Do Batch and User Evaluations Give the Same Results?

William Hersh, Andrew Turpin, Susan Price, Benjamin Chan,
Dale Kraemer, Lynetta Sacherek, Daniel Olson
{hersh; turpina; prices; chanb; kraemerd; sacherek; olsondan}@ohsu.edu
Division of Medical Informatics & Outcomes Research
Oregon Health Sciences University
Portland, OR, USA

Do improvements in system performance demonstrated by batch evaluations confer the same benefit for real users? We carried out experiments designed to investigate this question. After identifying a weighting scheme that gave maximum improvement over the baseline in a non-interactive evaluation, we used it with real users searching on an instance recall task. Our results showed the weighting scheme giving beneficial results in batch studies did not do so with real users. Further analysis did identify other factors predictive of instance recall, including number of documents saved by the user, document recall, and number of documents seen by the user.

1. Introduction

A continuing debate in the information retrieval (IR) field is 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 [2] and continuing through the Text Retrieval Conference (TREC) [3] 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 [9]. Some have argued against this view, maintaining that the real world of searching is more complex than can be captured with such studies. They point out that relevance is not a fixed notion [6], interaction is the key element of successful retrieval system use [10], and relevance-based measures do not capture the complete picture of user performance [4].

If batch searching results cannot be generalized, then system design decisions based on them are potentially misleading. The goal of this study therefore was to assess whether IR approaches achieving better performance in the batch environment could translate that effectiveness to real users. As the study also entailed data collection of other user attributes related to interactive searching, we were also able to assess the association of other factors with successful searching.

SIGIR 2000 7/00 Athens, Greece

The experimental milieu for assessing the study question was the TREC-8 interactive track. As with its predecessors from the two years previous (TREC-6 and TREC-7 interactive tracks), an "instance recall" task was employed, where users were asked to identify instances of a topic [5]. Instance recall was defined as the fraction of total instances (as determined by the NIST assessor) for the topic that were covered by the documents saved by the user. Figure 1 shows two example TREC-8 interactive track queries.

The overall plan for this study was to transform queries, documents, and relevance judgments from the TREC-6 and TREC-7 interactive tracks into a test collection that could identify highly effective batch performance compared to a baseline. In particular, we focused on the newer weighting schemes that have shown to be effective with TREC data over the standard TF*IDF baseline. This allowed the identification of a weighting approach that could be assessed in interactive user experiments.

This paper reports four experiments:
1. Establishment of the best weighting approach for batch searching experiments using previous TREC interactive track data.
2. User experiments to determine if those measures give comparable results with human searchers with new TREC interactive track data.
3. Verification that the new TREC interactive track data gives comparable batch searching results for the chosen weighting schemes.
4. Analysis of other factors predictive of successful searching from data collected by the user experiments.

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

2. Finding an effective weighting scheme for experimental system

The goal for 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 [11]. MG allows queries to be entered in either Boolean or ranked mode. If ranking is chosen, the ranking scheme can be varied according to the Q-expression notation introduced by Zobel and Moffat [12].

```
Number:
  414i
Title:
  Cuba, sugar, imports
Description:
  What countries import Cuban sugar?
Instances:
  In the time alloted, please find as many DIFFERENT countries of
  the sort described above as you can. Please save at least one
  document for EACH such DIFFERENT country.
  If one document discusses several such countries, then you need
  not save other documents that repeat those, since your goal
  is to identify as many DIFFERENT countries of the sort described
  above as possible.

Number:
  428i
Title:
  declining birth rates
Description:
  What countries other than the US and China have or have had
  a declining birth rate?
Instances:
  In the time alloted, please find as many DIFFERENT countries of
  the sort described above as you can. Please save at least one
  document for EACH such DIFFERENT country.
  If one document discusses several such countries, then you need
  not save other documents that repeat those, since your goal
  is to identify as many DIFFERENT countries of the sort described
  above as possible.
```

Figure 1 - Sample queries from the TREC interactive track.

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 [12]. 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, several 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 [7]. Varying the fifth character alters the document length normalization scheme. Letters greater than H use pivoted normalization [8].

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 1991-1994 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.

Results

Table 1 shows the results of our batch experiments using TREC-6 and TREC-7 Interactive Track data. The first column shows average precision, while 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.

