# Sheffield Interactive Experiment at TREC-9

M. Beaulieu, H. Fowkes and H. Joho
*Department of Information Studies*
*University of Sheffield, U.K.*
m.beaulieu@sheffield.ac.uk

## Abstract

The paper reports on the experiment conducted by the University of Sheffield in the Interactive Track of TREC-9 based on the Okapi probabilistic ranking system. A failure analysis of results was undertaken to correlate search outcomes with query characteristics. A detailed comparison of Sheffield results with the aggregate for the track reveals that the time element, topic type, and searcher characteristics and behaviour are interdependent success factors. An analysis of the ranking of documents retrieved by the Okapi system and deemed relevant by the assessors also revealed that more than 50% appeared in the top 10 and 80% in the top 30. However the searchers did not necessarily view these and over half of the items deemed relevant by the assessors and examined by the searchers were actually rejected.

## 1. Introduction

The experiment for TREC-9 as in previous rounds in which Sheffield has participated was based on the Okapi system. Although the experimental design included two versions of the system, one with relevance feedback and one without, it was envisaged that the five minute time limit for searching each of the interactive queries for Trec-9 would offer little opportunity for searchers to use the feedback facility for query reformulation. Our aim was thus to focus on the characteristics of the two types of queries introduced for the TREC-9 interactive task and assess their relative impact on the performance of both the searchers and the system.

The graphical user interface of the Okapi systems remained exactly the same as in the last three rounds of TREC and includes:

- a query box
- a working query window containing system generated candidate terms for query expansion
- a scrollable window displaying a ranked list of the top fifty retrieved items
- a window for collecting documents marked as relevant and saved by the searcher
- a separate overlapping window for viewing items selected from the hitlist where searchers have to make a relevance judgement.

The standard questionnaires for the interactive track were used for data collection including: session entry, pre-search, post search, post-system and session exit. In addition transaction logs and talk aloud protocols provided system data and user perceptions in the course of the search. However subjects were not very forthcoming in talking aloud due to

the time constraint imposed on them and consequently the protocols provided very limited insight.

Sixteen searchers participated in the experiment and fourteen were Masters students in the Information Studies Department. None had used the Okapi system before, although most had some knowledge of ranking systems either through their general use of search engines on the Web or through a course on information retrieval in their programme of study. Half had between two to three year's online experience, with searching the Web and library catalogues being the most common types of systems. The other half was deemed to be novice users with a year or less experience.

## 2. Results and query characteristics

The TREC-9 interactive task included two types of topics. For the first set 901-904, searchers had to find a given number of different answers to a question, e.g. *three* national parks, *a* Roman site in France, *four* Orson Welles films, and *three* countries importing Cuban sugar. In essence these topics were not dissimilar from those used in the Interactive Track for TREC-7 and TREC-8, where searchers had to find as many different instances or answers as possible. The main difference in TREC-9 was that searchers had only five minutes to complete the task as opposed to twenty minutes in the previous rounds.

The second set of queries 905-908 required a single correct answer between two possible choices, e.g. the *longest* running TV programme, the painting completed *first*, the *last* Chinese dynasty, the country with the *larger* population. In arriving at a correct answer searchers had to find appropriate supporting evidence in different documents and save those documents.

The results of the Sheffield respondents compared to the aggregate performance of the participants in the track are presented in Tables 1a, 1b. The following will discuss these results in relation to the characteristics of each of the eight individual topics.

