IBM Search UI Prototype Evaluation at the Interactive Track of TREC6

Birgit Schmidt-Wesche, Robert Mack, Christian Lenz Cesar
(bwesche @ us.ibm.com, maier @ watson.ibm.com, cesar @ watson.ibm.com)
IBM Thomas J. Watson Research Center, Hawthorne
and
David VanEsselstyn
Teachers College, Columbia University, New York

INTRODUCTION

Our search application used in the TREC6 Interactive Track was developed as part of a User-Centered Design (UCD) program aimed at prototyping UI approaches for using different search technologies being investigated at IBM Research. The search UI we used in the TREC6 Interactive Track focused specifically on four UI issues:

1. The value of web-like +/- query syntax, provided in the sense of the tacitly accepted web standard (with '+' meaning that a term must appear in result documents, and '-' meaning a term must not appear in the returned documents).
2. Support for query refinement using additional terms generated by a Context Thesaurus (CT) program (see Cooper and Byrd (1997)).
3. Display options such as the highlighting of query terms in a result document.
4. Ability to save relevant documents in a separate browser across queries.

The basic components of our search prototype were the IBM NetQuestion search engine (IBM, 1997), the CT component (both running on AIX servers), and a Java application UI (Win95 client) consisting of four tilable windows with the following functions:

1. Query Window with a text field for forming queries and a list of terms related to the user-defined query terms.
2. Result Window, with a list of document results, shown in order, a checkbox for selecting a document, asterisks indicating relative importance ("5-star" rating system), and the document title.
3. Document Window, showing document content, with query terms highlighted.
4. Saved Documents Window, listing documents users selected to save in the Result Window or the Document Window.

When a document is selected for viewing, the title in the Result Window changes from black to blue to indicate that the document has been viewed. Documents can be saved to a separate browser by clicking on a checkbox next to the title in the Result. The "related terms" panel in the Query Window actually showed two levels of related terms: one list displays all terms that co-occur with the query terms in the document collection, and a second window pop-up when the mouse moves over a first-level term. This popup window shows second-level terms related to the selected co-occuring term. The popup window terms change as users move the mouse pointer over co-occuring terms in the first-level list. Both sets of terms are prefixed by check boxes, which, when clicked, add the co-occuring or related terms to the query terms. Unchecking removes the terms from the query specification.

NetQuestion, a search engine especially designed for web search, combines Boolean with Free Text search functions. The '+'/'-' query syntax was designed to exploit these two underlying functions with

each query. What we refer to here as the Context Thesaurus component actually consists of a number of components, each focusing on the various degrees of relatedness between query terms. The type of relationships are determined by frequency, degree of vicinity, and specific syntactic structure in which any two given terms occur to each other in the documents of the collection to be searched. For a more detailed discussion, see Cooper and Byrd (1997).

TEST PARTICIPANTS

We hired five temporary employees from an employment agency used by IBM Research. Appendix 1 summarizes the characteristics of the four participants we used in the evaluation. We rejected one participant because the logging program for our experimental prototype failed, and we lost this person's data. Two participants were "under 40", and two "over 40", and we had one male and one female in each age group. While we had access to professional librarians from another source, we chose to use people from the temporary agency. We specified that these people should be computer literate, and that experience with the World-Wide Web (including Web search) was desirable. We were interested in how usable and useful our prototype UI was for ordinary computer literate people, and not professional librarians. Participants generally conformed to these requirements, although one participant had limited experience with a mouse-based interface, and searching on the Web.

EVALUATION METHOD

We followed the prescribed test procedure with two exceptions noted below. As per the prescribed method, we adhered to the following 10 steps:

1. The participants filled in the questionnaire in which they were asked about their educational background, and their experience with other online search systems and with mouse-based applications.
2. Experimenters read the test instructions to the participants. They could keep them for further reference.
3. Participants worked through the tutorial of the system whichever they had to deal with first (either the experimental system or the control system). The tutorial handout also was available for further reference during the experiment.
4. Experimenter read the TREC6 experiment instructions to the participants.
5. The participants started each task, first reading the topic descriptions and then searching for specific aspects as requested in the narrative and aspects description, writing down on paper the aspect they had in mind when they saved a document.
6. After each 20 min. task, the participants were handed the questionnaire in which they were asked to judge how easy it was to deal with this task.
7. After having completed the first three tasks we took a break.
8. After resuming, the participants worked through the tutorial of the second system, performed the searches, and filled out the questionnaires as before for the first system.
9. Following the three searches of our experimental system, participants completed a questionnaire in which they were asked to evaluate this system.
10. Finally, the last questionnaire asked the participants to compare the two systems in a couple of respects.

In addition, however, we deviated slightly from the overall method in two ways. First, our system hung up during several searches for various reasons. In all but one case, we were able to quickly restart the application, and allowed the subject an additional minute of search time at the end of the session. One

case was more severe. But since the hang-up occurred during the first three minutes of that search session, we stopped it, restored our search system, and resumed again, allowing only the remaining time of 17 minutes to complete that search. We did not reject the data from these participants, although, had we had more time to conduct the evaluation, we might have replaced these participants.

