Further Analysis of Whether Batch and User Evaluations Give the Same Results With a Question-Answering Task

William Hersh, Andrew Turpin, Lynetta Sacherek, Daniel Olson
Susan Price, Benjamin Chan, Dale Kraemer
Division of Medical Informatics & Outcomes Research
Oregon Health Sciences University
Portland, OR, USA

In the TREC-8 Interactive Track, our results indicated that the better performance obtained in batch searching evaluation do not translate into better performance by users in an instance recall task. This year we pursued this investigation further by performing the same experiments using the new question-answering task adopted in the TREC-9 Interactive Track. Our results once again show that better performance in batch searching evaluation does not translate into gains for real users.

A continuing unanswered question in information retrieval (IR) research is whether batch and user searching evaluations give the same results. We explored this question in the TREC-8 Interactive Track, where we found that the better results obtained in batch studies using the Okapi weighting scheme over the standard TFIDF approach did not accrue to real users for an instance recall task.[1] This work was limited by the small number of queries as well as the use of a single retrieval task, the recall of specific instances for a topic. Since the TREC-9 Interactive Track would be using a different task - question-answering - we decided to use the same research question again with this changed task. Although we would still have a small number of queries, it would provide another IR task to assess this research question.

As with the TREC-8 Interactive Track we performed three experiments. The first experiment was to identify an IR approach that achieved the best possible performance in the batch environment. In the second experiment, we used that best weighting measure as the "experimental" system to be compared with the "control" system using baseline TFIDF weighting. In the final experiment, we verified that the better batch performance of the experimental system held up with the new TREC-9 Interactive Track data.

Experiment 1 - Identifying the "best" weighting scheme

In TREC-8, the best weighting scheme was chosen by turning Interactive Track data from TREC-6 and TREC-7, which also used an instance recall task, into a test collection. All documents which had one or more instances were deemed relevant, and many runs using variants of TFIDF, Okapi, and pivoted normalization were used. The collection was that used by the instance recall task, the Financial Times 1991-1994 (FT91-94) from Disk 4 of the TREC CD-ROMs. The queries used were derived from the Description field of the topic. The Okapi weighting gave the best mean average precision (MAP), which was 83% over the standard TFIDF baseline.

All of our batch and user experiments used the MG retrieval system. [2] 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. [3] 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 of each triple 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. [3] 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 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. [4] Varying the fifth character alters the document length normalization scheme. Letters greater than H use pivoted normalization. [5]

Methods

For the question-answering task of the TREC-9 Interactive Track, we had no prior Interactive Track data to use. Instead, we used almost all queries from the ad hoc collection dating back to TREC-2 (051-450) along with 20 prior instance recall queries (from the past three years of the Interactive Track) and the 200 queries from the TREC-8 question-answering track. For the latter, we deemed any document which had an answer string as relevant. Mean average precision was calculated using the trec_eval program.

Results

While the version of Okapi used in TREC-8 (AB-BFD-BAA) did better on instance recall queries from past Interactive Track experiments (using the FT91-94 collection), it did not perform as well on the other query-collection sets. The weighting scheme giving the best results over all of the query sets was a version of Okapi that employed pivoted normalization (AE-BFM-ABA) as shown in Table 1.

The new best Okapi weighting calculates the similarity between a document and query as

$$\sum_{t \in T_{q,d}} f_{q,t} \times \ln\left(\frac{N - f_t}{f_t}\right) \times \frac{f_{d,t}}{\left(f_{d,t} + \frac{W_d}{av(W_d)}\right)}$$

where

W_d	$(1 - slope) + slope \times f_d$
$av(W_d)$	average W_d over all documents
N	number of documents in the collection
f_t	number of documents containing term t
$f_{q,t}$	frequency of the term in the query
$f_{d,t}$	frequency of the term in the document
$T_{q,d}$ =	Set of terms both in q and d

Table 1 - Batch results for ad hoc, instance recall, and question-answering tasks using cosine TFIDF, Okapi weighting, and Okapi + pivoted normalization weighting.

Query set	Collection	Cosine	Okapi (% improvement)	Okapi + Pivoted normalization (% improvement)
303i-446i	FT91-94	0.2281	0.3753 (+65)	0.3268 (+43)
051-200	Disks 1&2	0.1139	0.1063 (-7)	0.1682 (+48)
202-250	Disks 2&3	0.1033	0.1153 (+12)	0.1498 (+45)
351-450	Disks 4&5 minus CR	0.1293	0.1771 (+37)	0.1825 (+41)
001qa-200qa	Disks 4&5 minus CR	0.0360	0.0657 (+83)	0.0760 (+111)
Average improvement			(+38)	(+58)

Table 2 - Mean average precision for various slopes, with 0.6 obtaining the best results.

