.0300
	Post-search satisfied with results	0.0353	0.0035
	Number of search cycles	−0.0071	0.0937
	Number of documents seen	0.0072	0.0083
	Number of search terms per cycle	0.0158	0.0310

[a] Reference group.

positive association for users who were satisfied with their results, viewed more documents on the screen, and used a larger number of search terms per cycle. There was a negative association with the number of search cycles used as well as experience with a point-and-click interface. The latter result, however, is dubious, since virtually all searchers (21 of 24) chose the highest experience option on the questionnaire and the one person who indicated the lowest point-and-click experience actually had the highest average instance recall of any searcher. If this individual had been excluded from the analysis, or his or her point-and-click experience score were changed from three to four, then there would be no significant difference in instance recall due to point and click experience.

3.3. Discussion

This experiment assessed the ability of highly experienced information professionals to identify instances of topics in an on-line database. The pre-search questionnaire showed they had a great deal of searching experience and computer experience in general. They performed on-line searching as one of their professional functions and carried it out on a daily basis.

The results showed that although these searchers strongly preferred a Boolean interface, there was little difference in success whether a Boolean or natural language interface was used. Analysis of other factors showed that success was associated with several user attributes. In particular, it was seen that searchers who worked in special libraries did better as a group than those from academic libraries, who in turn outperformed those from public libraries. There was also a positive association with successful searching and satisfaction with search results, number of documents shown on the screen, and the number of search terms used per cycle.

There was a negative association for the number of search cycles and experience with a point-and-click interface, though the latter was likely to be a statistical artifact. The advantage of fewer search cycles is probably due to the fact that successful searchers were likely to find good documents quickly. There were no associations between any of the cognitive or personality attributes that were assessed, contrary to a number of previous studies mentioned in the introduction.

This study gives further credence to the majority of studies that already show Boolean and natural language searching interfaces to achieve comparable results (e.g., Hersh et al., 1994, 1995, 1996; Hersh & Hickam, 1995; Robertson & Thompson, 1990). Expert searchers' preferences for Boolean interfaces may be a result of long-standing familiarity or a sense of more control over document output. Further research focusing on qualitative assessment may uncover situations where Boolean searching is indeed more effective.

4. Assessing the validity of batch-oriented retrieval evaluations

The goal of our TREC-8 experiment was to assess whether IR approaches achieving better performance in batch evaluations could translate that effectiveness to real users. This was done by a three-stage experiment. In the first stage we identified an "improved" weighting measure that achieved the best results over "baseline" TF $*$ IDF with previous (TREC-6 and TREC-7) interactive track queries and relevance judgments. Next, we used the TREC-8 instance recall task to compare searchers using the baseline and improved measures. In the final stage, we verified that

the performance of the improved measure over baseline held up with TREC-8 interactive track queries and relevance judgments. As the study also entailed data collection of other user attributes related to interactive searching, we also assessed the association of other factors with successful searching.

This section reports three iterative experiments:

1. Establishment of the best weighting approach for batch searching experiments using previous (TREC-6 and TREC-7) interactive track data.
2. User experiments to determine if those measures give comparable results with human searchers with new (TREC-8 interactive track) data. This analysis also looked at other factors predictive of successful searching from data collected by the user experiments.
3. Verification that the new (TREC-8 interactive track) data gives comparable batch searching results for the chosen weighting schemes.

Each experiment is described in a separate section, with appropriate methods introduced as they were used for each.

4.1. Finding an improved weighting scheme for experimental system

The goal of the first experiment was to find the most effective batch-mode weighting scheme for interactive track data that would subsequently be used in interactive experiments. All batch and user experiments in this study used the MG retrieval system (Witten et al., 1994). MG allows queries to be entered in either Boolean or natural language mode. If natural language mode is chosen, its relevance ranking scheme can be varied according to the Q-expression notation introduced by Zobel and Moffat (1998).

A Q-expression consists of eight letters written in three groups, each group separated by hyphens. For example, BB-ACB-BCA, is a valid Q-expression. The two triples describe how terms should contribute to the weight of a document and the weight of a query, respectively. The first two letters define how a single term contributes to the document/query weight. The final letter of each triple describes the document/query length normalization scheme. The second character of the Q-expression details how term frequency should be treated in both the document and query weight, e.g., as inverse document/query frequencies. Finally, the first character determines how the four quantities (document term weight, query term weight, document normalization, and query normalization) are combined to give a similarity measure between any given document and query. To determine the exact meaning of each character, the five tables appearing in the Zobel and Moffat paper must be consulted (Zobel & Moffat, 1998). Each character provides an index into the appropriate table for the character in that position.

