Why Batch and User Evaluations Do Not Give the Same Results

Andrew H Turpin
School of Computing
Curtin University of Technology
Perth, WA, Australia
+61 8 9266 3014
andrew@cs.curtin.edu.au

William Hersh
Division of Division of Medical Informatics
& Outcomes Research
Oregon Health Sciences University
Portland, OR, USA
+1 503 494 4563
hersh@ohsu.edu

ABSTRACT

Much system-oriented evaluation of information retrieval systems has used the Cranfield approach based upon queries run against test collections in a batch mode. Some researchers have questioned whether this approach can be applied to the real world, but little data exists for or against that assertion. We have studied this question in the context of the TREC Interactive Track. Previous results demonstrated that improved performance as measured by relevance-based metrics in batch studies did not correspond with the results of outcomes based on real user searching tasks. The experiments in this paper analyzed those results to determine why this occurred. Our assessment showed that while the queries entered by real users into systems yielding better results in batch studies gave comparable gains in ranking of relevant documents for those users, they did not translate into better performance on specific tasks. This was most likely due to users being able to adequately find and utilize relevant documents ranked further down the output list.

Categories and Subject Descriptors

Evaluation, Building Test Collections, User Studies, Experimental Design and Metrics

General Terms

Algorithms, Performance

Keywords

Information retrieval evaluation, Text Retrieval Conference (TREC), interactive retrieval

1. INTRODUCTION

A common view held by many information retrieval (IR) researchers is that "batch" retrieval evaluations are reliable in determining the efficacy of one retrieval system versus another [1]. In batch evaluations, a test collection with fixed topics, documents, and relevance judgments is used to simulate the activity of real searchers. This approach has been used since the 1960s in the Cranfield experiments [2] and has continued through the Text Retrieval Conference (TREC) [3]. Some researchers have challenged this approach to system evaluation, maintaining that real world searching is more complex and that system efficacy cannot be accurately assessed with such studies. These authors point out that relevance is not a fixed notion [4], interaction is the key element of successful retrieval system use [5], and relevance-based measures do not capture the complete picture of user performance [6]. If batch searching results cannot be generalized, then system design decisions based on them are potentially misleading.

Unfortunately, there has been little research to address this problem. We have begun to address it in the context of the TREC Interactive Track. Our results have challenged the view that batch searching results are applicable to real-world searching. Our first experiments reported that not only did a system performing better in batch results not yield better results with a real user task, but that there was also no correlation between batch and searching results on most topics [7].

There were three criticisms of this study. First, the number of queries was small. Some recent results have shown that, at least in the case of recall-precision metrics, stability of results is not achieved when there are less than 25-50 queries [8]. The TREC-8 Interactive Track utilized only six queries [9]. Second, there was only one type of user task, namely the instance recall task that was used in the interactive track from TREC-6 to TREC-8. Finally, there was no data provided that explained why batch and user queries did not give the same results.

These criticisms have been addressed in two ways. First, we assessed whether batch and user evaluations give the same results with a new user task and eight new topics in the new question-answering task of the TREC-9 Interactive Track [10]. Second, we have carried out further analysis of our data, aiming to determine the reason for our results from both TREC-8 and TREC-9. This paper reports on the latter, analyzing the data from these two experiments to determine why batch and user evaluations give divergent results.

2. THE INTERACTIVE TRACK EXPERIMENTS

We first provide an overview of our TREC Interactive Track experiments, which provide a context for the analysis reported in this paper. The TREC Interactive Track is an activity within TREC where research groups instruct human users to perform a

SIGIR'01, September 9-12, 2001, New Orleans, Louisiana, USA.
Copyright 2001 ACM 1-58113-331-6/01/0009…$5.00.

designated user task. Each group uses the same task, the same document collection and the same information needs (“topics”). Once the experiments are complete, the documents that users deem relevant for each topic are sent to NIST for independent relevance judgment. Upon return of the relevance judgments, various metrics regarding the user’s performance on the task can be calculated for each research group.

The TREC-6 through TREC-8 Interactive Tracks employed an instance recall task, where users were asked to find “instances” relating to a topic within a 20 minute period [9]. The TREC-9 Interactive Track changed the user task to question-answering, where users were required to give explicit answers to topic questions within a 5 minute period[11].

