Interactive Okapi at Sheffield - TREC-8

M. Beaulieu, H. Fowkes, N. Alemayehu and M. Sanderson

Department of Information Studies
University of Sheffield
m.beaulieu@sheffield.ac.uk

Abstract

The focus of the study was to examine searching behaviour in relation to the three experimental variables, i.e. searcher, system and topic characteristics. Twenty-four subjects searched the six test topics on two versions of the Okapi system, one with relevance feedback and one without. A combination of data collection methods was used including observations, verbal protocols, transaction logs, questionnaires and structured post-search interviews. Search analysis indicates that searching behaviour was largely dependent on topic characteristics. Two types of topics and associated search tasks were identified. Overall best match ranking led to high precision searches and those which included relevance feedback were marginally but not significantly better. The study raises methodological questions with regard to the specification of interactive searching tasks and topics.

1. Experimental objectives and setting

The University of Sheffield's participation in the Interactive Track is a continuation of the work initiated at the very outset of TREC at City University based on the Okapi system. With respect to the stated high level goal of the Interactive Track in TREC-8, which is to examine the process as well as the outcome, the Sheffield experiment focused principally on the process. The aim was to investigate interactive information seeking behaviour and user perceptions of the retrieval process using two versions of the highly interactive Okapi IR system, one with relevance feedback and one without relevance feedback. The specific objectives were threefold, each relating to the different experimental variables, i.e. searcher, system and task, as follows:

- to examine information seeking patterns of behaviour and determine how behaviour is shaped by the characteristic of the task and the functionality of the system;
- to determine how the different interactive searching features of the Okapi system namely, the best-match ranking, best-passage retrieval and incremental query expansion facility impacted on searching behaviour;
- to consider how searcher perceptions of the searching task are supported by the functionality of the interface.

The same configuration of the Okapi system was used as in TREC-6 and -7. A full description is found in (1). Searchers were subjected to two experimental conditions over the six topics. Each of the 24 searchers performed three searches on the system with relevance feedback and three on the system without relevance feedback, with 144 searches being carried out in total.

1.1 Data collection methods

In order to capture the multiple dimensions of the interactive searching process for qualitative analysis, i.e. searcher/topic, searcher system and topic/system interactions, other data collection methods were used in addition to the standard Interactive Track questionnaires. The test instruments included:

Observations: a structured approach was adopted to enable the experimenter to record the search process in four stages corresponding to the retrieval sub-tasks, i.e. search formulation and reformulation, viewing and evaluating results.

Transaction logs: the systems' extensive logging facility provided quantitative data on search interactions complementing the qualitative observational data.

Verbal protocols: searchers were instructed to 'think aloud' as they interacted with the system in order to get some insight into their perceptions, problems, strategies and understanding of the task in hand. The protocols were also used to gain a better understanding of any inconsistencies that emerged between the observational and interview data.

Questionnaires: four types of questionnaires common to all participants in the Interactive Track were administered by the experimenter. The pre-session questionnaire established searcher skills and experience. The post-search questionnaires ascertained the level of familiarity and ease/difficulty of the six individual topics. The post-system questionnaire gathered information on the ease of use and learnability of the two versions of the system. The final post-session questionnaire collected data on searcher preferences and views of the experimental conditions.

Interviews: following the standard post-search questionnaires, additional more probing questions were asked in order to gain more insight into searchers' perceptions of the individual topics and search tasks. A final post-session semi-structured interview provided further information on the system's interactive search features as well as the overall experimental session.

2. Searching behaviour

2.1 Query formulation

In over half of the searches, subjects formulated initial search queries by simply extracting keywords from the given topic descriptions. The single exception was for the Tropical Storms topic where two thirds of searchers also generated their own query terms. It appeared that there was some ambiguity with this topic. Some searchers interpreted it as searching for different types of storms, e.g. hurricanes, typhoons, as indicated in the topic description, whilst others were looking for actual named tropical storms.

Overall the norm was to enter between two and four single query terms which corresponded to the number of keywords in the actual topic descriptions (Table 1). The highest number of terms were entered for Tropical Storms and Tourism Violence, the reason being in part because more keywords appeared in the topic itself.