Slope	Mean Average Precision
0.550	0.0740
0.575	0.0781
0.600	0.0782
0.650	0.0780
0.675	0.0776

and the sum is over all terms that occur both in the query and document.

As this new Okapi approach uses pivoted normalization, we needed to determine the best slope. As shown in Table 2, a slope of 0.6 was determined to be best.

The baseline TFIDF Q-expression was the same as for TREC-8 (BB-ACB-BAA), which calculates similarity between a document and the query as

$$\sum_{t \in T_{q,d}} \frac{\left(1 + \ln\ f_{d,t}\right) \times \ln\left(1 + \frac{N}{f_t}\right)}{\sqrt{\sum_{t \in doc} \left(1 + \ln\ f_{d,t}\right)^2}}$$

where

N	number of documents in the collection
f_t	number of documents containing term t
$f_{q,t}$	frequency of the term in the query
$f_{d,t}$	frequency of the term in the document
$T_{q,d}$ =	Set of terms both in q and d

Experiment 2 - Interactive retrieval experiments

Based on the results from Experiment 1, the goal of our interactive experiment was to assess whether the AE-BFM-ABA weighting scheme provided benefits to real users in the TREC interactive setting. The OHSU TREC-9 Interactive Track experiments were carried out according to the consensus protocol developed for the track. 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.

Methods

As noted above, the TREC-9 Interactive Track used a question-answering task. A set of eight questions was developed (see Table 3). Questions were of two types. The first type required users to find a small number of instances for a topic, e.g., the number of parks in the United States containing redwood trees. The second type required users to select the correct answer from two given, e.g., which country had a larger population, Denmark or Norway. Searchers from all sites were asked to answer the questions by searching, recording the answer, and recording all documents that contributed to the answer. Assessors at NIST scored each answer as being completely correct, partially correct, or not correct, with the documents saved by the user being judged as completely answering the question, partially answering the question, or not answering the question. For our analysis, a question was deemed correct if the assessor found the answer completely correct and the answer was supported by all documents saved by the user.

The collection used for these experiments was the same as that used by the question-answering track, consisting of Disks 4 and 5 (minus the Federal Register) of the TREC CD-ROM collection.

Both the baseline and the Okapi plus pivoted normalization systems used the same Web-based, natural language interface shown in Figure 1. MG was run on a Sun Enterprise 250 with 1 gigabyte of RAM running the Solaris 2.7 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 Internet Explorer or Netscape Navigator.

Table 3 - Questions for interactive question-answering task.

1. What are the names of three US national parks where one can find redwoods?
2. Identify a site of Roman ruins in present day France?
3. Name four films in which Orson Welles actually appeared.
4. Name 3 countries that imported Cuban sugar during the period of time covered by the collection.
5. Which children's TV program was on the air longer, the original Mickey Mouse Club or the original Howdy Doody Show?
6. Which painting did Edvard Munch complete first, "Vampire" or "Puberty"?
7. Which was the last dynasty of China: Qing or Ming?
8. Is Denmark larger or smaller in population than Norway ?