Table 1a: Sheffield results compared to the aggregate for Type 1 topics, 901-904

| Response / Topic number | 901 | | 902 | | 903 | | 904 | |
|---|---|---|---|---|---|---|---|---|
| All answers are supplied and supported *(2,2)* | Shef<br>- | Agg<br>8 *(7%)* | Shef<br>4 *(25%)* | Agg<br>18 *(18%)* | Shef<br>- | Agg<br>3 *(3%)* | Shef<br>1` *(6%)* | Agg<br>29 *(27%)* |
| All answers are supplied and some supported *(2,1)* | - | - | - | - | - | 2 *(2%)* | *1* *(6%)* | 7 *(7%)* |
| All answers are supplied and none are supported *(2,0)* | - | 3 *(3%)* | - | - | - | 1 *(1%)* | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Some answers are supplied, and all are supported *(1,2)* | 1 *(6%)* | 38 *(35%)* | - | - | 8 *(50%)* | 44 *(41%)* | 7 *(44%)* | 35 *(33%)* |
| Some answers are supplied and some supported *(1,1)* | 1 *(6%)* | 2 *(2%)* | - | - | 3 *(19%)* | 23 *(22%)* | 2 *(13%)* | 14 *(13%)* |
| Some answers are supplied and none are supported *(1,0)* | 1 *(6%)* | 6 *(6%)* | - | - | 3 *(19%)* | 27 *(25%)* | 1 *(6%)* | 8 *(8%)* |
| No answers are supplied and none are supported *(0,0)* | 13 *(82%)* | 50 *(47%)* | 12 *(75%)* | 80 *(82%)* | *2 (12%)* | 6 *(6%)* | 4 *(25%)* | 13 *(12%)* |
| Total number of searchers | 16 | 107 | 16 | 98 | 16 | 106 | 16 | 106 |

Table 1b: Sheffield results compared to the aggregate for type 2 topics 905-908.

| Response / Topic number | 905 | | 906 | | 907 | | 908 | |
|---|---|---|---|---|---|---|---|---|
| All answers are supplied and supported *(2,2)* | Shef *8 (50%)* | Agg 65 *(61%)* | Shef 8 *(50%)* | Agg 41 *(41%)* | Shef 9 *(56%)* | Agg 77 *(74%)* | Shef - | Agg 9 *(25%)* |
| All answers are supplied and none are supported *(2,0)* | 3 *(19%)* | 9 *(9%)* | *4 (25%)* | 32 *(22%)* | 7 *(44%)* | 15 *(14%)* | 5 *(31%)* | - |
| No answers are supplied and none are supported *(0,0)* | 5 *(31%)* | *32 (30%)* | 4 *(25%)* | *37 (37%)* | - | 13 *(12%)* | 11 *(69%)* | 78 *(75%)* |
| Total number of searchers | 16 | 106 | 16 | 100 | 16 | 105 | 16 | 101 |

*901: What are the names of the three US national parks where one can find redwoods?*

Sheffield respondents performed poorly on this query compared to the aggregate with 13 out of 16 finding no correct answers and the remaining 3 providing only partial answers, one of which was unsupported. The nil answers, which were twice as high as the aggregate, appear to have been influenced by some ambiguity in differentiating the meaning between "national" and "state" parks. The question may have presented some cultural bias, as a high proportion of our searchers were international students with no previous knowledge of the topic as indicated in the pre-search questionnaire.

*902: Identify a site with Roman ruins in present day France?*

The Sheffield results are comparable to the rest of the track with a quarter successful answers and three quarters of searchers unable to find a correct answer. The polarised

results may be due to the combination of evidence required to arrive at an answer, i.e. the name of the country, the specific location as well as the type of ruin. Furthermore only 7 documents were identified by the TREC assessors as providing an answer in the retrieved pool, a small number compared to other topics (See Table 3).

*903: Name four films in which Orson Welles appeared*

Once again our searchers produced comparable results with 13 out of 16 producing partial answers, but only half provided partial supporting evidence. This question was somewhat of a trick question in that most references referred to films directed by Welles and it would appear that there was some amount of guesswork in identifying films in which he was also an actor. The one Sheffield searcher, who got all the correct answers with supporting documents, had a special interest in film studies and was confident about the answer prior to searching the system. Three other searchers had also indicated pre-knowledge on this topic with a high degree of confidence.