The goal of both of our experiments within this framework was to assess whether IR approaches achieving better performance in batch evaluations could translate that effectiveness to real users. Both experiments consisted of the following three stages.

Stage 1: batch experiments. This stage set out to identify two ranking schemes using an underlying retrieval engine based on the vector space retrieval model. The resulting two systems were dubbed *baseline* and *improved.* The *baseline* system was fixed as a basic Cosine TF*IDF weighting scheme, the exact details of which are shown in the Appendix. The *improved* system was chosen as the system with the greatest improvement in mean average precision (MAP) over the baseline system as calculated from a batch run of a set of topics against a document collection. These batch runs were designed to mimic IR experiments as they are typically reported in venues such as TREC and SIGIR. Queries were generated automatically from topic descriptions, fed into the IR system, and the resulting top 1000 document descriptors matched against relevance judgments to calculate MAP.

In order to choose an improved system, we took on the role of a naïve purchaser of an IR system, choosing the system with the greatest improvement in MAP on the collection and topics that were as similar as possible to the actual collections and topics used in the Interactive Track experiments in the two subsequent stages. The limitation, of course, was that only collections and topics that have relevance judgments could be employed in the batch experiments, so the actual topics and collection could not be used. This process mimics a “real world” application of IR batch experimental results, where exact queries, and perhaps even collections, are not known in advance of a system being deployed.

Stage 2: user experiments. Once the two systems were chosen by Stage one, we sat a group of users in front of each system to answer the Interactive Track topics. Our user group was composed of 13 medical librarians and 12 graduate students with a health science background. Each user was asked to fulfill the requirements of each search topic using one of the two systems in the allotted time limit. The assignment of topic-system pairs to each user was randomized subject to the constraints that each user answered the same number of topics with either system, each topic was answered in equal numbers by each system, and each topic was answered by the same number of librarians and students. Users were not aware which system was baseline and which was improved, although they were aware that they were using two different systems.

The interface provided was a simple Web-based natural language searching interface to the MG system [13] that was identical for all users and systems. The single browser window contained three frames: one a query entry box, the second a list of document titles, and the third a display area for the full text of a document. Users could enter a query in the query box, whereupon a list of document titles ranked in order of relevance according to the weighting scheme of the appropriate system would appear in the title list section of the window. The user could then open the full text of the document by clicking on its title. Users were required to record any document they thought relevant to the topic both on paper, and by clicking a “Save Document” button on the browser window. All user actions were recorded in a log file.

Stage 3: performance assessment. Upon receipt of the relevance information from NIST, the user’s performance with each system was calculated. Furthermore, the batch experiments from Stage 1 were performed on the actual topics and collections used in the user trials of Stage 2. Examining the batch results on the actual topics and collection used provides a safety check to ensure that our original decision on an improved system in Stage 1 was not wildly inaccurate.

2.1 TREC-8 instance recall results

The TREC-8 interactive track used the task of instance recall to measure success of searching. Instance recall was defined as the number of instances of a topic retrieved [9]. For example, a searcher might be asked to identify all the discoveries made by the Hubble telescope; in this case each discovery was an instance, and the proportion of instances correctly listed was instance recall. This was in contrast to document recall, which was measured by the proportion of known relevant documents retrieved. Instance recall is probably a more pertinent measure of user success in this IR task, since users are less likely to want to retrieve multiple documents covering the same instances.

The common experimental protocol, used by all groups participating in the track, required the use of the Financial Times 1991-1994 database and the following (abridged) six topics [9]:

1. What tropical storms hurricanes and typhoons have caused property damage and or loss of life?
2. What countries import Cuban sugar?
3. What countries other than the US and China have or have had a declining birth rate?
4. What are the latest developments in robotic technology and it use?
5. What countries have experienced an increase in tourism?
6. In what countries have tourists been subject to acts of violence causing bodily harm or death?

Users were asked to identify and save at least one document for as many instances as they could identify in 20 minutes for each topic.

