Do Clarity Scores for Queries Correlate with User Performance?

A. Turpin

School of Computing
Curtin University of Technology
Perth, Australia.
Email: andrew@cs.curtin.edu.au

W. Hersh

Department of Medical Informatics & Clinical Epidemiology
Oregon Health and Science University
Portland, OR, USA.
Email: hersh@ohsu.edu

Abstract

Recently the concept of a *clarity score* was introduced in order to measure the ambiguity of a query in relation to the collection in which the query issuer is seeking information [Cronen-Townsend *et al.* Proc. ACM SIGIR2002, Tampere Finland, August 2002]. If the query is expressed in the "same language" as the whole collection then it has a low clarity score, otherwise it has a high score, where the similarity is the relative entropy of the query and collection models. Cronen-Townsend *et al.* show that clarity scores correlate directly with average precision, hence a query with a high clarity score is likely to produce relevant documents high in a list of resulting documents. Other authors, however, have shown that high precision does not necessarily correlate with increased user performance. In this paper we examine the correlation between user performance and clarity score. Using log files from user experiments conducted within the framework of the TREC Interactive Track, we measure the clarity score of all user queries, and their actual performance on the searching task. Our results show that there is no correlation between the clarity of a query and user performance. The results also demonstrate that users were able to slightly improve their queries, so that subsequent queries had slightly higher clarity scores than initial queries, but this was not dependent on the quality of the system they used, nor the user's searching experience.

Keywords: information retrieval, user study, entropy, language models, clarity score

1 Introduction

Most of us have issued a query to a database and got some unexpected results. For example, when issuing the query "ACSW" on the World Wide Web the highest ranking page returned may be Academic Software Inc, rather than the home page for this conference as may have been intended. In this case it is difficult for the search engine to distinguish between the many pages that contain the acronym ACSW in different contexts: the query is ambiguous. If, however, we issued the query against only ".edu.au" domains of the web, then the top 50 pages, say, would all discuss this conference. The query is only ambiguous with respect to the collection on which it is issued.

The concept of a clarity score was recently introduced in an attempt to quantify this type of ambiguity in queries [Cronen-Townsend et al., 2002]. Simply put, a clarity score is the difference between a probabilistic model of the language used in the document collection (for example, web pages on the www), and a model of the language used in the query relative to the collection. If the models are similar, then the query has a low clarity score: the query describes the entire collection. If the models are disparate, then the query identifies a subset of the collection which is likely to be the answer set for the query.

Cronen-Townsend *et al.* use a simple language model to implement their clarity scores, which is explained in detail in Section 2. They report that their clarity scores correlate closely to the *average precision* of ranked lists returned by various search engines. Precision has long been used as a metric for quantifying the performance of search engines [Salton and McGill, 1983], and is the main metric employed for evaluating systems in the IR community [Voorhees and Harman, 1999, Voorhees and Harman, 2000]. The precision of a system for a particular query is the proportion of documents relevant to that query (as judged by some third party) that appear in the ranked list. Precision can be calculated for any number of documents in the list. For example "Precision at Ten" (p@10) is the proportion of relevant documents in the top 10 of the result list. The quantity "Precision at One" is either 1.0, if the top ranked document is relevant, or 0.0 if the top ranked document is irrelevant to the query. Average precision is calculated as the mean of all precision values calculated after each relevant document in the ranked list. Typically, IR systems are ranked against each other using the mean of average precision (MAP) across fifty or more queries [Voorhees and Harman, 1996].

In the last few years, however, some researchers have shown that if users employ an IR system with a high MAP score (the "Okapi" system) they are not guaranteed to perform better than their colleagues using a system with low MAP ("Cosine") [Hersh et al., 2000, Turpin and Hersh, 2001, Turpin and Hersh, 2002]. These experiments took two IR systems, Okapi and Cosine, and ran the same queries with a different group of users for each system. Although the system with the higher MAP clearly returned more relevant documents higher in the ranked lists of results (as would be expected), the users still managed to perform their tasks as effectively with either system. Furthermore, the users did not notice any significant difference in burden between using the two systems.

The experiments were conducted as part of the Interactive Track of the TREC conferences [Voorhees and Harman 1999, 2000]. Briefly, Experiment 1 was an instance recall task, where users were instructed to find as many different answers to a question as possible in a 20 minute time limit [Hersh et al., 2000]. For example, "What tropical storms (hurricanes and typhoons) have caused property damage and/or loss of life?" For this experiment, a user's performance was quantified by *instance recall*: the number of different instances they found out of the total possible

number of known instances in the collection. The collection used was a database of Financial Times articles from 1991 to 1994, and contained 210,158 documents. There were six different questions, and two sub-groups of users (postgraduate students and librarians) [Hersh et al., 2000].