The second deviation is that we sat with participants and allowed them to ask us brief "help" questions which we also answered briefly. If participants looked puzzled or voiced confusion, we also asked them to describe their problem and then we helped them if we could. This kind of interaction is more common in explicit "formative" evaluations where experimenters are more interested in what participants are thinking about as they use a system, perhaps for purposes of developing a help system, or discovering system problems, than in gathering quantative data under strictly controlled conditions for benchmarking purposes (see Landauer, 1989).
We also video-taped the computer screen as participants worked, using a camera in the same room as the participants worked and the experimenters sat.

RESULTS

Here we briefly introduce the quantitative and qualitative results of the evaluation, and briefly introduce specific results requested by the TREC6 Interactive Track methodology, summarized in Appendices 2 through 6. We also contrast in somewhat more detail the search aspects ("topic keywords") our participants identified to those identified by NIST from the collection of submitted documents.

I. Recall and Precision: Quantitative Performance

NIST pooled data from each participating site, and carried out various statistical analyses over the pooled data. There is not enough data from individual participant site evaluations to analyze performance at this level. The experimental design, independent and dependent measures, and statistical analysis model are described in detail at the NIST Web site (general TREC home page at 'http://trec.nist.gov/'; TREC6 Interactive home page at 'http://www-nlpir.gov/~over/t6i/'), and in the overall proceedings of which this report is a part, and we do not repeat this here. Instead we summarize a few key design points and results, and draw conclusions with respect to our prototype performance.

The design is a Latin Square design, which means that the design is an incomplete orthogonal design: Search System is blocked with respect to Search Participant and Search Topic. The statistical analysis is based on derived dependent measures defined by the difference between a Participating Site Search System, and a control system provided by NIST (ZPRISE), for recall and precision. This difference in turn is defined by pairing performance on different Search Topics, that occur in the same order within a test block. Recall is actually *aspectual recall,* and is defined as the proportion of topic aspects identified by searchers, in relation to the total amount available in the corpus, as determined by NIST for a given topic. Precision is defined as the proportion of identified aspects to the relevant topic aspects, as determined by NIST. Again, the details are best understood by consulting the NIST Web site.

Essentially, the design and statistical analysis address the question: For each Participating Site, is there a significant difference between the Search System (for that site), and the NIST ZPRISE control system, with respect to precision and recall? Questions about the effect of Search Topics, and Site Participants of course can be addressed only at the level of pooled data over sites.

The analysis of variance (ANOVA) performed by NIST indicates that there are main effects of Participating Site, Search Participants, and Search Topics for the recall dependent variable. A

comparable analysis has not been carried out for precision. However, while there is an overall effect of Participating Site, most Search System differences by Participating Site are not statistically significant. For example, in our case (the IBM Participating Site), performance was better on the the ZPRISE control system than our prototype, for both recall and precision: Mean recall for the 12 Experimental-Control differences across 3 topic x 4 participant blocks = -0.114, and mean precision for the 12 Experimental-Control differences across 3 topic x 4 participant blocks = -0.019. However, the 95% confidence interval for both mean differences contains zero. And this was the case for all 10 Participating Site comparisons for precision, and 9 of 10 comparisons for recall. Nonetheless, we suspect that with more data points, the difference, for recall in any case, would attain significance. Hence, we conclude that our experimental system does not provide a more effective search UI than the ZPRISE control, and that we have improvements to make. The next section points to some details of query and browsing performance that may guide these improvements.

Two other quick points: First, there was a significant main effect of Search Topic. That is, search topics varied widely in difficulty: the "International Art Crime" topic (322) was hardest, the "New Hydroelectric Projects" topic (307) was easiest across Site Participants. Therefore, while the data are too noisy to statistically distinguish differences between Experimental and Control systems across sites, the differences between Search Topics are large enough to achieve statistical significance. With respect to the search systems and search topics evaluated here, differences attributable to Search Topic difficulty are a much more important factor in relative performance than differences in Search Systems. Finally, there was also a significant main effect of Search Participant. That is, people vary widely in their search performance (measured in terms of recall and precision). This is characteristic of behavioral studies. During the NIST Workshop on the Interactive Track, it was estimated by one site representative that more than four Search Participants would be needed (the minimum prescribed by the experimental design) to achieve enough statistical power for the Search System differences to achieve statistical significance, given the error variance in the data, and the experimental design. Consequently, future evaluations need to use more test participants, and/or use a more powerful experimental design.

II. Query and Browsing Performance

Appendices 5 and 6 summarize frequency of certain query and browsing activities involving our experimental system and the ZPRISE control system. Appendix 5 shows for each Search Participant, the number of queries submitted for each Search Topic, the number of times the Context Thesarus related terms function was used, average number of terms per query, number of documents viewed, aspects recorded, etc., along with aspectual recall outcome for each Search Topic. The tables in Appendix 6 summarize performance over Search Participant, and Search Topic.
The tables indicate that search participants using both the ZPRISE control and experimental systems generated multiple queries, and iteratively refined them by adding or deleting terms. In the case of our experimental system, all participants invoked related terms (generated from the Context Thesarus) at least once, although they tended not to modify query content using these related terms on subsequent search topics. Of course the ZPRISE control system did not provide related terms, or a "+/-" query language, so these activities are only represented for our control system.