*904: Name three countries that imported Cuban sugar during the period of time covered by the document collection*

Only one Sheffield searcher identified three countries compared with over a quarter of the aggregate. In addition twice as many Sheffield searchers did not succeed in finding any answers at all, 25% compared to 12%. This topic was also undertaken in TREC-8. The performance then was equally poor even though searchers had twenty minutes to search. It was found that although searchers were essentially looking for labels, i.e. names of countries, they had to engage with the content of the document to ensure that the correct context was covered. Although the time limit may have been a factor in TREC-9, it obviously doesn't account for the poorer performance compared to other participants in the track.

*905: Which children's TV program was on air longer: the original Mickey Mouse Club or the original Howdy/Doody show?*

Comparable results were obtained with the overall track with half of the searchers choosing the correct answer with supporting evidence. However a third of all searchers in the track provided no answer at all. As in question 902 on Roman ruins in France, few relevant documents provided the answer (See Table 3). In fact the searchers commonly saved two documents, one was deemed by the assessors to support the answer whereas the other didn't.

*906: Which painting did Edward Munch complete first: Vampire or Puberty?*

Sheffield performed slightly better than the aggregate with 50% getting the right answer with the correct supporting documents and 25% not finding the answer. Surprisingly 25% provided the right answer with no correct supporting evidence. Since only three

documents were judged to be relevant by the assessors (see Table 3), it would appear that searchers were able to make correct deductions or an informed guess.

*907: Which was the dynasty of China: Qing or Ming?*

Sheffield searchers outperformed the aggregate on this query with all identifying the correct answer, although just under half did not back it up with correct documents. Five of our searchers indicated that they knew the answer before searching, which may in part account for this discrepancy.

*908: Is Denmark larger or smaller in population than Norway?*

Just under a third of Sheffield searchers got the right answer but without supporting evidence compared with a quarter in the overall track who did provide correct supporting evidence. However in both cases around three quarters failed to find the answer all together. Again the high failure rate could have been related to the need to piece together different evidence over multiple documents in a short space of time.

Table 2 presents a summary of the adjusted score obtained for each answer which was correctly identified and supported by an appropriate document. The difference in the level of performance between the two different types of topics 901- 904 and 905-908 are clearly demarcated and reflect the overall pattern of performance in the track. It may be that the time limit was a critical success factor whereby it was more difficult to find correct multiple answers in the first type of topic and easier to find single answers in the second type.

Table 2: Sheffield adjusted score for correct supported answers for each topic out of the maximum obtainable score.

| Topic no | 901 | 902 | 903 | 904 | 905 | 906 | 907 | 908 |
|---|---|---|---|---|---|---|---|---|
| Adjusted score | 3 out of *48* *(6%)* | 4 out of *16* *(25%)* | 19 out of *64* *(30%)* | 23 out of *48* *(48%)* | 8 out of *16* *(50%)* | 8 out of *16* *(50%)* | 9 out of *16* *(56%)* | 0 out of *16* *(0%)* |

## 3. Searcher performance vs system performance

In an attempt to isolate user effect on system performance, the session logs were analysed to ascertain what proportion of relevant documents identified by the assessors were actually retrieved by the system. Table 3 compares the number of documents judged as relevant by the assessors for each topic and the average retrieved by the system in the

initial ranked hitlist of the top 50 documents retrieved for all of the searches. It would appear that poor searcher performance reported in Table 2 for topics 901 and 908 are not really borne out it terms of the average number of actual relevant documents retrieved by the system in the retrieved sets of 50 documents displayed to the searcher. Three quarters or more unique assessed relevant documents are retrieved by the system in all but one topic (908) in the first iteration which provides some evidence of the system's high level of performance.