Although the Q-expressions permit thousands of possible permutations to be expressed, the following generalizations can be made:

- Q-expressions starting with a B use the cosine measure for combining weights, while those starting with an A do not divide the similarity measure through by document or query normalization factors.
- A B in the second position indicates that the natural logarithm of one plus the number of documents divided by term frequency is used as a term's weight, while a D in this position indicates that the natural logarithm of one plus the maximum term frequency divided by term frequency is used.

- A C in the fourth position indicates a cosine measure based term frequency treatment, while an F in this position indicates Okapi-style usage (Robertson & Walker, 1994).
- Varying the fifth character alters the document length normalization scheme.
- Letters greater than H use pivoted normalization (Singhal, Buckley, & Mitra, 1996).

4.1.1. Methods

In order to determine the best batch-mode weighting scheme, we needed to convert the prior interactive data (from TREC-6 and TREC-7) into a test collection for batch-mode studies. This was done by using the description section of the interactive query as the query and designating all documents as relevant to the query where one or more instances were identified within it. The batch experiments set out to determine a baseline performance and one with maximum improvement that could be used in subsequent user experiments. Each Q-expression was used to retrieve documents from the 1991 to 1994 Financial Times collection (used in the Interactive Track for the past three years) for the 14 TREC-6 and TREC-7 Interactive Track topics. Average precision was calculated using the `trec_eval` program.

4.1.2. Results

Table 5 shows the results of our batch experiments using TREC-6 and TREC-7 Interactive Track data. The first column shows average precision. The next column gives the percent improvement over the baseline, which in this case was the BB-ACB-BAA (basic vector space TF $*$ IDF) approach. The baseline was improved upon by other approaches shown to be effective in other TREC tasks (e.g., ad hoc), in particular pivoted normalization (second and third rows – with slope of pivot listed in parentheses) and the Okapi weighing function (remaining rows). The best improvement was seen with the AB-BFD-BAA measure, a variant of the Okapi weighing function, with an 81% increase in average precision. This measure was designated for use in our user experiments because it had maximal improvement over the TF $*$ IDF baseline.

4.2. Interactive searching to assess best batch weighting scheme with real users

Based on the results from the first experiment, the explicit goal of the interactive experiment was to assess whether the AB-BFD-BAA (Okapi) weighting scheme provided benefits to real users

Table 5
Average precision and improvement for different Q-expressions (with corresponding weighting type) on batch runs using TREC-6 and TREC-7 interactive data

Q-Expression	Weighting type	Average precision	% Improvement
BB-ACB-BAA	TFIDF	0.2129	0%
BD-ACI-BCA (slope = 0.5)	Pivoted Norm.	0.2853	34%
BB-ACM-BCB (slope = 0.275)	Pivoted Norm.	0.2821	33%
AB-BFC-BAA	Okapi	0.3612	70%
AB-BFD-BAA	Okapi	0.3850	81%
AB-BFE-BAA	Okapi	0.3517	65%

in the TREC interactive setting over the TF * IDF baseline. We performed our experiments with the risk that this benefit might not hold for TREC-8 interactive data (though as seen in the following experiment 3 below, this was not the case).

4.2.1. Methods

This experiment was carried out according to the consensus protocol developed by track participants. We used all of the instructions, worksheets, and questionnaires developed by consensus, augmented with some additional instruments, such as tests of cognitive abilities and a validated user interface questionnaire. Both the baseline and Okapi systems used the same Web-based, natural language interface shown in Fig. 3. MG was run on a Sun Ultrasparc 140 with 256 megabytes of RAM running the Solaris 2.5.1 operating system. The user interface accessed MG via CGI scripts which contained JavaScript code for designating the appropriate weighting scheme and logging search strategies, documents viewed (title displayed to user), and documents seen (all of document displayed by user). Searchers accessed each system with either a Windows 95 PC or an Apple PowerMac, running Netscape Navigator 4.0.