The frequencies of activities that are comparable across systems vary across participants, but are actually comparable across our experimental and control systems, with one exception. We had some intense browsers (up to 105 result documents per topic), as well as quick scanners; some used only 2-3 query terms, while others entered up to 10 terms per query. With respect to aspectual recall, two of our searchers did better with the control system, while the other two did better with our experimental system.

Looking at the average recall per topic, we see that all four participants did relatively well on the final topic 339 ('Alzheimer's Drug Treatment'), compared to other topics.

The only striking difference between search systems and search participants can be observed with topic 303 ('Hubble Telescope Achievements'). Two participants using ZPRISE scored significantly better than the other two using our experimental system. It turned out that several "Hubble" documents in the NIST corpus are very short, and our experimental system ranked these more highly than longer documents which provided more aspects for the search topic. Compared to the ZPRISE result list, participants using our experimental system had relatively more difficulty identifying relevant aspects from the result list.

It is not yet clear how differences in these activity profiles account for the difference in recall and precision performance between systems at our site, or across sites. The "+/-" syntax and use of related terms involved extra activities and time not represented in the control system, that would be compensated for only if these activities ultimately resulted in better recall and precision for search topics. At this point, we cannot claim that either feature provided for more effective search in an overall sense, and we do not know whether the *functions per se* are problematic, or the *UI implementation*, or both.

Finally, the data pooled by NIST does not include information at the level of categories we summarize in Appendices 5 and 6, so we cannot compare this level of analysis to other site participants.

III. Summary of Data Requested by NIST

Appendix 1 summarizes the participants' characteristics, in particular their online search experience. Appendix 2 provides the users' judgements of the six topics of the experiment. Appendix 3 provides the questionnaire results for the experimental system, and Appendix 4 the results of the final questionnaire comparing NIST ZPRISE and the experimental system.

Finally, in Appendix 7, we provide the full transcript of participant S1 for the search scenario topic 326 "Ferry Sinking". This participant carried out 9 query iterations, consulting the CT related terms twice, and adding one related term in one query iteration. We note that the participant spends a lot of time expressing questions about the "+/-" syntax for queries. She also repeatedly expresses uncertainty about the type and amount of aspects to save.

IV. Individual Differerences in Aspects Identified by Participants

The aspects our participants identified are largely identical with what the NIST evaluators judged as relevant aspects. But there were some individual differences in how the aspects were recorded. These differences may imply different search strategies, and may also suggest a need for clearer instructions about what constititutes a "successful" search.

1. Brief keyword descriptions versus more detailed descriptions

While two participants quickly described the aspects they identified by simple keywords, the other two chose to describe the aspects in more depth. For example, for the topic 339 'Alzheimer's Drug Treatment', one participant just noted 'Cognex', or 'Tacrine', while another wrote down 'Cognex shows good results for FDA approval', or 'American Medical Association finds good hope in drug Tacrine'. Obviously, the former shorter descriptions were closer to the descriptions in NIST's outline.

2. Which aspects to look for?

The narrative for the topic 326 'Ferry Sinkings' read as follows: 'To be relevant, a document must identify a ferry that has sunk causing the death of 100 or more humans...' Instead of looking for various ferry sinking events, one of our participants looked for only one sinking event, and then concentrated on saving different aspects of this specific sinking event (in this case the Estonia ferry sinking in the Baltic Sea). When asked, the searcher explained that he was just following the task instructions in which it said to'...identify **a** ferry...'.

3. Is recording aspects a realistic aspect of search?

Whether described in brief or in a detailed way, all four participants felt the task of writing down the aspects they identified as an additional task burden. They had difficulties incorporating this task with the general search task, found it distracting, and had to be reminded of it a number of times.

4. Not enough time to search for all aspects

Only one of our participants was content throughout with the 20 minute time limit for a given search session. The other three believed that they were not finding all the aspects possible, and complained at one time or the other about not having had enough time to do so.

DISCUSSION

First, we give a tentative assessment of the performance of our participants, and which factors may have caused it. Then, we comment on concerns with the TREC6 Interactive Track methodology.

I. How did participants perform using our experimental search UI?

Searching is difficult in general, and judgments about the usability of our specific prototype system must be relative, of course, to other systems. We participated in the TREC6 program to get access to such comparative data.

However, from a qualitative point of view, we note that the search functions we provided, specifically the "+/-" syntax, and the related terms from the Context Thesaurus likely require more experience on the part of searchers to be used effectively. In terms of a system evaluation, this means evaluating longer-term use of these features and/or using more experienced search participants. There are also UI improvements we can make (e.g., performance), but even with these we suspect that there will be a learning curve for all but the most experienced searchers such as professional librarians. Also, adding new functions like related terms to basic search consumes extra time, and so must provide more benefit than doing other things, such as making another query or browsing more documents, or whatever. We had obviously hoped that our participants' performance would improve with these features compared to a system like the control system ZPRISE, which does not provide such features (at least not in the version that was adopted for the experiment). But it is not clear that this was the case. Also, participants may need to develop search strategies for when to do regular keyword search, and when to take the time to use additional features, like using the CT related terms. Once again, this points to the need to evaluate usage over a longer-term, or find radically usable implementations of these new techniques whereby users can immediately make effective use of the features with no training or practice.