Table 1. Initial number of query terms entered

No query terms	1	2	3	4	5	6	7	Total no searches
Total no	7	35	50	26	14	9	3	144
searches	5%	24%	35%	18%	10%	6%	2%	100%

2.2 Query reformulation

Overall queries were reformulated for just over half of the searches carried out on both versions of the system. There was little incentive to modify a query if searchers were still finding instances of the required information in initial results, as for example for the topics on Birth Rates, Robot Technology and Tourism. Likewise, searchers were more likely to modify an initial query when they were finding few relevant documents. This was the case for Tropical Storms, Cuba Sugar and Tourism Violence, where a higher number of negative relevance judgements were made in relation to the total number of items viewed (Table 2, 3). Generally there was a strong correlation between the number of negative relevant judgements and the number of iterations in a search session.

Table 2. Number of negative relevance judgements

No. non relevant documents	0- 5	6-10	11-15	16-20	Total no. searchers
Tropical storms	3	10	7	4	24
Cuba sugar	2	5	12	5	24
Birth rates	10	10	3	1	24
Robots	20	4	-	-	24
Tourism	11	11	2	-	24
Tourism violence	8	10	4	2	24
Total no searches	54	50	28	12	144

Table 3. Number of positive relevance judgements and saved documents

No. saved documents	0-5	6-10	11-15	16-20	Total no. searchers
Tropical storms	4	18	2	-	24
Cuba sugar	18	6	-	-	24
Birth rates	-	10	14	-	24
Robots	2	14	8	-	24
Tourism	-	7	13	4	24
Tourism violence	11	9	4	-	24
Total no searches	35	64	41	4	144

With respect to the use of relevance feedback for query modification, there was no evidence to show that the availability of the relevance feedback facility encouraged searchers to reformulate queries compared to searches carried out on the system with no relevance feedback. Half of the searches undertaken on the system with relevance feedback were reformulated. Out of the 36 reformulated searches, in 17 cases the queries were expanded without any modification to the term list presented in the working query, which in effect can be considered as a form of automatic query expansion. For the remaining 19 searches, where users manipulated the candidate list of terms for query expansion, they were more likely to add and experiment with their own query terms than to experiment with those suggested by the system. This was particularly true for the three topics where it was difficult to identify relevant documents.

2.3 Viewing document hitlists

Three aspects of searching behaviour were examined in relation to how searchers viewed and selected documents from hitlists. Firstly, we consider the relationship between document rankings and items saved. Secondly, we compare how far down searchers worked through an original hitlist before modifying or stopping a search, and thirdly whether searchers worked through the hit lists in a comprehensive or selective manner.

The system displays the top fifty documents from each retrieved set. In 85% of searches more documents were viewed and saved from the top 25 items than from the bottom half of the ranked list. Moreover in 28% of searches documents were saved only from the top 25. Whilst this may imply agreement between the searcher and the system, some aspects of the experimental design may have created a bias towards more documents being saved from the top of the list. Firstly searchers were instructed to ignore documents which duplicated information already saved. Hence this resulted in the exclusion of relevant lower ranking documents. The design of the interface also made it difficult to substitute documents previously saved with an item found further down the ranking. The time limit would also have prevented searchers from exploring the full list.

How far searchers actually worked down the ranked list varied a great deal across the six topics. In half of the searches subjects went as far down as the 30th item in the ranked list. However for the Tropical Storms topic they were more likely to go to the bottom of the list. This is possibly explained by the number of duplicate material which occurred and could be skipped over. Duplicate material was also a feature of the

Tourism Violence topic, but by contrast over half of the searchers did not go beyond the 30th ranking. In this case the display of the term occurrence information provided an effective indicator of document relevance. Searchers could assume that if the keyword `Tourism' did not appear, the document was unlikely to be relevant. However for the Tropical Storms topic the display of query terms associated with an individual item was less helpful and it was difficult to judge promising documents.

In another topic, Robot Technology, two-thirds of searchers only examined items from the top 20 documents in the list. This was largely due to the time searchers took to view and engage with the documents. In some cases there were a number of `instances' to be considered in an individual document and more time was required to read the content and get to grips with the topic.