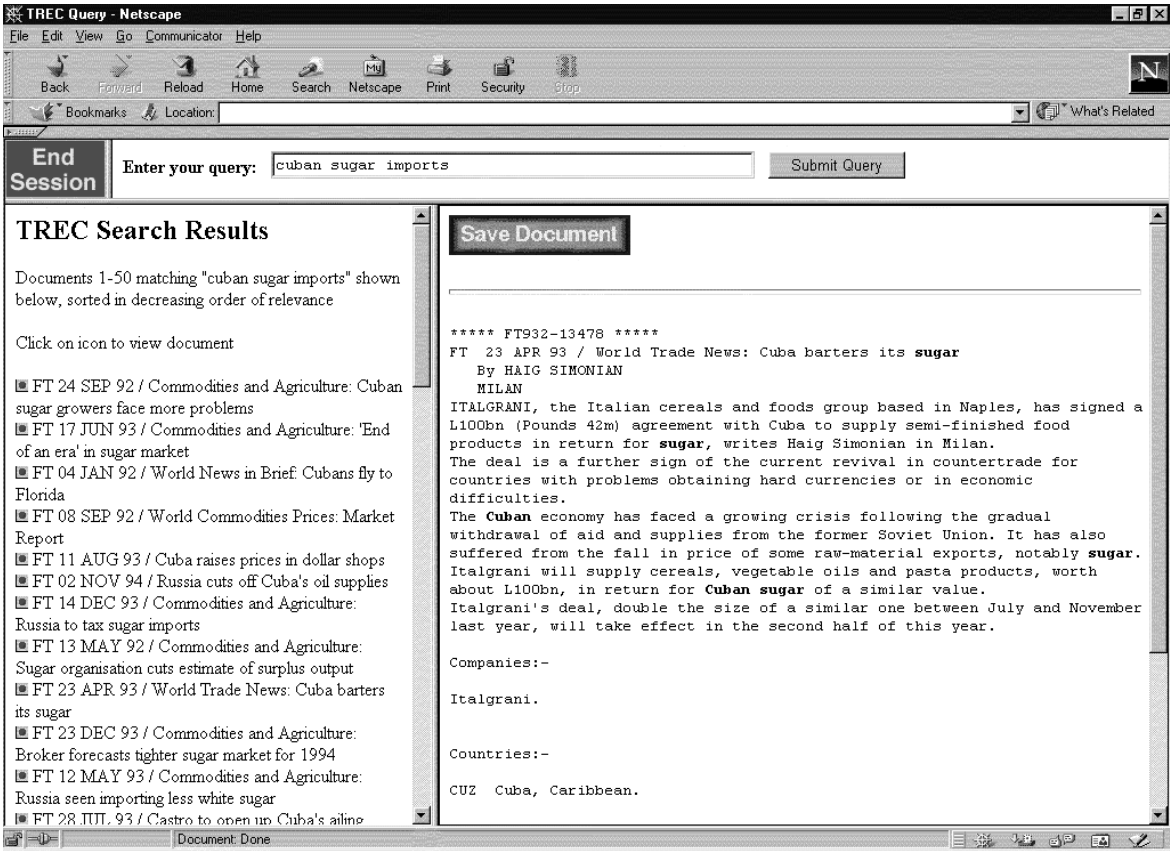

Fig. 3. Searching interface for the TREC-8 interactive experiment (both TFIDF and Okapi weighting).

Librarians were recruited by advertising over several librarian-oriented listservs in the Pacific Northwest. The advertisement explicitly stated that we sought information professionals with a library degree and that they would be paid a modest honorarium for their participation. Graduate students were recruited from the Master of Science in Medical Informatics Program at OHSU. They had a variety of backgrounds, ranging from physicians or other health care professionals to having completed non-health undergraduate studies.

The experiments took place in a computer lab. Each session took three and an half hours, broken into three parts, separated by short breaks: pre-searching data collection and orientation, searching with one system, and searching with the other system. The pre-searching data collection consisted of collection of demographic and experience data, followed by the administration of cognitive trait tests. Next was an orientation to the searching session and retrieval system, with the demonstration of a search and a practice search using a topic from a previous interactive track.