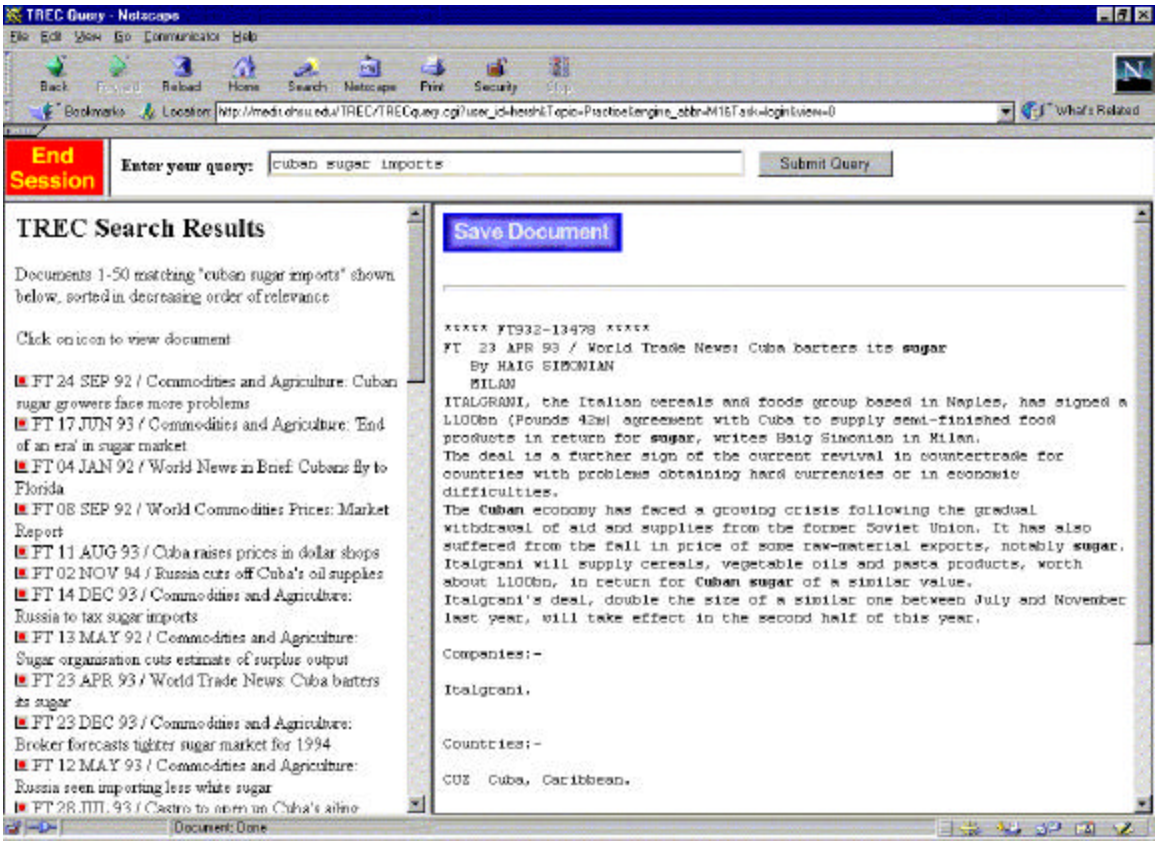

Figure 1 - Searching interface for baseline and Okapi weighting systems.

Subjects 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. We also recruited graduate students from the Master of Science in Medical Informatics Program at Oregon Health Sciences University (OHSU). They had a variety of backgrounds, from being a physician or other health care professionals to having completed only undergraduate studies.

The experiments took place in a computer lab. Each session took two 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 entire process included the following steps:

1. Orientation to experiment (10 minutes)
2. Administration of Pre-Search Questionnaire (10 minutes)
3. Orientation to searching session and retrieval system (10 minutes)
4. Practice search (10 minutes)
5. Short Break (5 minutes)
6. Searching on first 4 topics with assigned system (30 minutes)
7. Short break (10 minutes)
8. Searching on second 4 topics with assigned system (30 minutes)
9. Administration of Exit Questionnaire (5 minutes)

Each participant was assigned to search four questions in a block with one system followed by four questions 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.)

Per the consensus protocol, each participant was allowed five minutes per question. Participants were instructed to write their answer on the searcher worksheet and save all documents that supported their answers (either by using the "save" function of the system or writing its document identifier down on the searcher worksheet). The results of several participants had to be discarded for failing to follow these instructions.

The exit questionnaire was augmented from the consensus protocol to include the Questionnaire for User Interface Satisfaction (QUIS) 5.0 instrument [6]. 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.

For statistical analysis, we fit a series of mixed-model analysis of variance models and covariance models to the data. Mixed models allow both fixed effects (system and questions) and random effects (subjects) to be fit in one model. Given the binary outcome (correct or not correct), we fit a logistic model using a generalized linear model approach. We fit the model using SAS® Version 8.0 MACRO GLIMMIX, which uses an iteratively reweighted likelihood approach to fit these models. [7]

Our base model included systems (TFIDF and Okapi plus pivoted normalization) and questions. In addition to system and questions, since each subject answered all questions, we included subject in the model as a random intercept term. We also allowed a separate variance structure for each subject using the mixed model approach. In additional analyses we also added one of 11 covariates to the analysis of variance model (one covariate per analysis) to determine if the covariate made a significant contribution to the model with systems and questions. The covariates represented the factors measured in the various questionnaires and were each based on a Likert scale with values of one to five used as scale variables. The covariates and the variables they represent are listed in Table 4.