3. Interactive searching to assess weighting scheme with real users

Based on the results from Experiment 1, the explicit goal of the interactive experiment was to assess whether the AB-BFD-BAA (Okapi) weighting scheme provided benefits to real users 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 Experiment 3 below, this was not the case.

Methods

The main performance measure used in the TREC-8 interactive track was instance recall, defined as the proportion of true instances identified by a user searching on the topic. Relevance assessors at NIST defined the instances from pooled searching results from all experimental groups, as described in the past [5]. The experiment was carried out according to the consensus protocol developed by track participants (described in detail at trec.nist.gov). 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. Table 2 lists all of the data collected for each search in the experiment.

Both the baseline and Okapi systems used the same Web-based, natural language interface shown in Figure 2. 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.

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, 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 one-half hours, broken into three parts, separated by short breaks: personal data and attributes collection, searching with one system, and searching with the other system. The personal data and attributes collection consisted of the following steps:
1. Orientation to experiment (10 minutes)
2. Collection of Demographic/Experience data listed in Table 2 (10 minutes)
3. Collection of Cognitive data listed in Table 2 (40 minutes)
4. Orientation to searching session and retrieval system, with demonstration of a search (10 minutes)
5. Practice search using a topic from a previous interactive track (10 minutes)

The cognitive data was obtained by using tests from the Educational Testing Service (ETS) shown in past IR research to be associated with some aspect of successful searching.

The personal data and attributes collection was followed by a 10 minute break. The searching portion of the experiment consisted of searching on the first three topics assigned, taking a 15-minute break, and searching on the second three topics assigned. Per the consensus protocol, each participant was allowed 20 minutes 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 listed in Table 2 was administered. After each search of three topics were searched using one system, the Post-System data from Table 2 was collected. After the experiment was over, the Post-Experiment data from Table 2 was collected. We also administered the Questionnaire for User Interface Satisfaction (QUIS) 5.0 instrument [1]. 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. QUIS was given only at the end as a measure of overall user interface satisfaction since the interfaces for the two systems were identical.

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%

Table 1 - Average precision and improvement for different Q-expressions (with corresponding weighting type) on batch runs using TREC-6 and TREC-7 interactive data.

Variable	Definition
Study Design	
Type	Librarian vs. medical informatics graduate student
Topic	Topic number
System	Search system used (Okapi vs. TF*IDF)
Intermediate Outcomes	
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
QUIS	Average of all QUIS scores
Demographic/Experience	
Gender	Male vs. female
Age	In years
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	
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
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)
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)
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)

Table 2 - Data collected during interactive searching experiments.