The personal data and attributes collection was followed by a 10 min break. The searching portion of the experiment consisted of searching on the first three topics assigned, taking a 15-min break, and searching on the second three topics assigned. Per the consensus protocol, each participant was allowed 20 min per query. Participants were instructed to identify as many instances as they could for each query. They were also instructed for each query to write each instance on their worksheet and save any document associated with an instance (either by using the "save" function of the system or writing its document identifier down on the searcher worksheet).

Each participant was assigned to search three queries in a block with one system followed by three queries with the other system. A pseudo-random approach was used to insure that all topic and system order effects were nullified. (A series of random orders of topics with subject by treatment blocks were generated for balance and used to assign topics.)

After each search, a brief questionnaire collecting the post-topic data was collected. After each search of three topics were searched using one system, the post-system data was collected. After the experiment was over, the post-experiment data was collected. We also administered QUIS in this experiment, but it was given only at the post-experiment stage as a measure of overall user interface satisfaction, since the user interfaces for the two systems were identical.

After the experiments were completed, data were organized into a per-question format with all associated attributes. To address the question of whether there was a significant difference between the Okapi and TF $*$ IDF systems, an analysis of variance (ANOVA) model was fit to instance recall for study design data. The factors in the model included type of searcher, the individual ID (nested in type), system, and topic. In the analysis, ID and topic were random factors, while type and system were fixed factors. Two-factor interactions (among system, topic, and type) were also included in the analysis.

To assess other factors associated with successful searching, all of the other predictor variables were treated as covariates in the base ANOVA model, including subject demographic characteristics, cognitive test results, post-searching questionnaire responses, and exit questionnaire responses. Each individual covariate was added one at a time to examine its contribution to the model. Each was treated as a scale variable, even if it was ordinal or categorical. We also focused explicitly on the intermediate outcomes of documents saved, document recall, number of documents viewed, and number of documents seen by developing a separate ANOVA model to assess their association with instance recall.

4.2.2. Results

A total of 24 searchers consisting of 12 librarians and 12 graduate students completed the experiment. The average age of the librarians was 43.9 years, with seven women and five men. The average age of the graduate students was 36.5 years, with eight women and four men. All searchers were highly experienced in using a point-and-click interface as well as on-line and Web searching.

Table 6 shows instance recall and precision comparing systems and user types. While there was essentially no difference between searcher types, the Okapi system showed an 18.2% improvement in instance recall and an 8.1% improvement in instance precision, both of which were not statistically significant. Table 7 shows the *P*-values for the ANOVA model. Of importance was that while the difference between the systems alone was not statistically significant, the interaction between system and topic was. In fact, as shown in Fig. 4, all of the difference between the systems occurred in just one query, 414i. The task of this query was to identify countries to whom Cuba exports sugar.

A number of variables were associated with instance recall in a statistically significant manner. Intermediate outcome measures that were associated in a statistically significant manner included:

1. The number of documents saved by the user as containing an instance ($P < 0.001$).
2. Document recall (with document relevance defined as one containing one or more instances) ($P < 0.001$).
3. The number of documents seen by the user ($P = 0.002$).

Fig. 5(a)–(c) show the linear fit of the intermediate outcome variables. The first result raises the possibility that an intermediate measure, number of documents saved by the user, could be used to

Table 6
Instance recall and precision across TFIDF and Okapi systems and user types

	Instance recall	Instance precision
System		
TFIDF	0.33	0.74
Okapi	0.39	0.80
Type		
Librarian	0.36	0.76
Graduate student	0.36	0.78

Table 7
Summary of analysis of variance model for TFIDF and Okapi systems

Source	*P*-value
System	0.226
Topic	0.0516
Type	0.914
ID (Type)	0.0516
System * Topic	0.0269
System * Type	0.0881
Topic * Type	0.108

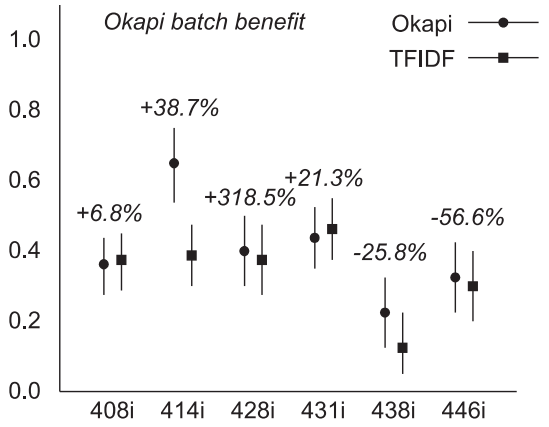

Fig. 4. Instance recall for each topic. The point values show the mean and confidence intervals for users with Okapi (circular point) and TFIDF (square point) weighting. In italics are the change in average precision for Okapi over TFIDF weighting.