Experiment 2 was a question-answering task, where users had 5 minutes to answer questions of the type: "Which was the last dynasty of China: Qing or Ming?" The collection used in these experiments was larger, containing 978,952 documents from 6 different newspaper collections. User performance on this task was measured by whether they answered the question correctly or not. There were eight different questions [Turpin and Hersh, 2001]. Mode detail of these experiments are provided in Section 3

In both of these experiments, the users performed equally well with both systems, despite the MAP scores for the systems based on the queries issued by the users being obviously (and statistically significant) higher for the Okapi system. Given, then, that MAP may not be a good predictor of user performance with an IR system, and that clarity scores correlate closely with average precision values, it remains an open question whether clarity scores correlate with user performance. If a user issues a query with a high clarity score, are they more likely to fulfill their information need more efficiently than a user who issues queries with low clarity scores? This research investigates whether clarity scores correlate positively with user outcomes on the instance recall task and the separate question answering task. We aim to answer four questions:

1. Was there a correlation between the number of instances discovered in the instance recall task, and the clarity score of the queries issued?
2. Were the clarity scores of queries issued by users who answered the question correctly in the question answering task higher than those who answered the question incorrectly?
3. Did user's clarity scores improve over the course of the experiments as they formed a better internal model of the language used in the test collections? If so, did the system they used affect this?
4. Did librarians improve their queries more than postgraduate students over the course of the experiment?

2 Computing Clarity

Cronen-Townsend *et al.* define a "model" of a query or a document as simply a probability distribution over all *terms* in the collection. A term is a stemmed (morphologically-normalised) version of a word, so words like "search", "searching", and "searched" are all represented by the single term "search". A query model, therefore, is a probability mass function p_Q such that $p_Q(t)$ is the probability of term t occurring in the query model. Similarly, p_C is the collection model and p_D the model for a document D. The clarity score is then the difference between the query model and the collection model as measured by the Kullback-Leibler divergence, or the relative entropy of the distributions:

$$\text{clarity}(Q) = \sum_{\forall t} p_Q(t) \log_2 \frac{p_Q(t)}{p_C(t)}. \qquad (1)$$

Another way of thinking of this divergence measure is that it gives the average number of bits wasted if the query is compressed using the collection model, rather than the more accurate query model. A "clear" query should identify some specific set of terms/documents in the collection, rather than the entire collection, so the more the query and collection models differ, the higher the clarity score.

Computing clarity, therefore, requires estimating the two distributions p_Q and p_C. The easiest to estimate is p_C, as $p_C(t)$ is the probability of term t occurring in the collection. In data compression much more complex models based on word and character co-occurrence are often used, but Cronen-Townsend *et al.* report that the simple zero-order Markov term-based model gives a useful clarity score.

Estimating the query model, p_Q, is more involved. A naive approach is just to use the same process as that of estimating the collection model, a simple frequency count of terms in the query. In the case of queries with a small number of terms, like "ACSW" used in the introduction, however, the resulting model is not very useful. In this instance, every $p_Q(t)$ would be zero except for $p_Q(\text{"ACSW"})$, which would be one. Moreover, it does not model the query with respect to the collection; the very relationship the clarity score is endeavoring to measure. What is more useful is a query model defined in terms of the language used in actual documents in the collection that contain the query terms. Presumably these are the documents most likely to be returned to the user, hence represent the information contained in the collection relevant to the query.

A next approximation at a query model, therefore, is to sum all the document models for documents in the set R of all documents that contain query terms:

$$p_Q(t) = \sum_{D \in R} p_D(t). \qquad (2)$$

The probability $p_D(t)$ can be estimated by the relative frequency with which term t occurs in D. Cronen-Townsend *et al.* also add an extra component to $p_D(t)$ based on the frequency with which term t occurs in the whole collection, which has the effect of smoothing out large fluctuations between documents, as follows:

$$p_D(t) = 0.6\frac{f_{t,D}}{f_D} + 0.4\frac{f_t}{F},$$

where

$f_{t,D}$	number of occurrences of term t in D;
f_D	number of terms in document D;
f_t	number of occurrences of t in the collection; and
F	total number of term occurrences in the collection.