Stage 1 of this experiment identified the Okapi weighting scheme [12] as the improved system (see Appendix for exact weighting formula), with an 81% improvement in MAP over the baseline system. These batch experiments were carried out on the same document collection as the user experiments, Financial Times 1991-1994, using 14 topics with relevance judgments from the previous two TREC Interactive Tracks, which also employed an instance recall task. In Stage 2, twelve librarians and twelve graduate students searched on each of the six topics. While users of the Okapi-based system had 15% better instance recall, all of the improved performance came from just one of the six topics and the overall difference was not statistically significant. Stage 3 of this experiment verified that the performance of the improved system over baseline held up (by 18%) with the new TREC-8 Interactive Track topics and relevance judgments.

2.2 TREC-9 question-answering results

In the TREC-9 Interactive Track, we asked the same research question and applied the same methodology with a different user task, and eight new topics on a different collection [10]. The new user task was question-answering, with two different types of questions [11]. The first type of question required users to find a small number of answers for a topic, e.g., the number of parks in the United States containing Redwood trees. The second type asked users to select the correct answer from two given, e.g., which country had a larger population, Denmark or Norway. Per the common experimental protocol, there were eight questions, with half of each type. The eight questions were:

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? The document collection was several times larger than the first experiment, containing not only the Financial Times database, but also newswire from the Associated Press, Wall Street Journal, San Jose Mercury News, and Los Angeles Times.

As there was no previous Interactive Track question-answering data to employ on Stage 1 of our three stage methodology, we performed the batch experiments with all previous TREC topics and relevance judgments (including the Interactive, Ad Hoc, and Question Answering tracks). In these experiments, the improved system was found to be Okapi with a pivoted normalization component [14]. This approach achieved over 65% improvement in MAP above the baseline. The slope of the pivot component was tuned empirically on these TREC topics and collections (see Appendix for exact weighting formula).

In the second stage, twelve graduate students and thirteen librarians searched on each of the eight topics using the same Web-based natural language searching interface as described above.

For this task, 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 preliminary 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. Using this performance measure, the user's rate of answering questions correctly per the common protocol was a statistically non-significant 6% lower with the improved system. The final stage verified that the performance of the improved measure over baseline held up (by 32%) with TREC-9 Interactive Track topics and relevance judgments.

2.3 Summary

Table 1 summarizes the results of the batch and user experiments of Stage 3 once relevance data is known for the topics and collections employed in the user experiments. The batch evaluations performed in Stage 3 of each of the experiments confirm that the systems performed differently in a batch setting for both experiments as they are commonly measured in venues such as TREC and SIGIR. We note that the results were not statistically significant, but this is not surprising due to the small number of topics. We address this issue below in our analysis. Users, however, performed equally well with both systems, with the paired t-tests indicating that any differences were likely due to chance. This statistical difference in the user studies is more meaningful than in the batch environment since the analysis was based on all user-system pairs and as a result has a much larger sample size.

3. ANALYSIS

The goal of this paper was to determine why users did not achieve better outcomes in searching tasks with the improved systems, where an improved system is defined as one which gives better performance with relevance-based metrics such as mean average precision in a batch setting. There are two possible explanations why user performance was not enhanced by the improved system:

1. The systems studied do not give better performance with real users. That is, the actual queries entered by users for each topic do not lead to improved relevance-based metrics, unlike the queries used in the batch studies.

2. The systems studied do give better performance with relevance-based metrics, but this does not translate to improvement in performance with the searcher's task.

Table 1 - Batch and user performance on the two tasks with the two systems averaged over all topic-system pairs.

	Instance recall experiment		Question-answering experiment	
	Batch mean average precision	User instance recall	Batch mean average precision	User % of questions correct
Baseline system	0.2753	0.3230	0.2696	66%
Improved system	0.3239	0.3728	0.3544	60%
Change	+18%	+15%	+32%	-6%
p-value (paired t-test)	0.24	0.27	0.06	0.41

We carried out several analyses to address these questions. We first examined searching performance with the actual queries users entered into the systems. As our retrieval system captured all documents returned to users in a log file, we were able to calculate MAP, precision at 10 documents, and precision at 50 documents for all user queries. There is also the possibility that comparing precision results from the batch experiments with instance recall from the user experiments introduces some error. Again using the log files, we calculated the number of instances retrieved at 10 documents and at 50 documents for the TREC-8 instance recall experiment, just as could be done in a batch setting. Our analysis was done using both the average and maximum value for all queries entered by a single user for a given topic. As both yielded similar results, we report only the average values in this paper.