Table 3 Assessed relevant documents retrieved in the top 50.

| Topic no | Total number of unique assessed relevant docs out of the possible maximum | Average no of assessed relevant docs retrieved |
|---|---|---|
| 901 | 10/13 *(77%)* | 5.6 *(43%)* |
| 902 | 6/ 7 *(86%)* | 1.87 *(27%)* |
| 903 | 13/17 *(76%)* | 5.6 *(33%)* |
| 904 | 29/39 (74%) | 11 *(44%)* |
| 905 | 7/7 *(100%)* | 3.2 *(46%)* |
| 906 | 3/3 *(100%)* | 2.4 *(80%)* |
| 907 | 20/23 *(87%)* | 7 *(30%)* |
| 908 | 9/15 *(60%)* | 2.6 *(17%)* |

Tables 4a, 4b compare the total number of assessed relevant documents examined by searchers for each of the topics with the number actually saved or deemed relevant by the searchers and those which were not deemed to be relevant. Overall 53% of documents deemed relevant by the assessors were examined but actually rejected by searchers. There were more documents rejected for type 1 topics than for type 2, 46% compared to 39%.

Table *4a:* Assessed relevant documents viewed and saved in the top 50, Type 1 Topics 901-904

| Topic no | No of relevant docs viewed | No of relevant docs saved | No of relevant docs not saved |
|---|---|---|---|
| 901 | 29 | 13 *(45%)* | 16 *(55%)* |
| 902 | 12 | 4 *(33%)* | 8 *(67%)* |
| 903 | 15 | 11 *(73%)* | 4 *(27%)* |
| 904 | 24 | 10 *(42%)* | 14 *(58%)* |

| | | | |
|---|---|---|---|
| Total | 70 | 38 *(54%)* | 42 *(46%)* |

Table *4b*: Assessed relevant documents viewed and saved in the top 50, Type 2 Topics 905-908

| Topic no | No of documents viewed | No of documents saved | No of documents Not saved |
|---|---|---|---|
| 905 | 25 | 14 *(56%)* | 11 *(44%)* |
| 906 | 18 | 9 *(50%)* | 9 *(50%)* |
| 907 | 9 | *8 (89%)* | 1 *(11%)* |
| 908 | 2 | 2 *(100%)* | - |
| Total | 54 | 33 *(61%)* | 21 (39%) |
| Overall Total | 134 | 71 *(47%)* | 63 *(53%)* |

Table 5 presents the ranking position of all the assessed relevant documents retrieved by the system but not necessarily viewed by the searchers. More than half appeared in the top 10 of the hitlist displayed to the searchers and 80% in the top 30.

Table 5: Assessed relevant document ranking for all searches for each topic.

| Topic | Top 10 | Top 20 | Top 30 | Top 40 | Top 50 |
|---|---|---|---|---|---|
| 901 | 55 *(65%)* | 8 *(9%)* | 3 *(4%)* | 1 *(1%)* | 18 *(21%)* |
| 902 | 19 *(76%)* | 4 *(16%)* | 2 *(8%)* | - | - |
| 903 | 37*(44%)* | 15 *(18%)* | 13 *(15%)* | 10 *(12%)* | 9 *(11%)* |
| 904 | 53 *(36%)* | 30 *(20%)* | 29 *(20%)* | 17 *11(%)* | 20 *(13%)* |
| 905 | 38 *(70%)* | 7 *(13%)* | 5 *(9%)* | 3 *(6%)* | 1 *(2%)* |
| 906 | 31 (97%) | 1(3%) | - | - | - |
| 907 | 36 *(43%)* | 14 *(17%)* | 13 *(15%)* | 10 *(12%)* | 11 *(13%)* |
| 908 | 19 *(59%)* | - | 7 (22%) | 3 *(9.5%)* | 3 *(9.5%)* |
| Totals | 288 *(53%)* | 79 *(14%)* | 72 (*13%)* | 44 (*9%)* | 62 *(11%)* |

## 4. Discussion of success factors

Although more failure analysis can be carried out on the data, a number of interdependent success factors appear to contribute to the above results including: time, topic type, and searcher characteristics and behaviour. Firstly there appeared to be some degree of correlation between topic type and the amount of time available for searching. 28% of searchers indicated that they didn't have enough time to undertake type 1 topics and 14% deemed they had enough time. In the case of type 2 topics there was little difference between those who felt they had enough or not enough time, 20% as opposed to18%. Searchers' perceptions regarding the time available also related to their level of satisfaction with the search outcome. There appeared to be a higher degree of confidence in the outcome of type 2 topics.