measure searching outcome without the labor-intensive relevance judgments to measure instance recall. Our findings also indicate that the quantity of relevant (containing an instance) documents retrieved is associated with ability to perform the instance recall task. They also indicate that success at the instance recall task is related to the number of documents that the user pulls up the full text to read, adding credence to the (unpublished) observation that the ability to succeed at the instance recall task is related to reading speed.

While none of the demographic/experience, cognitive, post-searching, or post-experiment variables were associated with higher instance recall, three of the post-searching variables were. We found that the higher familiarity the user expressed with a topic, the lower instance recall they obtained ($P < 0.001$), as shown in Fig. 6. The meaning of the inverse relationship between familiarity with the topic and instance recall is unclear, though perhaps suggests that users knowledgeable about the topic were less likely to search comprehensively. Ease of performing the search ($P = 0.003$) and confidence that all instances were identified ($P = 0.01$) were, however, associated with successful searching.

4.3. Verifying best weighting scheme with TREC-8 data

The next experiment was to verify that the improvements in batch evaluation detected with TREC-6 and TREC-7 data held with TREC-8 data. It may have been possible that the benefit of Okapi weighting did not materialize with the latter, thus rendering the result in the second experiment not applicable to determining whether improvements in batch searching results hold up with real users.

4.3.1. Methods

The batch runs for the baseline and Okapi systems from the first experiment were repeated using the same approach of developing a test collection by designating all documents as relevant to the query where one or more instances were identified within it.