Table 4 - Covariates and the variables they represented.

Covariate	Definition
Familiar	User familiar with topic of question
Certainty	User certainty of answer
Easy Start	Easy to get started on question
Easy To Do	Question easy to answer
Satisfied	User satisfied system helped answer question
Time Adequate	Time was adequate to answer question
Terms	Number of unique terms used in all searchers for question
Cycles	Number of search cycles for question
Viewed	Number of document surrogates viewed for question
Seen	Number of documents for which full-text viewed for question
Saved	Number of documents saved as answering questions

Results

A total of 25 individuals followed instructions well enough for their data to be included in the analysis. Although a "pure" statistical analysis would only include the 16 subjects who have been balanced for query and system order, we have included the results from all 25 searchers in this initial analysis. The make-up of the participants was 18 librarians and seven others who were graduate students or research assistants. The average age of the librarians was 38.6 years. All but three were female. The average age of the remaining subjects was 34.4 years, with four males and three females.

The Pre-Search Questionnaire showed this was a group with a great deal of searching experience. Most had been searching for over half of their adult life. Virtually all reported high experience with point-and-click user interfaces, on-line library card catalogs, on-line searching, and Web searching. All indicated they frequently conducted searches and enjoyed doing it. Because of the heterogeneity of this data, no further analysis of this per-user data was performed. We instead focused our analysis on attributes measured on a per-question basis as described below.

The rate of correctness varied widely across the questions. Table 5 shows the results for each question based on the correctness criteria defined above, with results shown for all participating groups and OHSU searchers only. For the statistical analysis, we deleted two of the eight questions (numbers 3 and 8) because all searchers gave the same answer. Including a question in the analysis for which all subjects have the same answer, either correct or incorrect, causes problems for the iterative statistical algorithm. No subject answered either of the two deleted questions correctly. No question was answered correctly by all subjects. For OHSU searchers, the differences across questions was statistically significant using a Chi-square test ($p < .0001$). The rate of correctness did not vary, however, across systems. As shown in Table 6, it was virtually identical for the two retrieval systems. There was no statistically significant difference between systems.

Table 5 - Results for each question for all participants and OHSU participants only.

	All Groups			OHSU only		
Question	Incorrect	Correct	% Correct	Incorrect	Correct	% Correct
1	99	8	7.5%	21	4	16.0%
2	80	18	18.4%	20	5	20.0%
3	103	3	2.8%	25	0	0.0%
4	77	29	27.4%	10	15	60.0%
5	41	65	61.3%	5	20	80.0%
6	59	41	41.0%	6	19	76.0%
7	28	77	73.3%	4	21	84.0%
8	92	9	8.9%	25	0	0.0%
Total	579	250	30.2%	116	84	42.0%

Table 6 - Results for each question per system.

	TFIDF			Okapi + Pivoted Normalization		
Question	Searches	#Correct	%Correct	Searches	#Correct	%Correct
1	13	3	23.1%	12	1	8.3%
2	11	0	0.0%	14	5	35.7%
3	13	0	0.0%	12	0	0.0%
4	12	7	58.3%	13	8	61.5%
5	12	9	75.0%	13	11	84.6%
6	15	13	86.7%	10	6	60.0%
7	13	11	84.6%	12	10	83.3%
8	11	0	0.0%	14	0	0.0%
Total	100	43	43.0%	100	41	41.0%

The results of the analyses of covariance are shown in Table 7. None of the variables assessed were statistically significant by system and all were statistically significant by question. The latter, of course, represented the large variation in rate of correctness per question. There was a significant association with the following covariates: certainty, easy to do, satisfied, time adequate, seen, and saved. For satisfied and time adequate, the inclusion of the covariate resulted in a change in the p-value for questions. While this p-value was still significant at a 5% level, the p-values were much closer to the 5% than without the covariate. This suggests that the covariate was explaining some of the variation formerly explained by questions alone. There did not appear to a meaningful association between the other six covariates and the likelihood of being correct.

Experiment 3 - Verifying "best" weighting scheme

The final experiment was to determine whether the question-answering data for the TREC-9 Interactive Track gave better results in batch searching evaluation. This would allow us to determine whether the user evaluation in Experiment 2 gave the same or different results than batch searching experiments.