Searchers adopted two strategies for working through a hitlist. They either viewed documents sequentially or were more selective. For topics where there appeared to be a high level of potentially relevant items, e.g. Robot Technology, Tourism, Birth Rate, the tendency was to view each item in turn whereas for topics where there appeared to be fewer relevant items, e.g. Tropical Storms, Cuba Sugar, Tourism Violence, searchers were more likely to skip through and be highly selective.

2.4 Viewing full documents

2.4.1 Passage retrieval

On viewing a document the searcher is taken to a highlighted `best' passage which represents the section of the document which scores the highest in relation to query term occurrence. The function of the highlighted best passage is not only to assist users in determining relevance, but it also serves as a source for extracting terms for query expansion and the searcher is given the choice to make a relevance judgement on the full document or the passage only. An analysis of searching behaviour in relation to passage retrieval was undertaken for three of the topics where the feature came most into play, i.e. Birth Rates, Robot Technology and Tourism. Documents related to the other topics were generally much shorter, and the `best passage' was not an option.

Searchers inevitably started by scanning or reading the information in the highlighted passage. However in over half of the searches, subjects examined other parts of the document before making a relevance judgement based on the passage only. It appeared that searchers sought more contextual information and evidence outside the highlighted passage before making a relevance judgement. Alternatively the motivation was also to look for additional instances outside the highlighted passage. In the case of the Tourism topic searchers were more likely to make a judgement on the basis of the passage only. Searchers were in effect looking for the right combination of labels. If a sentence include the keywords `tourism' , `increase' and a number, the document was deemed to be relevant. However on the whole searchers were not confident in making relevance judgements on best passages only. The verbal protocols also revealed that the passage only option was perceived as a means of making a weak relevance judgement, i.e. an indication that the whole document was only partially relevant.

2.4.2 Scanning vs reading

Scanning was the most prevalent strategy for evaluating the contents of documents. This was in part dictated by the time limit for the search task and the need to find as many different instances as possible. Query term highlighting was essential for document viewing. However scanning and reading were usually carried out simultaneously. Inevitably searchers had to supplement scanning with reading in order to establish contextual relationships between query terms. For example the terms `Tourism' and `Violence' retrieved documents on the impact of political violence on tourism as well as those on violence targeted directly against tourists. Hence the level of engagement with the topic had a major influence on the procedural level of interaction. When labels were explicit (e.g. names, numbers) searchers would extract instances by plotting highlighted keywords. For topics such as finding the latest developments and applications of Robot

Technology, which was more cognitively demanding, scanning keywords did not provide enough clues. Other factors which contributed to the number of searchers who read as the primary activity for viewing documents included the density of the 'instances' found in the document as well as the lack of familiarity with the topic.

2.4.3 Interpreting topics

Overall searchers had little familiarity with the topics, Birth Rates being the most familiar and Cuba Sugar and Robot Technology being the least familiar. As searches were carried out subjects expressed different levels of certainty regarding the task in hand. Through the behavioural observations and the verbal protocols it was possible to identify different cognitive states and levels of engagement associated with the different topics. The descriptions below provide some insight into how searchers interacted conceptually with the topics on Tropical Storms, Cuba Sugar, and Robot Technology.

Tropical Storms

The different interpretations for this topic seemed to have an influence on the levels of certainty and patterns of searching behaviour. Searchers who interpreted the question as finding different types of storms ended up questioning and doubting their search goal as the search progressed. They made more negative relevance judgements than those searchers who understood the task to be looking for names of tropical storms.

The more common interpretation of the topic was to search for the names and locations of different types of tropical storms. This led to a different level of engagement with the topic as well as uncertainty. Some of the searcher-system interaction for this topic was very superficial involving pinpointing names of storms. In fact 6 out of the 24 subjects never read any of the documents at any point in the search. Others by contrast adopted a deeper level of engagement with the topic, treating the property damage/loss of life as a separate criteria that documents had to fulfil in order to consider a document as being relevant. This resulted in another level of questioning and doubting as searchers were unsure whether to save a document if the evidence of property damage/loss of life was only implied. A typical example of this was the dilemma over whether it was possible to infer property damage from references to `heavy landslides' or damage to `communication infrastructure'. A high number of searchers engaged either in trying to predict whether a document met the relevance criteria or in trying to avoid documents from the hitlist which were concerned with a specific storm already identified.