We next analyzed what users did with documents retrieved by looking at the average number of queries issued by users during the time available to work on each topic. The next step was to assess what the users did with documents once they clicked on the title to open the full text of the document for reading. We looked at the average rank of documents opened for reading, and then assessed the number opened that were relevant and non-relevant. Our final step was to look at the length of documents read which were retrieved by both systems.

4. RESULTS

Our first analysis showed that users unquestionably achieved better results, as measured by relevance-based metrics, using the improved system. As shown in Tables 2 and 3, the improvement in mean average precision was not only more pronounced for real users (46% and 67% respectively for the instance recall and question-answering tasks), but was also statistically significant. So while the batch runs in Stage 3 of both experiments did not overwhelmingly confirm that the improved system was superior to the baseline system, when a larger number of user queries are used as input, rather than a single automatically generated query from the topic description, the improved system is clearly superior using the MAP metric in both cases. The precision at 10 and 50 documents verified that that the improved systems were better at putting relevant documents towards the top of the ranked output.

Another potential criticism of the instance recall experiment is that the batch run uses precision as a predictor of system success, whereas user outcomes are measured using instance recall. The final two columns of Table 2, however, confirm that the improved system placed a larger number of instances towards the top of the output for the instance recall experiment.

Other measures also demonstrated the seemingly superior searching output of the improved systems. Table 4 shows that users of the improved system issued fewer queries, most likely because each query found more relevant documents. Table 5 shows that the users retrieved fewer non-relevant documents, averaged over all topic-system pairs, using the improved system.

Table 2 - Three precision based metrics and two instance recall based metrics calculated from documents returned to users during the instance recall experiment averaged over all topic-system pairs.

	Mean average precision	Precision @ 10 documents	Precision @ 50 documents	Instances @ 10 documents	Instances @ 50 documents
Baseline system	0.36	0.35	0.27	3.30	10.08
Improved system	0.53	0.55	0.36	6.77	13.11
Change	+46%	+55%	+33%	+105%	+30%
p-value (paired t-test)	0.02	0.03	0.14	0.04	0.28

Table 3 - Three precision based metrics calculated from documents returned to users during the question-answering experiment averaged over all topic-system pairs.

	Mean average precision	Precision @ 10 documents	Precision @ 50 documents
Baseline system	0.25	0.15	0.10
Improved system	0.42	0.30	0.14
Change	+67%	+97%	+48%
p-value (paired t-test)	0.001	0.001	0.019

Table 4 - Number of queries issued by users averaged over all topic-system pairs.

	Instance recall experiment	Question-answering experiment
Baseline system	3.56	4.04
Improved system	2.98	2.69
Change	-16%	-33%
p-value (paired t-test)	0.16	0.04

Table 5 - Number of documents retrieved by users averaged over all topic-system pairs.

	Instance recall experiment		Question-answering experiment	
	Relevant	Nonrelevant	Relevant	Nonrelevant
Baseline system	129.7	158.7	29.5	74.1
Improved system	131.8	103.0	29.4	48.5
Change	+2%	-35%	0%	-35%
p-value (paired t-test)	0.93	0.01	0.97	0.02

Table 6 - Rank of documents opened for reading by user averaged over all topic-system pairs.

	Instance recall experiment	Question-answering experiment
Baseline system	32.3	9.8
Improved system	25.3	9.3
Change	-22%	-5%
p-value (paired t-test)	0.28	0.71

Table 7 - Proportion of relevant documents in the top 10 not opened for reading by user averaged over all topic-system pairs.

	Instance recall experiment	Question-answering experiment
Baseline system	37.9%	29.9%
Improved system	47.4%	54.8%
Change	+25%	+83%
p-value (paired t-test)	0.220	0.002

Table 8 - Length of documents retrieved by user averaged over topic-system pairs.

	Instance recall experiment	Question-answering experiment
Baseline system	1810	1581
Improved system	4275	4362
Change	+136%	+176%
p-value (paired t-test)	0.002	<0.000

But other aspects of the analysis demonstrate that users of the baseline system were not hampered in completing their tasks by the “inferior” system. Table 5 also shows users of each system read about the same number of relevant documents. So by issuing more queries with the baseline system, users were able to find as many relevant documents as with the improved system. And as shown in Table 6, at least for the TREC-9 question-answering task, the average rank in the output list of documents opened for full-text viewing was comparable, indicating the user was able to find documents at the same rank for both systems. The average rank of opened documents was slightly higher for the baseline system in the instance recall task, but the difference was not statistically significant.