Using Equation 2, however, does not allow that some documents are very good descriptions of the query–for example, containing all of the rare query words–while some may be very poor–containing only a single, very common term from the query. In order to boost the contributions of documents that closely match the query, and degrade the contribution of those that are not as closely related but still contain at least one query term, Equation 2 is modified to

$$p_Q(t) = \sum_{D \in R} p_D(t) \times w\text{eight}(D). \qquad (3)$$

The weight of a document in relation to the query Q is determined as the product of $p_D(q)$ for all query terms q in Q. For a document that contains all of the query words, this weight will be high. If the document contains all query terms numerous times, then the weight will be even higher. For a document that

	Clarity	Original question	User query
TREC8			
Highest	1.100	How much sugar does Cuba export and which countries import it?	`import cuban sugar0`
Lowest	0.248	Do any countries other than the U.S. and China have a declining birth rate?	`birth rate statistics`
TREC9			
Highest	1.021	Which children's TV program was on the air longer the original Mickey Mouse Club or the original Howdy Doody Show?	`Mickey Mouse Club`
Lowest	0.257		`howdy doody show`

Table 1: Highest and lowest clarity scores for the two experiments.

contains only one of the query terms, most of the contributions to the product are 0.4 times the relative frequency of the term in the collection, with a small addition from the term itself, resulting in a low weight.

Putting it all together, we have:

$$\begin{aligned} p_C(t) &= \frac{f_t}{F}, \\ p_D(t) &= 0.6\frac{f_{t,D}}{f_D} + 0.4p_C(t), \\ p_Q(t) &= \sum_{D \in R}(p_D(t) \times \prod_{t \in Q} p_D(t)), \text{and} \\ \text{clarity}(Q) &= \sum_{\forall t} p_Q(t) \log_2 \frac{p_Q(t)}{p_C(t)}. \end{aligned}$$

3 Interactive TREC Experiments

This section provides an overview of our past 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 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 [Hersh and Over, 1999]. 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 [Hersh and Over, 2000].

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 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.

In order to find an improved system that would be predicted to perform well on the actual data used in the user experiments of Stage 2, the collection and topics chosen for this stage 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 can be employed in 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. Our user group was composed of thirteen medical librarians and twelve graduate students, mainly from the medical field. 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 randomised 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 [Witten et al., 1999] 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 validates our original predictions of which system should be chosen as the improved system in Stage 1.

3.1 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 [Hersh and Over, 1999]. Two examples are shown in Table 1; in this case each country 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.

Stage 1 of this experiment identified the Okapi weighting scheme [Robertson and Walker, 1994] as the improved system, 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, the Financial Times 1991-1994, using fourteen 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.

3.2 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. The new user task was question-answering, with two different types of questions. The first type of question required users to find a small number of answers for a topic; for example, the number of parks in the United States containing Redwood trees. The second type asked users to select the correct answer from two given; for example the third query in Table 1.

As there was no previous Interactive Track question-answering data to employ on Stage 1 of our three step 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 [Singhal et al., 1996]. This approach achieved over 65% improvement in MAP above the baseline on the Question Answering Track data.

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.

TREC-8	Batch MAP	User
Baseline system	0.2753	0.3230
Improved system	0.3239	0.3728
Change	+18%	+15%
TREC-9		
Baseline system	0.2696	66%
Improved system	0.3544	60%
Change	+32%	-6%

Table 2: Original results from the 1999 and 2000 TREC Interactive Track user experiments which are used for comparison with clarity scores in this paper. TREC-8 user performance is measured in instance recall, while in TREC-9, user performance is measured in % questions answered correctly. None of the changes reached statistical significance at the $p = 0.05$ level.

3.3 Summary

Table 2 summarises the results of these batch and 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. Users, however, performed equally well with both systems, with 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.

4 Methods

Each action by users in both the instance recall and question answering experiments were logged in a file, allowing us to go back and compute clarity scores for all queries. We included data from 24 users in the instance recall experiments, and 25 users from the question answering task. In both experiments each user issued about 3.5 queries on average for each question, so to get a single clarity score for a user-question pair we took either the maximum clarity score over the user's queries for that question, labeled MAX, or the mean, labeled MEAN.

The highest and lowest clarity scores for both the instance recall task (TREC8) and the question answering task (TREC9) are shown in Table 1.

In order to examine the relationship between instance recall of users–the proportion of possible instances they found for a question–and their clarity scores, a correlation coefficient was computed over all user-question pairs using the Pearson Product Moment Correlation.