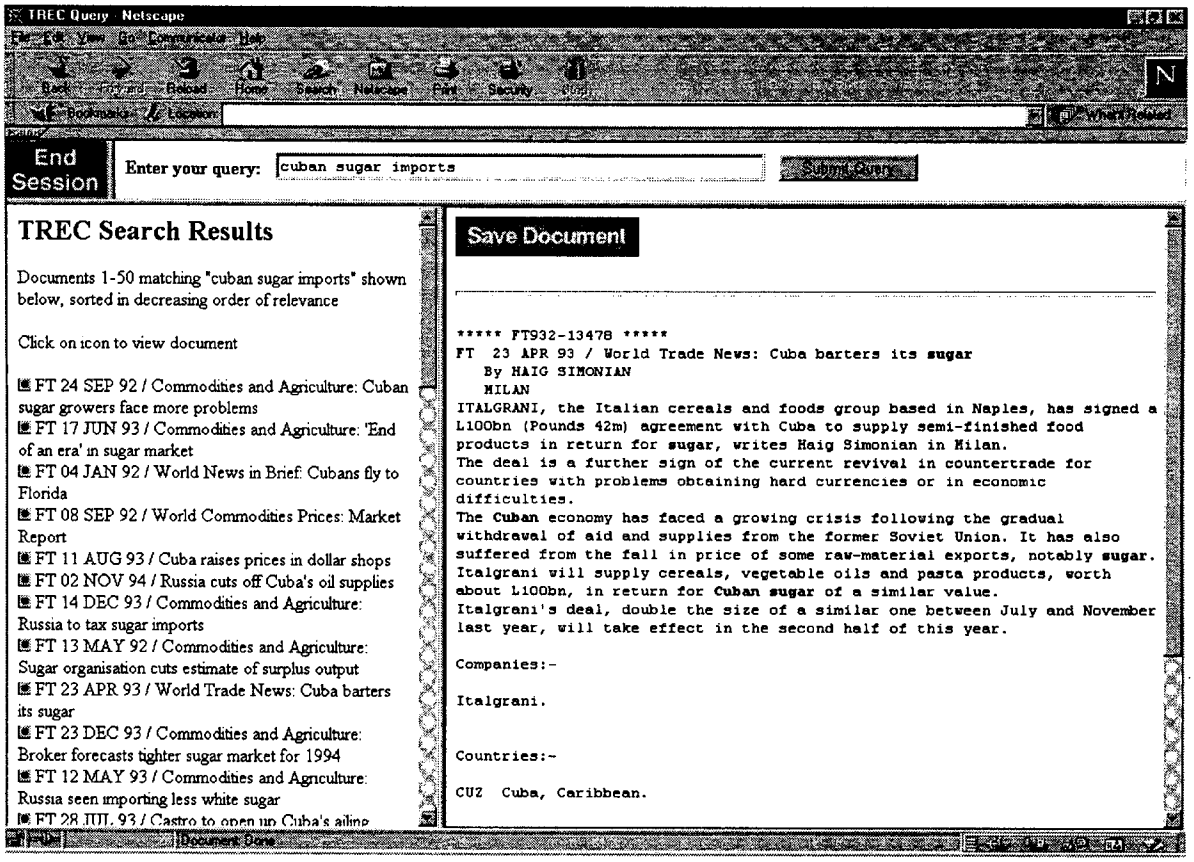


Figure 2 – Searching interface.

After the experiments were completed, data was organized into a per-question format with all associated attributes. Our initial analysis used instance recall as the dependent variable influenced by the data points in Table 2 as independent variables. 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. Residuals were examined for deviations from normality. All analyses were run in Version 6.12 of SAS for Windows 95.

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 3 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 4 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 by Figure 3, all of the difference between the systems occurred in just one query, 414i, which is one of the queries shown in Figure 1.

4. Verifying weighting scheme with current 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.

	Instance Recall	Instance Precision
System		
Baseline	0.33	0.74
Okapi	0.39	0.80
Type		
Librarian	0.36	0.76
Graduate Student	0.36	0.78

Table 3 - Instance recall and precision across systems and user types.

Source	P-value
System	0.226
Topic	0.052
Type	0.914
ID(Type)	0.052
System * Topic	0.027
System * Type	0.088
Topic * Type	0.108

Table 4 - Summary of analysis of variance model for instance recall.

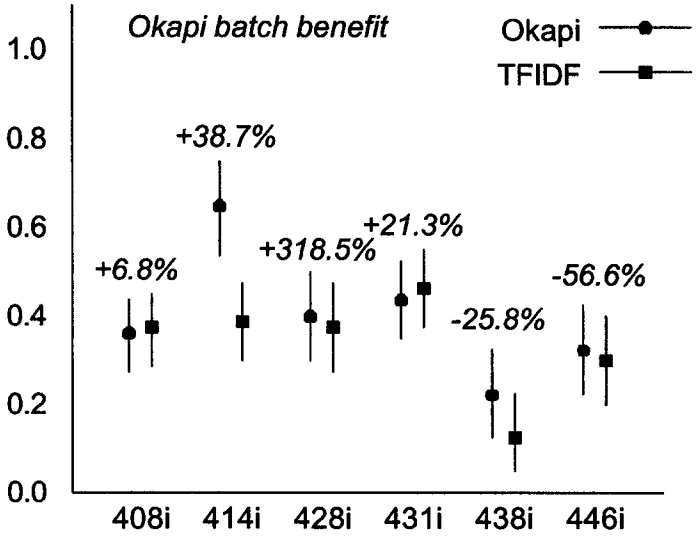


Figure 3 - Instance recall for each topic. The point values show the mean and confidence intervals for users with Okapi (circular point) and TFIDF (square point) weighting. In italics are the change in average precision for Okapi over TFIDF weighting.

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.

Results

Table 5 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 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.

5. Assessment of other factors predictive of searching success

Since the IR system used was not predictive of instance recall, we next looked at all of the variables listed in Table 2 to see if any of them was associated with successful searching.

Methods

All of the variables in Table 2 for each search 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.