Beyond potential UI issues for the features we included in the prototype UI, we note again that our search participants tended to be less experienced than those used by other site participants, and this may have led to some unproductive search strategies. One participant in particular had special trouble making effective queries because of his strategy of creating verbose query specifications, using many non-relevant terms. And, as we also noted earlier, we invited participants to ask questions (treat the experimenter as "help" system) which led to some level of conversational interchange. Lack of experience might account for the overall lower performance on both the control and experimental systems at our site, although the difference in performance between these two systems could still reveal one system as better than another.

We emphasize that the development of our prototype search UI is part of a longer-term design and evaluation process. We evaluated the prototype in order to iteratively improve the UI, and to provide feedback about various problems and options to developers. We note that independent of the quantitative results just summarized, the evaluation of our prototype UI provided much useful "qualitative" feedback, positive and negative about various aspects of the UI and the underlying search technology components. This feedback is of interest to our internal research and development colleagues, and will guide further refinement of the search UI, as will the overall recall and precision benchmarks of the TREC6 program.

II. TREC6 Interactive Track Goals and Methodology

We observe that in addition to the learning curve for the experimental system's UI and search functions, the TREC6 Interactive Track methodology also adds to the learning curve, e.g., reading the instructions, recording aspects, the format of the search scenarios. Streamlining the methodology might also make the evaluation's tasks easier for the participants. This section makes suggestions towards this end.

1. The Goal of TREC6 Interactive Track

As we noted above, we learned a lot about the search technology, and UI decisions we made independent of how our experimental system performed quantatively. The latter quantitative benchmarks established by other TREC6 participants will give us useful targets to achieve in the future. However, is this the only goal of TREC6 Interactive Track? Is there a way that specific search features and functions can be assessed in the context of overall performance? In some cases, of course, these features and functions may be proprietary but over time, it seems to us that the TREC6 community may want to contribute search UI insights to the larger Human-Computer Interaction community. Examples from our experimental prototype include a scratchpad for saving documents and lists of related terms for query refinement.

2. The Search Collection and Search Task

The Web sets users' expectations about what kinds of information can be searched for. The "Financial Times" collection is limited in timeliness and in topic coverage. Are there issues in generalizing what we learn about search from search tasks applied to this collection? For example, search on the Web leads to documents that can be browsed using hyperlinks. Hyperlinking and document browsing add new issues to searching for information.

3. A Critical View on the Written Test Instructions

We believe the written instructions used in the test could be improved and we identify potential areas below. Some items are repetitions of above mentioned ones, but are summarized here for completeness.

Practically **all** types of written instructions, including tutorials and help text (of the experimental system) gave rise to misunderstandings by at least one of the participants, but here we will concentrate on the experiment instructions which contained the general task instructions and the topic instructions, since they are the ones of general interest in this context.

3.1 Experiment instructions

a. How to record aspects
As mentioned above, some of the searchers rendered very detailed aspects descriptions. The instruction 'please write down a word or short phrase to identify the aspect - enough to keep track of which aspects you have found' just left too much room for different interpretations.

b. Misinterpretation of the instructions' example
The general task instruction provided by NIST presented an example to illustrate a typical aspects search that resulted in saving eight aspects (in five documents). As a consequence, two of our participants were ready to stop their search sessions after having found eight aspects. Only after intervention of one of the experimenters, they could be convinced to continue to find more aspects.

3.2 Topic instructions (description, narrative, aspects)

a. Not precise enough
Above we already mentioned that in one case the indefinite article 'a' was misunderstood as the numeral 'one'. Since the narrative required documents that '...identify **a** ferry...', this participant identified one specific ferry sinking event, and then searched for different aspects hereof. Not even the description in the aspects section could prevent him from adopting this view.

b. Too long and detailed?
Some of the topic instructions seemed to be too subtle to clearly draw a line between relevant and irrelevant. Especially the narratives of the topics 307 'New Hydroelectric Projects' and 303 'Hubble Telescope Achievements' seem to have confused some of our participants, who despite the exclusion specifications saved documents on aspects specified as irrelevant. On the other hand, we also observed the case where a participant tried to strictly follow the rules, but taking the topic instructions too literal by only saving documents that complied with **all** of the listed requirements. For example, only when a document was explicitly mentioning the name of a drug, its manufacturer, as well as its success rate, would this participant save this document for the topic 339 'Alzheimer's drug treatment', event though the last item, the success rate, did not seem to be imperative.

4. Controlling characteristics of Evaluation Participants: Should we Agree on Common Searcher Characteristics?

As we noted above, we hired the kind of people we expect to use IBM search technology in Web search services typical on the Web, as compared to professional librarians. However, we did not control or even measure the wide variety of individual differences we know exist between people. There were early, and inconclusive discussions amongst TREC6 Interactive Track participants about whether to provide guidelines on participant selection. We think there should be guidelines, and perhaps some standard measuring tools to address psychological attributes that may influence facility with electronic search systems. An example might be measures of computer literacy, or spatial ability for some UI features.

ACKNOWLEDGEMENTS

We are grateful to Jim Cooper (IBM Research) for letting us adapt Java code components for our prototype, and for advice on various aspects of the server programs used. We are also grateful to Eric Brown and Herb Chong (both of IBM Research) for helpful advice on the use of various server programs.