For the question answering task, the mean clarity score was calculated for each question over all user-question pairs where the user had answered the question correctly. Similarly, the clarity scores were averaged for those user-question pairs where the user had not answered the question correctly. A paired t-test was then used to compare these six means. Note that

Question	MEAN		MAX	
	ρ	p	ρ	p
408i	0.06	0.79	-0.18	0.41
414i	-0.27	0.20	-0.40	0.06
428i	0.15	0.49	0.07	0.76
431i	0.18	0.39	0.13	0.55
438i	0.04	0.84	-0.20	0.35
446i	0.13	0.55	0.37	0.08
All	0.04	0.65	-0.05	0.56

Table 3: Correlation coefficients (ρ)for clarity scores and instance recall values taken over all users using both the MEAN and MAX methods for the TREC8 experiment. In each case $n = 24$.

two of the questions were excluded from this analysis as no user answered them correctly.

To determine if user's clarity scores improved over the course of the experiments for each question, the difference between the first query and subsequent queries was taken and then averaged across all user-question pairs. For example, if a user answering a question issued five queries with clarity scores 1, 2, 3, 4 and 5 respectively, then four differences were generated: $2-1=1$, $3-1=2$, $4-1=3$, and $5-1=4$. These differences were then averaged over all user-question pairs. This method is dubbed FIRST. The difference between subsequent queries was also examined, called the LAST method. Using the same example, the differences for this method would be would be $2-1=1$, $3-2=1$, $4-3=1$ and $5-4=1$. These differences were also grouped by system, either Cosine or Okapi, for both experiments, and by user type for TREC8 (librarian or postgraduate student).

5 Results

Table 3 shows correlation coefficients between the clarity scores of user queries and their instance recall performance. As can be seen there is no significant correlation between the clarity score for a user's query and their ability to find instances relevant to the question. Perhaps for question 414i using the MAX method there would be a significant negative correlation if the sample size was increased.

Tables 4 and 5 show the clarity score taken over all users that answered each question correctly, or not. A paired t-test between the two mean columns indicates that there is no significant difference between the clarity of queries issued by either group for both methods of calculating clarity. That is, clarity scores for queries by users who answered a question correctly were the same as the clarity scores for those users who did not answer the question correctly.

Now we shall turn our attention to quantifying the degree with which users improved the clarity of their queries over the course of the experiment. Figure 1 shows a plot of the mean difference in clarity for each query using the FIRST method for differencing separated by system. For example, the value shown for "q4" is the mean of all differences between the clarity value of the fourth query and the first query issued by any user of that system. Hence a value above zero indicates that the query has a higher clarity score than the first query issued for that question, and a negative value shows that the query has a lower score than the first.

For the TREC8 experiment, it seems that users of the Okapi system were able to improve their queries so that the fourth query had significantly higher clarity than the first. Users of the Cosine system, however, faired little better with subsequent queries

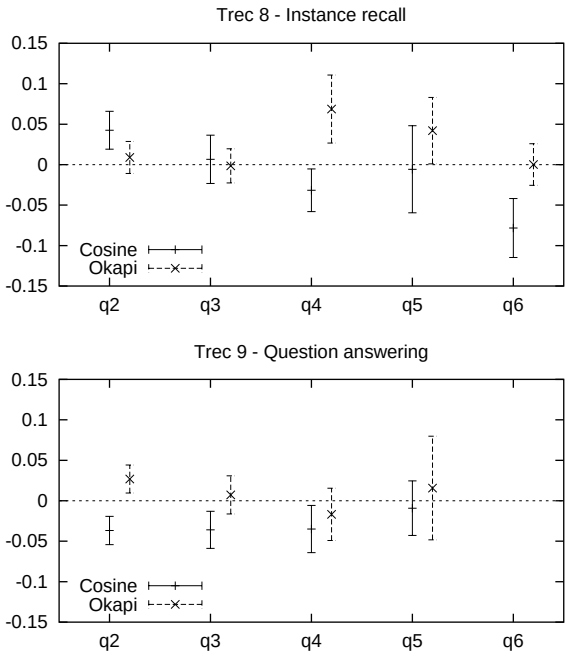

Figure 1: Mean difference between subsequent user queries using the FIRST method, where qx is query number x. Error bars are one standard deviation. For all cases $n > 10$.