Robot Technology

For this topic most searchers expressed a high level of uncertainty about the content of the information found in the initial stages of the search. Although there were a few searchers who adopted a common search pattern of scanning for names of applications, one distinctive feature of the searching behaviour was that searchers attempted to comprehend and make sense of the documents. This involved a great deal of reflection and interpretation whereby information was explicitly extracted, compared, categorised and summarised. This type of conceptual activities and reasoning was largely absent or was not evident from the searching behaviour associated with the other topics. Some searchers even attempted to grapple with the conceptual complexities of defining the latest developments and differentiating between current and future developments.

Cuba Sugar Imports

In contrast with the other topics, there was generally a high level of certainty about what to look for during the initial search process. This however gave way to doubt and uncertainty in the viewing/evaluation stage owing to the few relevant documents found in the database. A significant number of searchers thus changed their searching strategy in the course of the search. Two different approaches were adopted. Some searchers reformulated their query which resulted in a more positive state, as the subsequent duplication of named countries provided confirmation that there were only a few countries involved. Others however came to a similar conclusion but the uncertainty was resolved by a deeper level of engagement with the topic. By reading and comprehending the actual text, searchers assimilated information about the broader economic context and the trade embargo and were able to confirm that there were only a few answers to the topic request.

3. User perceptions of search tasks

Users experienced little difficulty in starting searching namely because as mentioned previously they simply extracted keywords from topic descriptions. Three quarters of searches were deemed to be easy or somewhat easy. Tourism and Birth Rates were the easiest and Cuba Sugar, and Robot Technology were classed as the most difficult. It appears that an inherent feature of a difficult topic is the level of familiarity and the understanding of the content and context of the issues. For example the technological content of the Robot technology topic and the economic or geographical issues addressed in the Cuba Sugar topic made it more difficult for the searchers to absorb the text. Another indicator of a difficult topic was the higher level of engagement or effort required from searchers. The Robot Technology topic comprised two elements, to find the latest developments as well as applications. Whilst for the Cuba Sugar topic, the task involved finding hidden labels and there were many false leads before finding a correct ‘instance’.

Perceptions of search ease/difficulty are closely linked with search satisfaction. Hence for the topics perceived as more difficult (Cuba Sugar and Robot Technology) searchers were more reluctant to rate search satisfaction very highly because they were uncertain of the scope of the topic. Search satisfaction is also a product of the quality of what was found rather than how much is found. Searchers expressed the lowest level of satisfaction for the Robot Technology topic even though overall they found the highest number of instances related to it.

In all of the topics, excluding tropical storms, two-thirds of searchers were very or quite confident that they had identified all of the instances. In the case of Cuba Sugar and Tropical Storms they were more confident because there appeared to be few instances anyway. There is also a clear relationship between how many instances are identified and searcher perceptions of whether or not they had enough time to do an effective search. For topics where the most instances were found three quarters of searchers said that they didn’t have enough time. Conversely for topics where more duplicate material was found, searchers indicated that they did have enough time to carry out the search. The results also reinforce the relationship between searcher confidence and the amount of time needed. In topics where searchers were most confident that they had identified all the instances, they were also likely to claim that they had enough time.

4. User perception of the system

There was little difference in the perception of the ease of use or learnability between the two versions of the system. Both systems were deemed to be easy. In seems that the difficulties that searchers had in manipulating the working query did not colour the overall perceptions of ease of use or learnability. This is possibly because after the first or second attempts most searchers abandoned the working query and treated the experimental system in exactly the same way as the control system.