Users were also likely to achieve a saturation effect for relevant documents (which might not occur in a task where more than 15-20 minutes were available for its completion). As shown in Table 7, the number of relevant documents ranked in the top 10 that were not opened by the user showed a statistically significant difference across systems in the question-answering experiment. Averaging over all users of the same topic-system pair, 30% of the relevant documents in the top 10 were not read using the baseline system, whereas 55% were not read using the improved system. These figures were calculated by pooling all users queries for any single topic, hence they should not be biased by users choosing not to read a document if they had already read it in a previous query for that topic. Performing a similar analysis for the instance recall experiment showed a similar trend, but the results were not statistically significant. Using the baseline system, an average of 38% relevant documents that appeared in the top 10 of any query were not read, with the figure for the improved system being 47%. It appears that the extra benefit of the improved system was ignored in the question-answering experiment and at best played a small part in the instance recall experiment.

One possible explanation for users of the baseline system being able to compensate for fewer relevant documents being retrieved is based on a “feature” of the Okapi and pivoted normalization weighting schemes, which is their downgrading of the importance of shorter documents. As a result, the baseline TF*IDF system ranks shorter documents higher. Table 8 shows this to be the case, with average length of retrieved documents nearly threefold shorter. Likewise, the baseline systems retrieved six times as many documents of length less than 1,000 bytes. Thus, users in the instance recall and question-answering tasks were able to read a larger quantity of documents using the baseline system and thus compensate for fewer relevant documents being ranked highly.

5. DISCUSSION

Our TREC Interactive Track experiments demonstrate that batch searching experiments of the type typically reported in TREC and SIGIR and those involving real user tasks do not achieve the same results, at least in the setting of controlled user experiments, with short tasks and times to search. In two different user tasks, our experiments have shown that users can perform comparably with a baseline system or one shown to be significantly "better" in relevance-based batch experiments of the type typically employed in IR research.

Why might this be so? One explanation would be that users do not issue queries that take advantage of the increased performance offered by the batch systems. This is most definitely not the case, as users of the improved system issue queries that result in more relevant documents being ranked higher in the output. However, the "inferior" system does retrieve the same number of relevant documents; they just happen to be ranked lower. So while the primary research goal of these experiments was based around common batch evaluation techniques based on MAP, if the batch experiments are "perfect", using exact user queries on the exact collection and using the same outcome metric as the user experiments, the improved systems show an even greater improvement than predicted using the standard techniques.

It seems, therefore, at least in the tasks studied in these experiments, users are able to find an ample number of documents to perform their tasks regardless of the batch performance of their IR system.

Users of the baseline system had to "work harder", however, to satisfy the requirements of the tasks by issuing more queries and reading more documents than users of the improved system. However, the baseline system returned shorter documents, so perhaps the burden of reading many short documents is less or equal to the burden of reading fewer longer documents. In fact, several user satisfaction metrics compiled from questionnaires completed after searching seem to support this hypothesis. Metrics such as "queries were easy to do" or "the system was easy to use," showed no difference between the baseline and improved systems in terms of user satisfaction. Further research is required to establish the validity of this observation.

The implications of these results are significant. Much research over the last few decades has compared the "performance" of retrieval systems based on batch evaluation studies. While perhaps useful in the tuning of systems, batch studies should not necessarily lead us to conclude that one approach to indexing or retrieval is more effective than another. Clearly users can easily compensate for differences found significant in batch studies, such as higher ranking of relevant documents, on the two tasks tested here.

There are, of course, limitations to the work in this paper. We have looked at two retrieval tasks done in a laboratory setting. These tasks do not necessarily require high recall or precision. There may well be retrieval tasks for which higher recall and precision would have an impact, and future work in the TREC Interactive Track and elsewhere should address where that might be the case. For these tasks, a batch result based on MAP showing one system is superior to another may indeed translate into improved user outcomes on the superior system.