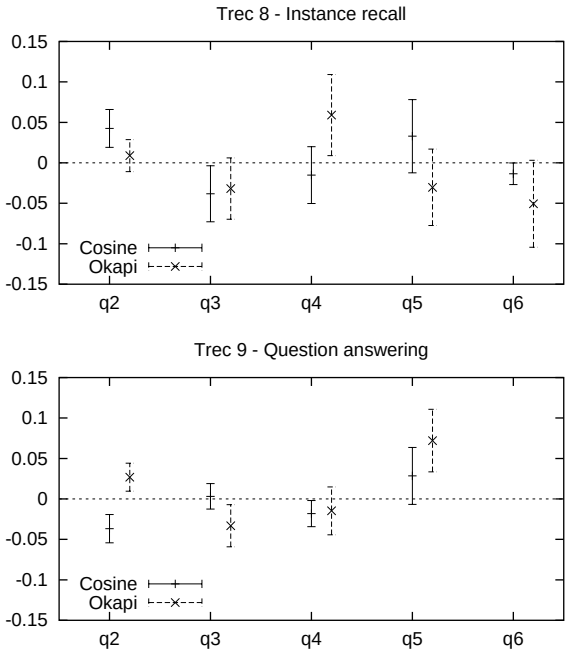

Figure 2: Mean difference between subsequent user queries using the LAST method, where qx is query number x. Error bars are one standard deviation. For all cases $n > 10$.

			Clarity			
			Correct		Incorrect	
Question	Correct	Incorrect	Mean	Stdev	Mean	Stdev
1	4	21	0.339	0.071	0.393	0.085
2	5	20	0.363	0.076	0.336	0.035
3	0	25	—	—	0.282	0.044
4	15	10	0.344	0.057	0.339	0.075
5	20	5	0.368	0.099	0.317	0.057
6	19	6	0.387	0.126	0.352	0.052
7	21	4	0.411	0.086	0.444	0.114
8	0	25	—	—	0.362	0.067
		Mean	0.369		0.353	

Table 4: Mean clarity scores using the MEAN method for each user-question pair taken over all users who answered the question correctly or not. For each question $n = 25$. A paired t-test reveals no statistical difference between clarity in the correct or incorrect columns ($p = 0.210$).

			Clarity			
			Correct		Incorrect	
Question	Correct	Incorrect	Mean	Stdev	Mean	Stdev
1	4	21	0.395	0.117	0.499	0.171
2	5	20	0.429	0.149	0.426	0.147
3	0	25	—	—	0.327	0.181
4	15	10	0.365	0.076	0.350	0.087
5	20	5	0.495	0.212	0.388	0.162
6	19	6	0.413	0.140	0.366	0.060
7	21	4	0.504	0.145	0.553	0.144
8	0	25	—	—	0.493	0.138
		Mean	0.433		0.430	

Table 5: Mean clarity scores using the MAX method for each user-question pair taken over all users who answered the question correctly or not. For each question $n = 25$. A paired t-test reveals no statistical difference between clarity in the correct or incorrect columns ($p = 0.918$).

over their first. Indeed as the number of queries increased, the clarity of their queries decreased significantly. Given that clarity is correlated with precision [Cronen-Townsend et al., 2002], it seems that users of the Cosine system would have to wade through more and more irrelevant documents with each query they issued. Surprisingly they could still find as many instances relevant to the question as their Okapi counterparts in the time limit of 20 minutes [Hersh et al., 2000].

In the TREC9 experiment there was little to separate the systems, with neither the Okapi or Cosine users able to do better than their first query.

Figure 2 shows the same figures calculated using the LAST method. This demonstrates more clearly than Figure 1 trends in user's query formulation behaviour. On this figure, a query is an improvement over the last if the error bars around a positive mean do not include zero. Each query is not significantly better than the previous, except for the fifth query in TREC9 using the Okapi system.

Figure 3 shows the differences in subsequent query clarity scores separated by user type. In the TREC8 experiment (instance recall) there were two user groups: postgraduate students and librarians. From the figure it seems that neither group outperformed the other in refining their queries to match the collection using clarity score as the metric. Both groups improved upon their first query with the second ("q2"). The third query was poorer, on average, than the first for librarians, while postgraduate students had no significant change from the second. The fourth and fifth queries issued were similar to the first (error bars include zero) for both groups, and it seems the sixth was worse than the first for both groups.