REFERENCES

Cooper, J. and Byrd, R. (1997). Lexical Navigation: Visually Prompted Query Expansion and Refinement. Proceedings of DIGLIB97, Philadelphia, PA, July, 1997 (pp. 237-246).

IBM (1997). Information about NetQuestion can be found at: http://www.software.ibm.com/data/mediaminer.

Landauer, T., (1987). Research methods in human-computer interaction. In M. Helander (Ed.), Handbook of Human-Computer Interaction. (pp. 905-928). North-Holland, Elsevier Science.

APPENDICES

Appendix 1: Search Participants' Characteristics

Education & Search Experience				
Judgements from 1..5:	None	... Some ...	A great deal	
	S1	S2	S3	S4
education	BA, MA	BA	Ass	Ass
age	under 21	60	25	45
gender	female	male	male	female
previous TREC	no	no	no	no
know test systems	no	no	no	no
online search. exp.	2 yrs.	no	5 yrs.	1 yr.
experience with:				
mouse-based IF exp.	5	3	5	5
comp. library catalogs	3	2	5	1
CD ROM sys.	1	1	5	3
comm. online sys. like Dialog, Lexis	1	1	1	1
web search	5	1	5	4
other systems	2	1		1
full-text DB	2	1	5	2
ranked IR	2	1	1	1
IR w. relev. feedback	2	1	5	2

Appendix 2: Users' Judgements of the Six Experiment Topics

Topics Judgements

Judgements from 1...5: Not at all... Marginally... Extremely

	S1 (ZP)	S2 (IBM)	S3 (ZP)	S4 (IBM)
Topic 303i - 'Hubble Telescope Achievem.'				
familiar topic?	3	3	1	1
difficult search?	2	1	1	4
satisfied w. results?	3	1	5	4
found all aspects?	4	1	4	4
enough time?	5	3	5	4
Topic 307i - 'New Hydroelectr. Projects'				
familiar topic?	4	1	2	1
difficult search?	2	1	1	2
satisfied w. results?	5	1	4	4
found all aspects?	5	1	3	2
enough time?	5	2	5	2
Topic 322i - 'Intern. Art Crime'				
familiar topic?	3	3	4	3
difficult search?	3	1	5	1
satisfied w. results?	4	1	1	5
found all aspects?	2	1	1	5
enough time?	4	1	1	5
Topic 326i - 'Ferry Sinkings'				
familiar topic?	1	1	3	3
difficult search?	3	1	1	3
satisfied w. results?	3	4	3	4
found all aspects?	3	1	3	4
enough time?	4	2	2	4
Topic 339i - 'Alzheimer's Drug Treatm.'				
familiar topic?	3	3	5	4
difficult search?	2	1	1	3
satisfied w. results?	3	1	4	5
found all aspects?	3	1	4	5
enough time?	5	3	4	5
Topic 347i - 'Wildlife Extinction'				
familiar topic?	2	3	5	4
difficult search?	2	4	1	4
satisfied w. results?	5	1	4	2
found all aspects?	5	1	2	2
enough time?	5	3	5	2

Appendix 3: Judgements of Experimental System

IBM's Search System

Judgemt from 1...5:	Not at all...	Marginally...	Extremely	
	S1	S2	S3	S4
easy to use	3	4	5	5
easy to learn	5	4	4	5
understand coverage	4	4	5	5
comments	"stop search" required		required: poss. to re-open docs from saved docs window	a bit slow...

Appendix 4: Judgement of NIST ZPRISE versus Experimental System

IBM Search	**versus**	**ZPRISE**		
Judgement from 1...5:	Not at all...	Marginally...	Completely	
	S1	S2	S3	S4
understand task	5	4	5	5
searches as usual	4	n/a	5	4
different systems	3	no answer	4	4
easier to use	IBM / ZP	ZP	IBM	ZP
easier to learn	ZP	ZP	IBM	ZP
like better	IBM / ZP	IBM	IBM	ZP
comments	ZP easier, since no + / - signs	IBM better look and feel	IBM better look and feel	ZP showed contained query terms

Appendix 5: Search Participants' Detailed Search Data

Subject 1	No. queries	No. q-terms; Av. per query	No. CT refer.s	No. CT usages	No. queries w. operators	No. docs viewed; Av. per query	No. docs saved; Av. per query	Aspectual recall
NQ-326	9	18 (2)	2	1	4	26 (2.88)	4 (0.44)	0.333
NQ-322	13	29 (2.2)	-	-	7	25 (1.9)	3 (0.23)	0.111
NQ-307	3	4 (1.33)	-	-	1	46 (15.33)	10 (3.3)	0.348
ZP-347	6	15 (2.5)	na	na	na	25 (4.16)	11 (1.83)	0.231
ZP-303	4	11 (2.75)	na	na	na	26 (6.5)	3 (0.75)	0.571
ZP-339	7	23 (3.28)	na	na	na	27 (3.85)	4 (0.57)	0.7

Subject 2	No. queries	No. q-terms; Av. per query	No. CT refer.s	No. CT usages	No. queries w. operators	No. docs viewed; Av. per query	No. docs saved; Av. per query	Aspectual recall
ZP-326	2	10 (5)	na	na	na	14 (7)	2 (1)	0.222
ZP-322	3	18 (6)	na	na	na	10 (3.3)	3 (1)	0
ZP-307	1	5 (5)	na	na	na	9 (9)	7 (7)	0.217
NQ-347	14	41 (2.9)	1	5	2	14 (1)	3 (0.214)	0.154
NQ-303	2	5 (2.5)	-	-	1	33 (16.5)	1 (0.5)	0
NQ-339	3	5 (1.66)	1	-	3	9 (3)	4 (1.33)	0.9