Results

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. Saved - the number of documents saved by the user as containing an instance ($p < .001$)
2. DocRec - document recall (with document relevance defined as one containing one or more instances) ($p < .001$)
3. Seen - the number of documents seen by the user ($p = .002$)

Figures 4a-4c 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 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 ($p < .001$). 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 doing the search ($p = .003$) and confidence that all instances were identified ($p = .01$) were, however, associated with successful searching.

6. Discussion

Our experiments show that batch and user searching experiments do not give the same results. This outcome is limited by the fact that we only assessed one type of user searching with only six queries. Nonetheless, it 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 work 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, 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.

The results of this paper 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 is 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.

References

[1] J. Chin, V. Diehl, and K. Norman, Development of an instrument measuring user satisfaction of the human-computer interface, *Proceedings of CHI '88 - Human Factors in Computing Systems*, New York, 213-218, 1988.

[2] C. Cleverdon and E. Keen, Factors determining the performance of indexing systems, Cranfield UK: Aslib Cranfield Research Project 1966.

[3] D. Harman, Overview of the first Text REtrieval Conference, *Proceedings of the 16th Annual International ACM Special Interest Group in Information Retrieval*, Pittsburgh, 36-47, 1993.

[4] W. Hersh, Relevance and retrieval evaluation: perspectives from medicine, *Journal of the American Society for Information Science*, 45: 201-206, 1994.

[5] E. Lagergren and P. Over, Comparing interactive information retrieval systems across sites: the TREC-6 interactive track matrix experiment, *Proceedings of the 21st Annual International ACM SIGIR Conference on Research And Development in Information Retrieval*, Melbroune, Australia, 162-172, 1998.

[6] C. Meadow, Relevance?, *Journal of the American Society for Information Science*, 36: 354-355, 1985.

[7] S. Robertson and S. Walker, Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval, *Proceedings of the 17th Annual International ACM Special Interest Group in Information Retrieval*, Dublin, 232-241, 1994.

[8] A. Singhal, C. Buckley, and M. Mitra, Pivoted document length normalization, *Proceedings of the 19th Annual International ACM Special Interest Group in Information Retrieval*, Zurich, Switzerland, 21-29, 1996.

[9] K. Sparck-Jones, *Information Retrieval Experiment*. London: Butterworths, 1981.

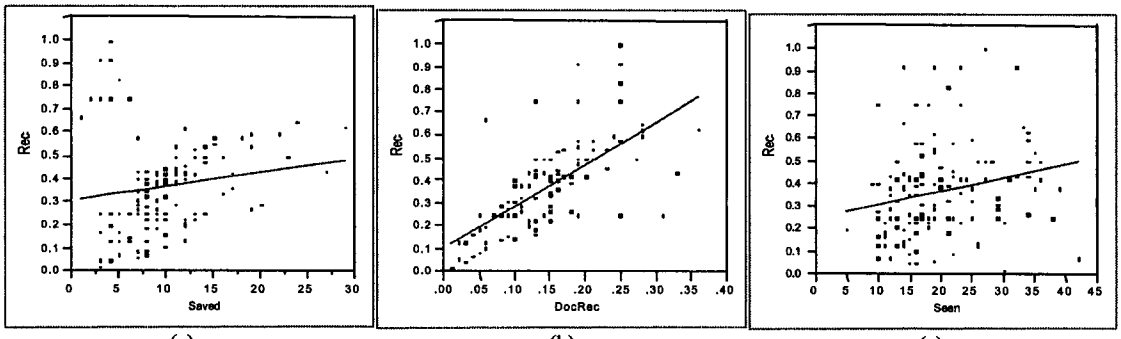
[10] D. Swanson, Information retrieval as a trial-and-error process, *Library Quarterly*, 47: 128-148, 1977.

[11] I. Witten, A. Moffat, and T. Bell, *Managing Gigabytes - Compressing and Indexing Documents and Images*. New York: Van Nostrand Reinhold, 1994.

[12] J. Zobel and A. Moffat, Exploring the similarity space, *SIGIR Forum*, 32: 18-34, 1998.

Query	Instances	Rel. 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%

Table 5 - Average precision and improvement for each query in the batch runs with TREC-8 data.



(a) (b) (c)

Figures 4a-4c - Linear fit of relationship between instance recall and (a) number of documents saved, (b) document-level recall, and (c) number of documents seen.