The discrepancy between the overall track performance in the two types of topics is not easily reconciled with our findings in the analysis of the degree of complexity of the TREC-8 topics and searching behaviour (1). In TREC-8 we found that in order to arrive at a relevance judgement, more complex topics required some interpretation on the part of the searcher and a higher degree of engagement with the contents of documents being examined. Less complex topics on the other hand were more easily understood at the outset and by enlarge relevant documents were identified by scanning for highlighted query terms in the documents. Hence it could be said that in general complex topics are likely to require more effort from the searcher than less complex ones. In the current round the differences in the level of engagement with the documents was not easily discernible in the time allowed for each search. However it would seem that the number of different answers required for type 1 topics was more demanding than the single answer required for type 2. The short time element may have been a more important success factor here than the complexity of the topic.

A third element, which contributed to the success/failure of the search outcomes, relates to the behaviour and the characteristics of the searchers themselves. Although type 2 topics required searchers to engage with the documents viewed to accumulate evidence for the correct answer, there was a substantial number of correct answers from Sheffield searchers which were not supported by appropriate documents 30% (Table 1b). The reason could be two-fold: firstly informed guesses could be made on partial evidence and secondly it may have been difficult for searchers to ascertain which document provided the correct evidence. The number of assessed relevant documents, which were viewed and rejected by searchers, would support this element of uncertainty, i.e. the difficulty in identifying the correct evidence and knowing which documents to save. In comparing the system performance with regard to actual assessed relevant documents retrieved by the system and the actual search outcomes for the topics, it is clear that search outcomes are highly dependent on the searchers themselves. Searchers either fail to examine relevant documents, or disagree with the assessors' judgements.

## 5. Conclusions

Since the Interactive Track was first established, much effort has been put into defining an appropriate search task. Although there is evidence to show that a realistic and reliable experimental setting can be created through simulated tasks (2), the search task for the current round of the Interactive Track was not ideal. In particular whilst it may be a common and realistic scenario for a searcher to want to find an answer as quickly as possible, the five minute time constraint in an experimental setting had an adverse effect. The participants in the experiment not only had to find the correct answer(s) but also had to provide the correct evidence, i.e. identify the documents which provided the right answer. Providing the evidence proved to be difficult and led to second guessing. With hindsight it may have been better for searchers to have had more time to engage with the documents to avoid readily rejecting items which did in fact contain the supporting evidence.

In addition to highlighting the limitations of the task, the current experiment also demonstrated the importance of comparing both user and systems performance in interactive searching. Although it is recognised a ranked output may not be the best way of presenting results (3), little research has been carried out to date on how searchers handle and interpret ranked output.

## References

1. Fowkes, H. & Beaulieu, M. Interactive searching behaviour: Okapi experiment for TREC-8. In: *Proceedings of the BCS-IRSG 22nd Annual Colloquium on Information Retrieval Research*, Cambridge, UK, 5th-7th April, 2000. 47-56.

2. Borlund, P. & Ingwersen, P. (1997) The Development of a Method for the Evaluation of Interactive Retrieval Systems In: *Journal of Documentation* 53 (3) 225-250.

3. Hearst, M. (1999) User Interfaces and Visualisation, In: *Modern Information Retrieval* (eds) Baeza-Yates, R. & Ribeiro-Neto, B., Addison-Wesley Longman Publishing Company.