Subject 3	No. queries	No. q-terms; Av. per query	No. CT refer.s	No. CT usages	No. queries w. operators	No. docs viewed; Av. per query	No. docs saved; Av. per query	Aspectual recall
NQ-326	3	11 (3.66)	1	1	1	67 (22.33)	4 (1.33)	0.111
NQ-322	11	29 (2.63)	-	-	6	105 (9.54)	3 (0.27)	0.111
NQ-307	1	2 (2)	-	-	-	47 (47)	8 (8)	0.174
ZP-347	3	5 (2.66)	na	na	na	44 (14.66)	4 (1.33)	0.077
ZP-303	1	3 (3)	na	na	na	28 (28)	7 (7)	1
ZP-339	1	3 (3)	na	na	na	30 (30)	5 (5)	0.7

Subject 4	No. queries	No. q-terms; Av. per query	No. CT refer.s	No. CT usages	No. queries w. operators	No. docs viewed; Av. per query	No. docs saved; Av. per query	Aspectual recall
ZP-326	3	21 (7)	na	na	na	26 (8.66)	5 (2.66)	0.333
ZP-322	8	28 (3.5)	na	na	na	25 (3.125)	8 (1)	0.222
ZP-307	2	7 (3.5)	na	na	na	25 (12.5)	10 (5)	0.261
NQ-347	4	13 (3.25)	3	3	-	15 (3.75)	4 (1)	0.038
NQ-303	6	11 (1.83)	2	3	-	23 (3.83)	3 (0.5)	0.286
NQ-339	3	4 (1.33)	2	-	-	21 (7)	4 (1.33)	0.6

Appendix 6: Performance over Search Participant and Search Topic

Average	recall per	subject:
	Average NQ recall	Average ZP recall
Subject 1	0.264	0.501
Subject 2	0.351	0.146
Subject 3	0.132	0.425
Subject 4	0.308	0.272
Total	0.288	0.336

Average	recall per	topic:
	Average NQ recall	Average ZP recall
326	0.222	0.2775
322	0.111	0.111
307	0.261	0.239
347	0.096	0.154
303	0.143	0.785
339	0.75	0.7
Total	0.263	0.378

Appendix 7: <u>Search Session Report - Topic: Ferry Sinkings</u>

(E = experimenter; S = searcher; SYS = experimental system responses)

E: 'Twenty minutes, here we go. And if you have any questions or comments just say so, all right?'
S: 'Okay.'

types the query: ferry

SYS: displays 50 out of 1083 found documents - [we will show 8 documents here for all result lists]

*** FT922-12800FT 15 APR 92 - Freight ferry
*** FT924-13535FT 15 OCT 92 - New ferry service
*** FT923-3664FT 11 SEP 92 - World News in Brief: Ferry blown up
*** FT934-15680FT 11 OCT 93 - World News in Brief: Fears for Korean ferry
*** FT911-5368FT 15 APR 91 - World News in Brief: Ferries disrupted
*** FT923-9034FT 07 AUG 92 - Ferry row settled
*** FT944-10832FT 09 NOV 94 - SeaCat and ferry collision probed
*** FT924-3846FT 05 DEC 92 - World News in Brief: Ferry blaze

E: (giving support for scanning the result list)
'You are quicker with putting your cursor on the scroll bar.'
S: 'Oh, and just dragging it?'
E: 'Yes.'

S: looks at 2 documents
requests related terms for the query ‘ferry’.
P4482 Ferries
O European Ferries
Ex Ferry
P4481 Deep Sea Passenger Transportation

‘Now, how do I get rid of these. I just go back up there [into the query entry field], right?’
E: ‘Exactly. They will go by themselves. As soon as you press “Documents”, it will empty out everything.’

S: changes query to: ferry +sinking +deaths

SYS: displays 50 out of 85 -
***** FT923-4272FT 08 SEP 92 - Arts: Elisabeth
**** FT944-166FT 31 DEC 94 - Swedes count the European costs of bananas and whisky: Hugh Carnegy reports on a nation already unsettled by changes, violence and unemployment as it takes on membership of the EU
**** FT944-6607FT 29 NOV 94 - Technology: Closing in on a serial killer - Cancer kills five people every minute of the day. In the first of a six-part series, Clive Cookson reports on the war against cancer and the encouraging trends
**** FT932-15625FT 10 APR 93 - Sport: The cars are the same, the countries have changed - Motor racing
**** FT911-2716FT 27 APR 91 - Love and death in the city of lost hopes: It is two years since tanks crushed the Peking rebellion. In the heady days before the uprising Peter Ellingsen fell in love with a young student - but the thugs
*** FT922-1708FT 20 JUN 92 - Books: Caught up in family sagas - Fiction
*** FT943-6370FT 30 AUG 94 - Spanish vow to banish drift-nets: Fishing disputes have risen up the diplomatic agenda. FT reporters examine the conflicts worldwide
*** FT943-6371FT 30 AUG 94 - Fleets fight in over-fished waters: Fishing disputes have risen up the diplomatic agenda. FT reporters examine the conflicts worldwide