4.1 Relevance feedback and query expansion

Two-thirds of searchers professed a high level of understanding of both systems. However given that three-quarters did not perceive any differences between the two, this might suggest that they did not understand the underlying relevance feedback mechanism and did not readily link the terms of the working query to the retrieved set of documents. Two-thirds of searchers declared a preference for the control system without relevance feedback. The most frequently cited reason given was related to the manipulation of the working query. Searchers confirmed that removing suggested terms individually was a very time-consuming process which didn’t warrant the effort. However the desire for more control over the search process was a deeper concern. It appeared that searchers were happy for the system to provide suggestions as long as it presented the candidate terms in a way that didn’t threaten to change the direction of the search.

4.2 Best match ranking

Searchers expressed confidence in the ranking of the hitlist as an overall guide to potential relevant documents but they also recognised that the ranking was not a guarantee that items would be relevant. The query term occurrence information was also considered to be useful in providing clues to potential relevancy.

4.3 Passage retrieval

Passage retrieval was perceived as an efficient way of identifying instances and minimising user effort for extracting relevant information. However for topics less familiar to the searchers and where they needed to understand the context of the document, jumping to the highlighted passage was deemed to be procedurally disorientating and counter intuitive. This generated reservations about the suitability and reliability of best passage retrieval in terms of a starting point as well as giving the searcher some indication about how far into the document they were.

At a more conceptual level about two thirds of searchers were sceptical about the support that passage retrieval offered for making positive relevance judgements. In some ways passage retrieval added to the cognitive burden and uncertainty in making relevance judgements. The dilemma was clearly expressed by one searcher as follows:

"All of it is relevant, but then what I actually want to know is in the passage, but then I only know that because I've looked at it all so it's difficult to decide whether I want to say it's fully or partly relevant".

5. Search outcomes

In this section we present the search outcomes as perceived by the searchers as opposed to the actual results or system performance in terms of precision and recall measures.

Table 4 shows that the average number of instances found by subjects for all the searches across the different topics for each of the experimental conditions were comparable, 12.0 instances per search for the system with relevance feedback and 11.7 for the version without relevance feedback. Equally the average number of instances for the searches which were *not* reformulated under both conditions was similar 13.8 and 14.0 respectively. However for searches which were reformulated, the average number of instances decreased significantly to 10.2 instances for searches using query expansion based on relevance feedback and 9.9 for searches not using relevance feedback.

Clearly, the conditions which led to reformulation differed for the different topics. Reformulation was beneficial for three of the topics which required a some degree of interpretation, e.g. Tropical Storms, Cuba Sugar, Tourism and Violence, whereas it appeared less fruitful for the seemingly more straight forward topics where the answer to the request was clearer, e.g. declining Birth Rates, Robot Technology developments, and increases in Tourism. It also appears that query modification is more likely to improve searches with initial poorer results than those which are already successful.

Although search queries which were expanded by using relevance feedback retrieved marginally more instances than search queries which were reformulated by searchers generating their own additional query terms, queries which were in effect expanded automatically (i.e. searchers accepted all the candidate terms presented by the system in the working query) compared to those expanded interactively (i.e. searchers added their own query terms to the working query), led on average to more instances, 11.2 as opposed to 9.3 (Table 5). Automatic query expansion (AQE) also seemed to be more effective for the topics where instances could be more easily identified (Birth Rates, Robot Technology, Tourism), whereas interactive query expansion (IQE) appeared to be as more productive for more complex topics (Tropical Storms, Cuba Sugar).

Table 4. Instances retrieved with relevance feedback on and off

Condtition	Relevance feedback on			Relevance feedback off		
No. instances	All searches	Searches not reformulated	Reformulated searches	All searches	Searches not reformulated	Reformulated searches
Tropical storms	142	54	88	135	54	81
Cuba Sugar	77	26	51	108	29	79
Birth rates	173	97	76	156	91	65
Robots	221	165	56	196	137	59
Tourism	167	138	29	170	109	61
Tourism violence	87	20	67	83	29	54
Total no. instances	867	500	367	848	449	399
No. searches	72	36	36	72	32	40
Average no. instances	12.0	13.8	10.2	11.7	14.0	9.9

Table 5. Instances retrieved with automatic and interactive query expansion

Condition	AQE	IQE
Tropical storms	13	75
Cuba sugar	12	39
Birth rates	64	12
Robots	41	15
Tourism	17	12
Tourism violence	44	23
Total no. instances	191	176
No. searches	17	19
Average no instances	11.2	9.3