Table 7 - Summary of p-values for base analysis of variance model and model with each potential covariate added to model individually.

Covariate	System	Questions	Covariate
None	0.73	<0.0001	N/A
Familiar	0.76	<0.0001	0.70
Certainty	0.92	<0.0001	<0.0001
Easy Start	0.82	<0.0001	0.18
Easy To Do	0.82	0.0021	<0.0001
Satisfied	0.76	0.034	<0.0001
Time Adequate	0.88	0.030	<0.0001
Terms	0.98	<0.0001	0.096
Cycles	0.94	<0.0001	0.44
Viewed	0.86	<0.0001	0.13
Seen	0.59	<0.0001	0.0417
Saved	0.23	<0.0001	<0.0001

Methods

For this experiment, we developed a test collection consisting of the collection used for Experiment 2, queries derived from the question statement, and relevant judgments derived by designating those determined to "support" the answer by NIST assessors. In their judgment of the results, the assessors selected the correct answers as well as listing which documents provided "supporting" evidence for those answers. We assumed these documents were relevant and used them in our batch experiments accordingly.

Results

As shown in Table 8, the Okapi (AE-BFM-ABA) weighting provided improved MAP over TFIDF for all but on query and by an overall average of 31.5%. This was similar to our TREC-8 Interactive Track experiments, where batch results showed improved performance for the better weighting scheme that did not occur with user experiments.

Conclusions

Our TREC-9 Interactive Track results paralleled our TREC-8 results, i.e., performance enhancement that occurred in batch evaluation studies was not associated with performance of real users. As with our TREC-8 Interactive Track study, this one had limitations as well. Like past experiments, the number of queries and users was small. Recent research suggests that evaluation measures are unstable when less than 25 or 50 queries are used in an evaluation, at least in the batch setting. [8] Nonetheless, there is some significance to the fact that comparable results have been obtained with two different retrieval tasks even with the small number of queries and users.

A number of factors were assessed to determine their effect on the rate of correctness, but the large variation in question correctness overwhelmed any differences in effects of the factors.

The next step in our research will be an investigation to determine why gains in batch evaluation performance do not occur in real user studies. There are really two possibilities: either real users do not get the kind of improved recall and precision seen in batch studies with the queries that they enter or they do get better recall and precision in their searches but it does not translate into better user performance with the specific task. We will assess this by calculating recall and precision on the actual queries entered by users to determine whether systems using Okapi weighting provide benefit to them.

Table 8 - Batch searching results.

Question	TFIDF	Okapi + Pivoted Normalization	% improvement
1	0.1352	0.0635	-53.0%
2	0.0508	0.0605	19.1%
3	0.1557	0.3000	92.7%
4	0.1515	0.1778	17.4%
5	0.5167	0.6823	32.0%
6	0.7576	1.0000	32.0%
7	0.3860	0.5425	40.5%
8	0.0034	0.0088	158.8%
Mean	0.2696	0.3544	31.5%

Acknowledgements

This study was supported in part by Grant LM06311 of the U.S. National Library of Medicine.

References

[1] Hersh W, et al. *Do batch and user evaluations give the same results?*, in *Proceedings of the 23rd Annual International ACM Special Interest Group in Information Retrieval.* 2000. Athens, Greece: ACM Press, 17-24.

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

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

[4] Robertson S and Walker S. *Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval*, in *Proceedings of the 17th Annual International ACM Special Interest Group in Information Retrieval.* 1994. Dublin: Springer-Verlag, 232-41.

[5] Singhal A, Buckley C, and Mitra M. *Pivoted document length normalization*, in *Proceedings of the 19th Annual International ACM Special Interest Group in Information Retrieval.* 1996. Zurich, Switzerland: ACM Press, 21-9.

[6] Chin J, Diehl V, and Norman K. *Development of an instrument measuring user satisfaction of the human-computer interface*, in *Proceedings of CHI '88 - Human Factors in Computing Systems*. 1988. New York: ACM Press, 213-8.

[7] Wolfinger R and O'Connell M, *Generalized linear mixed models: a pseudo-likelihood approach.* Journal of Statistical Computation and Simulation, 1993. 48: 233-43.

[8] Buckley C and Voorhees E. *Evaluating evaluation measure stability*, in *Proceedings of the 23rd Annual International ACM Special Interest Group in Information Retrieval.* 2000. Athens, Greece: ACM, 33-40.