S: changes query to: ferry +sinking

SYS: displays 50 out of 1383 -
***** FT942-14757FT 19 APR 94 - Letters to the Editor: Channel control overdue
**** FT931-12892FT 27 JAN 93 - Insurers given credit for fall in ships lost
**** FT911-5168FT 16 APR 91 - Threequarters of sunken oil tanker's cargo still on board
*** FT923-2875FT 16 SEP 92 - World News in Brief: Sparring partners
*** FT922-14951FT 03 APR 92 - World News in Brief: Ship's master faulted in Titanic sinking
*** FT923-6450FT 25 AUG 92 - World News in Brief: Malaysia hunts Taiwan trawler
*** FT924-12163FT 22 OCT 92 - Start on N-waste dump planned
*** FT934-14903FT 14 OCT 93 - Observer: Sinking feeling

S: looks at 1 document
‘I think I was better with just “ferry”.’
changes query back to: ferry

SYS: displays 50 out of 1083 found documents -
*** FT922-12800FT 15 APR 92 - Freight ferry
*** FT924-13535FT 15 OCT 92 - New ferry service
*** FT923-3664FT 11 SEP 92 - World News in Brief: Ferry blown up
*** FT934-15680FT 11 OCT 93 - World News in Brief: Fears for Korean ferry
*** FT911-5368FT 15 APR 91 - World News in Brief: Ferries disrupted
*** FT923-9034FT 07 AUG 92 - Ferry row settled
*** FT944-10832FT 09 NOV 94 - SeaCat and ferry collision probed
*** FT924-3846FT 05 DEC 92 - World News in Brief: Ferry blaze

S: looks at 3 documents

'So, am I supposed to write something about this if I am going to save it?'
E: 'Yes.'
S: 'Like write something about the title?'
E: 'No, write down something that says that this is about the Bangladesh Ferry disaster, that's fine.'
S: 'Okay.'
E: (commenting on UI issue) 'Actually, you only need to click on it (push button) once. I know we had some problems yesterday, like, when you click on it, it may not come back up again.'

S: saves 1 document

'So, I'm only supposed to save one?'
E: 'One per each sinking event with more than 100 dead people - I don't like this topic...'
S: 'So, it doesn't matter. I just have to save as many as I can that have 100 deaths and describe where it is?'
E: 'Yes. Have you ever heard about any ferry sinkings so that you can think of any?'
S: 'No, not really, no idea...'

looks at another 3 documents
saves 1 document
looks at another 2 documents

changes query to: ferry +europe

SYS: .... takes long time ....

E: 'I just know that the search engine is mainly looking for Europe now and...'
S: 'not ferries...'
E: 'among the documents of Europe, also for ferries...'
S: 'So, it would have been better if you wrote Europe plus ferries, then it would have looked for ferries more than Europe?'
E: 'The plus is always the most important.'
S: 'All right, is there any way I can stop this?'
E: 'No, unfortunately not. This Europe was prevalent and it had 36,000 documents, I just happen to know...'
S: 'Well, why wouldn't it do the ferry stuff first? Do you know what I'm saying?'
E: 'Good question.'
S: 'Like if I had put this first...I'll do it that way.'

SYS: displays 50 out 36711 found documents
*** FT941-5746FT 04 MAR 94 - International Company News: Write-off hits Norwegian ferry operator
*** FT922-8897FT 11 MAY 92 - Letter: Right place for locating EuroFed
*** FT911-4515FT 19 APR 91 - Survey of the Canary Islands (8): Green, fertile and volcanic - Western Islands
** FT922-8062FT 15 MAY 92 - Survey: FT Traveller, Genoa (6) - Symbol of lost opportunities - The Harbour, The flag of privatisation is now flying highest at the port - surprisingly, for a former centre of hard-left politics
** FT941-4041FT 12 MAR 94 - Travel: Light fantastic - Why artists colonised a remote Danish fishing village
** FT911-4516FT 19 APR 91 - Survey of the Canary Islands (4): Ports provide that vital economic lifeline - Santa Cruz de Tenerife and Las Palmas de Gran Canaria face competition from Agadir
** FT922-7181FT 20 MAY 92 - Arts: The feel-good factor - Television
** FT921-13774FT 20 JAN 92 - The Week Ahead

S: changes query back to: ferry
and adds a related term to the query: 'O European Ferries'
so query now reads:
ferry 'O European Ferries'

E: 'Take that "O" out of there, it's not going to do anything right now.'
S: 'Why did it put that?'
E: 'These related terms are processed automatically. Many funny terms come from the fact that the Financial Times has specific types of tags that they use.'
S: 'So, I should erase this?'
E: 'Yes.'