6. Search results

The comparative performance of searches undertaken with the system which included relevance feedback as opposed to those on the system without relevance feedback shows that the former led to marginally better precision but the difference is not significant (Table 6). An analysis by individual topics reveals the poorest overall performance for Tourism/Violence where searchers had difficulty in filtering out duplicates. Nevertheless, searches reformulated on the system with relevance feedback did lead to better results than those reformulated without relevance feedback. By contrast the Cuba sugar topic was the second worst in terms of precision but achieved the highest recall. It was perceived as difficult with many duplicates leading to negative relevance judgements and few instances were found. The best precision was achieved for the most straightforward topic on increase in Tourism. Although the Robot Technology topic required more effort and engagement, it led to the second best precision results. Both of these topics required little query reformulation.

Table 6. System Performance with and without relevance feedback

System condition	Precision	Recall
With RF	0.759	0.352
Without RF	0.728	0.393

7. Summary and conclusions

The experiment was primarily concerned with searching behaviour in the use of a highly interactive retrieval system and determining the interaction between the different variables, i.e. searcher, search task/topic and system characteristics, within the context of the TREC Interactive Track experimental framework.

7.1 Searcher characteristics

The twenty-four test subjects were quite a homogeneous group and there appeared to be no evidence of any significant individual differences in searcher characteristics in terms of their familiarity with the topics or the searching strategies they adopted. Moreover search results reveal no significant difference in search performance between individual searchers.

7.2 Topic and task characteristics

Searching behaviour was however largely determined by the nature of the different search topics themselves. Two types of topics and associated search task emerged. On the one hand some topics were clearly defined and instances could be easily identified from the outset. The search task for those topics involved scanning documents to spot keywords, e.g. names of countries, which made it possible for relevance judgements to be readily made. Other topics on the other hand were more complex, requiring some element of interpretation and they were also characterised by a degree of uncertainty. In such cases identifying instances was not simply a question of finding keywords, but relevance was more dependent on establishing the context in which the keywords appeared. Hence searchers had to read the documents, engage with the content and deliberate. Deliberation or uncertainty was largely concerned with defining the scope of the topic from the documentary evidence, e.g. whether or not a named storm was a tropical storm of not. Uncertainty also emerged when few instances could be found.

7.3 System characteristics

Overall the system best match ranking was effective in producing high precision searches rather than searches with high recall. Relevance feedback came into play in different ways depending on the type of topic. Automatic query expansion could improve results of simple straightforward topics whereas for more complex topics, interactive query expansion with contributions from both the searcher and the system appeared to be more effective. Displaying query term information in the hit list as well as highlighting best passages and query terms in documents, assisted users in making relevance judgements for the simple topics but they were less helpful on their own for the more complex topics where searchers had to engage with the content.

7.4 Experimental design and conditions

The results of the study raises methodological questions with regard to the specification of the interactive task and the topics. From a system's perspective it could be argued that the TREC interactive task of finding as many different instances on a topic as possible in twenty minutes is basically a recall task. However from a searcher's perspective it appears to be perceived as a precision task with the emphasis being on finding `new evidence' not just more evidence. Having the system discard duplication or documents which covered known evidence was the most frustrating element of the task. Thus more attention needs to given to consider other types of retrieval tasks which may be more appropriate for evaluating interactive searching.

The choice of topics also influenced the nature of the search task. Although the experiment included only six topics, which made it feasible to increase the number of test subjects, fruitful data was collected on the characteristics of topics. More research however is required not only in identifying different types of search topics, but also in defining more close what constitutes a simple and more complex topic and determining how the different elements should be taken into account in the experimental design. In order to deepen our understanding of interactive searching and its evaluation, a typology of search topics needs to be developed.

References

1. Beaulieu, M. M. and Gatford, M. J. Interactive Okapi at TREC-6. In: Voorhees, E. M.   and Harman, D. K. (eds). *The Sixth Text Retrieval Conference (TREC-6).* Gaithersburg, MD: NIST, 1998, 143- 167.