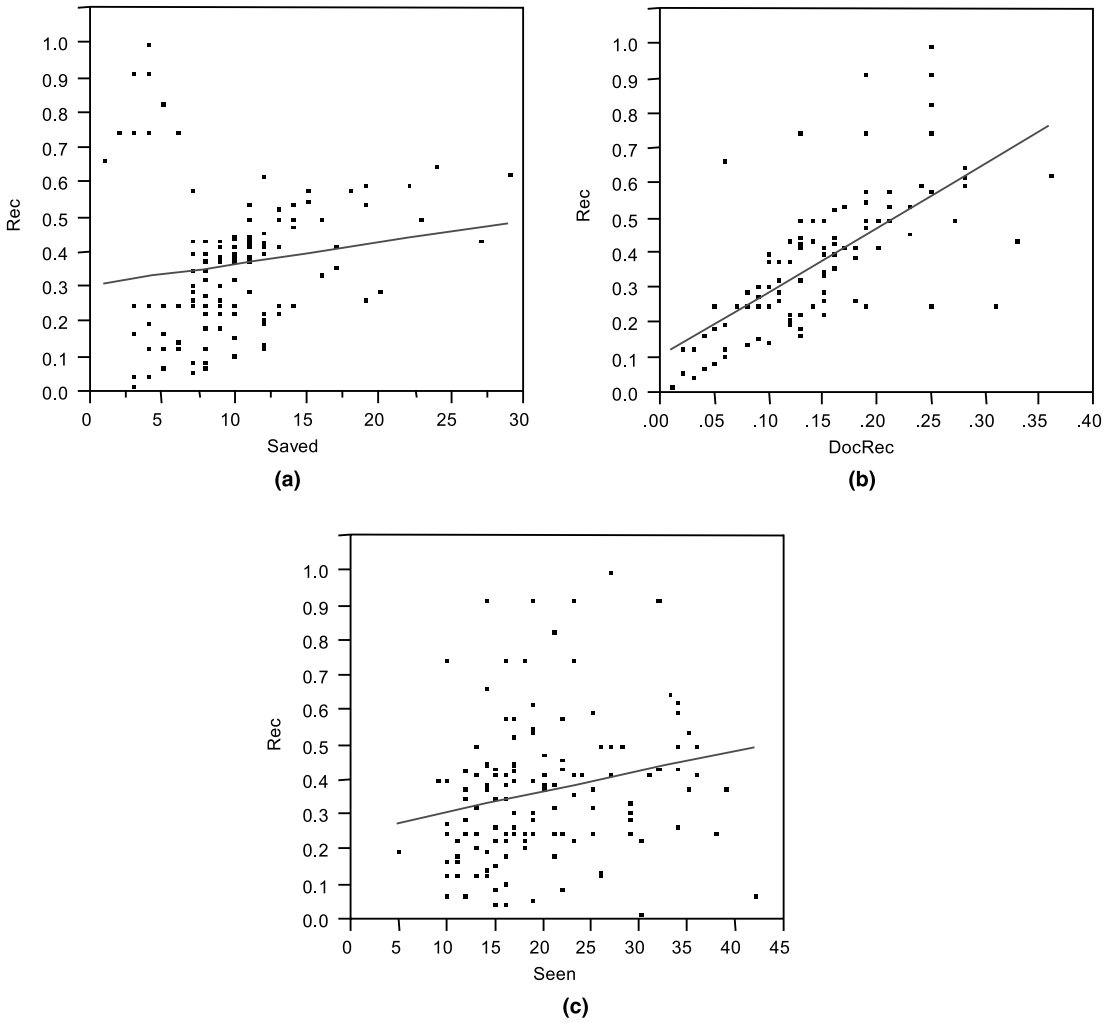

Fig. 5. Linear fit of relationship between instance recall and (a) number of documents saved, (b) document-level recall, and (c) number of documents seen.

4.3.2. Results

Table 8 lists the average precision for both systems used in the user studies along with percent improvement. The Okapi AB-BFD-BAA still outperformed the baseline system, BB-ACB-BAA, but by the lesser amount of 17.6%. This happened to be very similar to the difference in instance recall noted in the second experiment.

One possible reason for the smaller gains on the TREC-8 vs. TREC-6 and TREC-7 queries was that the average number of relevant documents for a TREC-8 query was three times higher than a query in the TREC-6 or TREC-7 sets. On average, TREC-6 interactive queries had 36 relevant documents, TREC-7 had queries 30 relevant documents, and TREC-8 queries had 92 relevant documents. The higher number of relevant documents may have given the baseline TF $*$ IDF

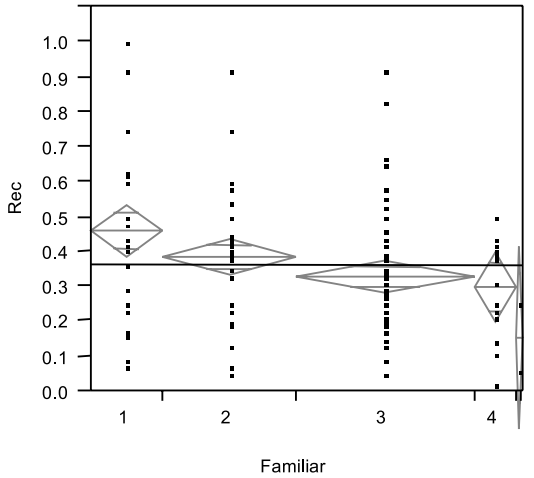

Fig. 6. Relationship between instance recall and familiarity with topic as designated by user. Each diamond represents the mean plus one and two standard deviations.

Table 8
Average precision and improvement for Okapi in batch runs for TREC-8 data

Query	Instances	Relevant documents	Baseline	Okapi	% Improvement
408i	24	71	0.5873	0.6272	6.8%
414i	12	16	0.2053	0.2848	38.7%
428i	26	40	0.0546	0.2285	318.5%
431i	40	161	0.4689	0.5688	21.3%
438i	56	206	0.2862	0.2124	−25.8%
446i	16	58	0.0495	0.0215	−56.6%
Average	29	92	0.2753	0.3239	17.6%

system a better chance of performing well, narrowing the gap between the different ranking schemes.

Also noteworthy in these results is that while query 414i achieved the second-best improvement of the six in average precision, it was far less than the improvement for 428i, which showed no improvement in the user studies. In fact, two queries showed a decrease in performance for Okapi with no difference in the user studies.