Another potential confound in these experiments is the design of the user interface, particularly the manner in which ranked document lists were presented. Users were presented with document titles upon which to make their initial judgments, but document titles may not be the best indicator of a document's content, or its relevance to the topic. We intend to conduct further analysis to determine what effect document titles had on decisions made by the users. By analyzing the log files we may be able to identify relevant documents that were consistently ignored across users, presumably because of their titles. An alternate approach would be to have the titles themselves independently judged for relevance to each topic, and examine the correlations between the judged titles, and the actions of users in selecting documents to read.

While it is clear that the users of the baseline system issued more queries than users of the improved system, we intend to examine more carefully the evolution of queries for individual topics. Questions that can be answered from the log files include: Did users begin with short queries and add more terms? Did they adapt their queries based on the documents that they read? (None of the systems in these experiments employed any automatic query expansion or relevance feedback techniques.) How did query evolution differ between the two systems? Was the reason for the higher number of queries using the baseline system due to that system returning the same set of relevant documents with different queries? Further research into factors effecting user searching performance may lead us to rethink our approach to ranking documents and evaluating IR systems.

6. ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for helpful comments that improved this manuscript.

7. REFERENCES

[1] Salton G and Buckley C, Term-weighting approaches in automatic text retrieval. Info Proc Mgmt, 1988. 24: 513-23.

[2] Cleverdon C and Keen E, Aslib Cranfield Research Project: Factors determining the performance of indexing systems (Vol. 1: Design, Vol. 2: Results). 1966: Cranfield, UK.

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

[4] Meadow C, Relevance? J Am Soc Info Sci, 1985. 36: 354-5.

[5] Swanson D, Information retrieval as a trial-and-error process. Library Quarterly, 1977. 47: 128-48.

[6] Hersh W, Relevance and retrieval evaluation: perspectives from medicine. J Am Soc Info Sci, 1994. 45: 201-6.

[7] 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.

[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.

[9] Hersh W and Over P. TREC-8 interactive track report, in Proceedings of the 8th Text REtrieval Conference (TREC-8). 2000. Gaithersburg, MD: NIST, 57-64.

[10] Hersh W, et al. Further analysis of whether batch and user evaluations give the same results with a different user task, in Proceedings of the Ninth Text Retrieval Conference (TREC-9). 2001. Gaithersburg, MD: NIST, in press.

[11] Hersh W and Over P. TREC-9 Interactive Track Report, in Proceedings of the Ninth Text Retrieval Conference (TREC-9). 2001. Gaithersburg, MD: NIST, in press.

[12] 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.

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

[14] 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.

APPENDIX

The three formula used to assign a similarity between a query *q* and a document *d* in the two experiments were:

	Q-expression	Description	Formula
Baseline	BB-ACB-BAA	Basic Cosine	$\sum_{t\in T_{q,d}} \frac{TF(t,d)\times IDF(t)}{\sqrt{\sum_{t\in T_d} TF(t,d)^2}}$
Improved TREC-8	AB-BDM-BAA	Okapi	$\sum_{t\in T_{q,d}} \frac{IDF(t)^2\times f_{d,t}}{f_{d,t}+W_d}$
Improved TREC-9	AE-BFM-ABA	Pivoted Okapi	$\sum_{t\in T_{q,d}} f_{q,t}\times\ln\left(\frac{N-f_t}{f_t}\right)\times\frac{f_{d,t}}{f_{d,t}+W_d^{'}}$

where

T_d	=	Set of terms in document *d*
$T_{q,d}$	=	Set of terms both in *q* and *d*
$TF(t,d)$	=	$(1+\ln f_{d,t})$
$IDF(t)$	=	$\ln(1+N/f_t)$
N	=	Number of documents in the collection
f_t	=	Number of documents containing term *t*
f_d	=	Number of terms in document *d*
$f_{q,t}$	=	Frequency of term *t* in query *q*
$f_{d,t}$	=	Frequency of term *t* in document *d*
W_d	=	$\sqrt{\lvert T_d\rvert}/av\left(\sqrt{\lvert T_d\rvert}\right)$
$W_d^{'}$	=	$((1-s)+s\cdot f_d)/av((1-s)+s\cdot f_d)$ and $s=0.6$
s		"Slope" used in pivoted normalization component, set to 0.6
$av(x)$	=	Average of *x* over all documents