[here, the system takes a long time, because of a user / UI error and the intermediate dialog is about this behaviour]

SYS: displays 50 out of 1083 found documents
***** FT941-732FT 29 MAR 94 - Netherlands ferry route may restart
**** FT934-15859FT 09 OCT 93 - P&O announces 240 job cuts
*** FT931-6871FT 26 FEB 93 - P&O negotiating sale of two ferries to Greek shipping companies
*** FT941-3869FT 14 MAR 94 - FT Guide to the Week
*** FT941-10270FT 10 FEB 94 - Management (Marketing and Advertising): Softly, softly approach - The French enlisted history-makers to market the Channel Tunnel, but UK advertising has had to be coaxing and humorous
*** FT941-6094FT 03 MAR 94 - Technology: Check-in for the Channel rail link - A new passenger ticketing system
*** FT922-12800FT 15 APR 92 - Freight ferry
*** FT924-14523FT 09 OCT 92 - Ferguson wins Pounds 16m ferries order

looks at 1 document

S: changes query to: ferry "European Ferries"

SYS: displays 50 out of 1083 found documents -
***** FT941-732FT 29 MAR 94 - Netherlands ferry route may restart
**** FT934-15859FT 09 OCT 93 - P&O announces 240 job cuts

*** FT931-6871FT 26 FEB 93 - P&O negotiating sale of two ferries to Greek shipping companies
*** FT941-3869FT 14 MAR 94 - FT Guide to the Week
*** FT941-10270FT 10 FEB 94 - Management (Marketing and Advertising): Softly, softly approach - The French enlisted history-makers to market the Channel Tunnel, but UK advertising has had to be coaxing and humorous
*** FT941-6094FT 03 MAR 94 - Technology: Check-in for the Channel rail link - A new passenger ticketing system
*** FT922-12800FT 15 APR 92 - Freight ferry
*** FT924-14523FT 09 OCT 92 - Ferguson wins Pounds 16m ferries order

S: looks at 5 documents
saves 1 document
looks at another 3 documents

E: 'Another point is that if you use more than one term, and because we really want to concentrate on ferry now, so put a "+" in front of ferry. Ferry is our basic one. If you only use ferry, then it knows you only want to look for ferry. If you have more than one term, you have to tell them, what is your most important aspect, or topic here.'

S: changes query to: +ferry sinking

SYS: displays 50 out of 1083 found documents -
***** FT944-15057FT 20 OCT 94 - Improved ferry safety urged
***** FT944-5084FT 06 DEC 94 - Ro-ro ferry study agreed
***** FT944-11367FT 07 NOV 94 - Pounds 45m car-ferry research planned
***** FT944-5248FT 05 DEC 94 - Sea safety review focuses on ferries
**** FT944-11048FT 08 NOV 94 - Bow doors faulty on 33% of ferries using UK ports: Government to increase safety checks on vessels
**** FT944-11013FT 08 NOV 94 - International Company News: Heavy loss in US pushes Trygg-Hansa into the red - Swedish insurer posts SKr813m deficit at nine months
*** FT944-10102FT 12 NOV 94 - Tighter ferry rules proposed
*** FT944-10109FT 12 NOV 94 - Tighter ferry rules proposed

S: 'It's hard to find them that say 100 people...a lot of them say "less than 100".'
E: 'That document is an 88 minus...'
S: 'Makes me not want to go on a ferry now.'

[intermediate dialog about a document the searcher thought she had saved, but was not recorded in the saved documents list]

looks at 5 documents

'O, nine hundred!'
E: 'That was a big European event. That was a very, very bad one.'

S: saves 1 document

'So, how many more do I have to look for? I already saved three, am I supposed to look for more than that?'
E: 'I think there are more than that. Now, if you try to do it with the plus ferry and whatever else...'

S: ‘Do I have to put the plus though?’
E: ‘No.’
S: ‘So, I'm not supposed to save them again right?’
E: ‘No. So, think of some other means of excluding Estonia. You don’t want to see Estonia anymore, because you saved it.’
S: ‘Oh, so I put ‘minus’, right?’
E: ‘Right.’

S: changes query to: +ferry sinking -estonia

SYS: displays 50 out of 1029 found documents -
**** FT943-1964FT 21 SEP 94 - Survey of Logistics (7): Flags of inconvenience - Competition flares in Europe's sea-lanes
*** FT922-12800FT 15 APR 92 - Freight ferry
*** FT923-3664FT 11 SEP 92 - World News in Brief: Ferry blown up
*** FT924-13535FT 15 OCT 92 - New ferry service
*** FT934-15680FT 11 OCT 93 - World News in Brief: Fears for Korean ferry
*** FT911-5368FT 15 APR 91 - World News in Brief: Ferries disrupted
*** FT934-1954FT 16 DEC 93 - Technology: Ships bridge the danger gap - Andrew Fisher concludes a series on transport safety with an investigation into innovations that may help prevent sea disasters and give clues to their
*** FT924-3846FT 05 DEC 92 - World News in Brief: Ferry blaze

S: looks at 6 documents

E: ‘Do you have the feeling that you’ll find any others?’
S: ‘They are all the same basically. Like all the ones that have more than 100 deaths seem to be all the same in every area that I put.’
E: ‘So, you don’t find new instances?’
S: ‘No, not with more than 100 people.’
E: ‘Well, if you think you are done... - we’ll be close to being done here anyhow. This thing (the alarm clock) will ring in a minute.’
S: ‘What do I do with the ones that I already saved? Just leave them?’
E: ‘Go back, say ‘Browse Saved Documents’.’

S: browses saved documents

[again a brief dialog about documents apparently missing from the saved documents list]

Buzzer rings.

E: So, let’s stop here.