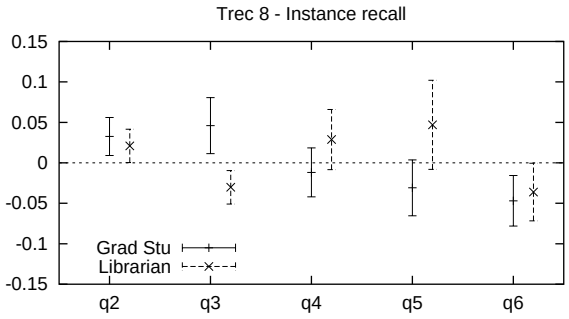

Figure 3: Mean difference between subsequent user queries in the TREC8 experiment separated by user type using the FIRST (top) and LAST (bottom) method, where qx is query number x. Error bars are one standard deviation. For all cases $n > 10$.

6 Discussion

There is no correlation between user performance and clarity score for the two tasks we examined. This is unsurprising, as a strong correlation between MAP and clarity scores has been reported [Cronen-Townsend et al., 2002], and there is no correlation between MAP and user performance in our data [Turpin and Hersh, 2001, Turpin and Hersh, 2002].

Perhaps more interesting is the analysis of the change in clarity of user's queries over time. It seems that generally the second query issued by a user had a higher clarity score than the first, but after that further queries did not improve in clarity score. This is despite the users issuing on average 3.5 queries per question, and having time to read through docu-

ments returned from the search engines to get a feel for the language of the collection. There is a hint in the TREC8 experiment that users of the Okapi system, that returned more relevant documents higher in the ranked list of documents, may have been able to modify their queries more effectively than users of the Cosine system.

There was no evidence to suggest that experienced searchers (librarians) were better at improving their queries with repeated searches than less experienced searchers (postgraduate students).

Another observation that arises from this work is that all of the clarity scores are fairly low, the maximum being 1.1. Cronen-Townsend *et al.* report clarity scores of 3.0 and above. Perhaps the questions asked in these two experiments were hard to answer, relative to the collections employed.

7 Acknowledgments

Thanks to Hugh Williams for discussions on the implementation of clarity and computing resources. Thanks to the anonymous reviewers for their helpful comments.

References

[Beaulieu et al., 2002] Beaulieu, M., Baeza-Yates, R., Myaeng, S., and Järvelin, K., editors (2002). *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Tampere, Finland. ACM Press, NY.

[Cronen-Townsend et al., 2002] Cronen-Townsend, S., Zhou, Y., and Croft, W. B. (2002). Predicting query performance. In [Beaulieu et al., 2002], pages 299–306.

[Hersh and Over, 1999] Hersh, W. and Over, P. (1999). Trec-8 interactive track report. In [Voorhees and Harman, 1999], pages 57–64.

[Hersh and Over, 2000] Hersh, W. and Over, P. (2000). Trec-9 interactive track report. In [Voorhees and Harman, 2000], pages 41–49.

[Hersh et al., 2000] Hersh, W., Turpin, A., Price, S., Chan, B., Kraemer, D., Sacherek, L., and Olson, D. (2000). Do batch and user evaluations give the same results. In *Proc. 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 17–24, Athens Greece. ACM.

[Robertson and Walker, 1994] Robertson, S. and Walker, S. (1994). Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 232–241, Dublin, Ireland. ACM Press, NY.

[Salton and McGill, 1983] Salton, G. and McGill, M. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

[Singhal et al., 1996] Singhal, A., Buckley, C., and Mitra, M. (1996). Pivoted document length normalization. In Frei, H.-P., Harman, D., Schäuble, P., and Wilkinson, R., editors, *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 21–29, Zurich, Switzerland. ACM Press, NY.

[Turpin and Hersh, 2001] Turpin, A. and Hersh, W. (2001). Why batch and user evaluations do not give the same results. In Croft, W., Harper, D., Kraft, D., and Zobel, J., editors, *Proc. 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 225–231, New Orleans, LA. ACM.

[Turpin and Hersh, 2002] Turpin, A. and Hersh, W. (2002). User interface effects in past batch versus user experiments. In [Beaulieu et al., 2002], pages 431–432.

[Voorhees and Harman, 1996] Voorhees, E. M. and Harman, D. K., editors (1996). Gaithersburg, MD. NIST Special Publication 500-238.

[Voorhees and Harman, 1999] Voorhees, E. M. and Harman, D. K., editors (1999). *Proceedings of the Eigth Text REtrieval Conference (TREC-8)*, Gaithersburg, MD. NIST Special Publication 500-246.

[Voorhees and Harman, 2000] Voorhees, E. M. and Harman, D. K., editors (2000). Gaithersburg, MD. NIST Special Publication 500-249.

[Witten et al., 1999] Witten, I. H., Moffat, A., and Bell, T. C. (1999). *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann Publishing, San Francisco, second edition.