4.4. Discussion

Our experiments showed that batch and user searching experiments do not give the same results. This outcome calls into question whether results from batch studies should be interpreted as a definitive assessment of system performance. The ultimate answer to the question of whether these two approaches to evaluation give the same results must ultimately be answered by further experiments that use a larger number of queries and more diverse user tasks.

Another observation from this experiment is that simple statistical analyses may obscure more complex situations. In particular, just performing a simple *t*-test on the overall means in Table 4 could lead one to conclude that retrieval systems which perform better in batch studies also do so in user studies. However, our more statistically proper ANOVA model showed that the difference was not statistically significant and occurred solely due to one query, 414i. The reason for this query being an outlier is not clear, as the subject matter for this query was not markedly different from the others. The only difference was that it had far fewer relevant documents than the rest (see Table 8), making it more likely to amplify random differences in user search strategies.

Additional analysis of the data also presents a more complex picture of real-user searching. For example, user familiarity with a topic was shown to vary inversely with searching performance. While this may be an artifact of the laboratory-based study design, it could also indicate that users may be lulled into a false sense of security about topics for which they have underlying knowledge. Further study of this outcome is indicated to address user performance under varying baseline familiarity with a topic.

5. Conclusions

The results of these experiments challenge some widely held assumptions by many researchers in the IR field, which are that natural language systems are superior to Boolean systems and that the results of batch searching experiments are generalizable to evaluations with real users. Our particular experiments showed that for experienced users searching TREC-style queries on TREC databases, Boolean and natural language searching yield comparable instance recall. Furthermore, results obtained from batch studies do not parallel the results obtained in this task.

There are, of course, limitations to these studies that warrant further investigation. For example, these studies only looked at the TREC interactive instance recall task. There are many other IR tasks which must be assessed, from simple answering of questions to comprehensive searches on a topic. Another limitation is the type of searcher we used, namely the experienced librarian or graduate student. Further studies must involve other types of searchers, including those from other professions as well as non-professionals, such as students and the general public.

But these experiments show that all notions of the efficacy of IR systems can be studied with real users. It is commonly heard that user studies cannot be used for every permutation of a system one might wish to study (e.g., the myriad of weighting schemes in the MG or SMART systems, or the many different types of users and questions they may pose to a system). It is true that real user studies are time-consuming and expensive. However, the disparity in results between batch and real-user studies demonstrates that outcomes from the former may not be generalizable.

These studies provided a few noteworthy insights into real-world searching. One observation is that subjects in this study used an average of more than six unique search terms per query. This is about three times higher than the average number of terms used by general users of Web search engines (Jansen, Spink, Bateman, & Saracevic, 1998). This indicates that experienced searchers, or at least searchers performing the sort of task used in this study, are different than general Web searchers. It also indicates that using more terms improves search success.

The TREC-7 experiments showed that instance recall was associated with user satisfaction of results, number of documents viewed, and number of terms used. Instance recall decreased, however, with the number of search cycles. These results would imply that successful searching is associated with user satisfaction (something that does not always occur with computer applications in general, e.g., Nielsen & Levy, 1994), the ability to read more documents, and the ability to think of more search terms. Poor searches, i.e., those requiring a larger number of search cycles, diminish performance in the instance recall task.

The TREC-8 experiments showed that instance recall was associated with number of documents saved, number of documents seen, and document-level recall. These results imply, similar to TREC-7, that ability to read more documents is an important ability in the instance recall task. These experiments also showed an unexplained negative association between instance recall topic familiarity. The reason for this is unclear, and it may well just be a chance event. In any case, further investigation is needed.

The results of these studies reinforce the need for more user studies and caution against over-reliance on results obtained in batch studies. They also show that the TREC evaluation milieu can be used for such studies. The advantage of TREC is that it provides a standardized data set and experimental methodology for experimentation. In addition to participating in future TREC interactive tracks, we also plan additional experiments with existing data to verify the results of this study.

Acknowledgements

This study was funded in part by Grant LM-06311 from the US National Library of Medicine.

References

Allen, B. (1992). Cognitive differences in end-user searching of a CD-ROM index. In *Proceedings of the 15th annual international ACM special interest group in information retrieval* (pp. 298–309). Copenhagen: ACM Press.

Chin, J., Diehl, V., & Norman, K. (1988). Development of an instrument measuring user satisfaction of the human–computer interface. In *Proceedings of CHI'88 – human factors in computing systems* (pp. 213–218). New York: ACM Press.

Cleverdon, C., & Keen, E. (1966). *Aslib Cranfield research project: Factors determining the performance of indexing systems* (Vol. 1: Design, Vol. 2: Results). Cranfield, UK.

Dumais, S., & Schmitt, D. (1991). Iterative searching in an online database. In *Proceedings of the human factors society 35th annual meeting* (pp. 398–403).

Gomez, L., Egan, D., & Bowers, C. (1986). Learning to use a text editor: some learner characteristics that predict success. *Human–Computer Interaction*, *2*, 1–23.

Harman, D. (1993). Overview of the first Text Retrieval Conference. In *Proceedings of the 16th annual international ACM special interest group in information retrieval* (pp. 38–47). Pittsburgh: ACM Press.

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

Hersh, W., Buckley, C., Leone, T., & Hickam, D. (1994). 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* (pp. 192–201). Dublin: Springer.

Hersh, W., Elliot, D., Hickam, D., Wolf, S., Molnar, A., & Leichtenstein, C. (1995). Towards new measures of information retrieval evaluation. In *Proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 164–170). Seattle: ACM Press.

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

Hersh, W., & Over, P. (2000). TREC-8 interactive track report. In *Proceedings of the eighth text retrieval conference (TREC-8) 1999*, (pp. 57–64). Gaithersburg, MD: NIST.

Hersh, W., Pentecost, J., & Hickam, D. (1996). A task-oriented approach to information retrieval evaluation. *Journal of the American Society for Information Science*, *47*, 50–56.

Hersh, W., Price, S., Kraemer, D., Chan, B., Sacherek, L., & Olson, D. (1998). A large-scale comparison of Boolean vs. natural language searching for the TREC-7 interactive track. *Proceedings of the seventh text retrieval conference (TREC-7)* (pp. 491–500), Gaithersburg, MD: NIST.

Hersh, W., Turpin, A., Price, S., Kraemer, D., Chan, B., Sacherek, L., & Olson, D. (1999). Do batch and user evaluations give the same results? An analysis from the TREC-8 interactive track. In *Proceedings of the eighth text retrieval conference (TREC-8)*. Gaithersburg, MD: NIST.

Jansen, B., Spink, A., Bateman, J., & Saracevic, T. (1998). Real life information retrieval: a study of user queries on the Web. *SIGIR Forum*, *32*, 5–17.

Meadow, C. (1985). Relevance?. *Journal of the American Society for Information Science*, *36*, 354–355.

Nielsen, J., & Levy, J. (1994). Measuring usability: preference vs. performance. *Communications of the Association for Computing Machinery*, *37*, 66–75.

Robertson, S., & Thompson, C. (1990). Weighted searching: The CIRT experiment. *Informatics 10: Prospects for intelligent retrieval*. New York.

Robertson, S., & Walker, S. (1994). 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* (pp. 232–241). Dublin: Springer.

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

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, *24*, 513–523.

Singhal, A., Buckley, C., & Mitra, M. (1996). Pivoted document length normalization. In *Proceedings of the 19th annual international ACM special interest group in information retrieval* (pp. 21–29). Zurich: ACM Press.

Sparck-Jones, K. (1981). *Information retrieval experiment*. London: Butterworths.

Staggers, N., & Mills, M. (1994). Nurse–computer interaction: staff performance outcomes. *Nursing Research*, *43*, 144–150.

Swan, R., & Allan, J. (1998). Aspect windows, 3-D visualization, and indirect comparisons of information retrieval systems. In *Proceedings of the 21st annual international ACM special interest group in information retrieval* (pp. 173–181). Melbourne, Australia: ACM Press.

Swanson, D. (1977). Information retrieval as a trial-and-error process. *Library Quarterly*, *47*, 128–148.

Turtle, H. (1994). Natural language vs. Boolean query evaluation: a comparison of retrieval performance. In *Proceedings of the 17th annual international ACM special interest group in information retrieval* (pp. 212–220). Dublin: Springer.

Witten, I., Moffat, A., & Bell, T. (1994). *Managing gigabytes – compressing and indexing documents and images*. New York: Van Nostrand Reinhold.

Zobel, J., & Moffat, A. (1998). Exploring the similarity space. *SIGIR Forum*, *32*, 18–34.