Rutgers Interactive Track at TREC 2002

N.J. Belkin, C. Cool*, D. Kelly, G. Kim, J.-Y. Kim, H.-J. Lee, G. Muresan, M.-C. Tang, X.-J. Yuan
School of Communication, Information & Library Studies, Rutgers University and
*Graduate School of Library and Information Studies, Queens College, CUNY
[belkin | diane | gkim | jaykim | hyukjinl | muresan | muhchyun | xjyuan]@scils.rutgers.edu *ccool@qc.edu

1 Introduction

Two important results came out of our investigations in the TREC 2001 Interactive Track (Belkin, et al., 2002). One was that the greater the amount of interaction that searchers engaged in, the lower their satisfaction with the results of the search. We understood this to mean that interaction effort was inversely related to search satisfaction, and therefore, that making interaction more effective would lead to increased search satisfaction. The second was that performance in the searching task increased with query length. We conjectured that this was due, at least in part, to the subjects having searched using a best-match search engine (Excite[1]), as well as longer queries being better able to express the information problem. These two findings became the basis for our systems and experiments in the TREC 2002 Interactive Track. We formed the following hypotheses:

1. A system designed to reduce the amount of interaction that a searcher has to engage in, by making it more effective, will lead to increased satisfaction with search results, and increased performance, as compared to a system not so designed;
2. A system which encourages long queries will lead to better performance in the search task than one which does not.

In order to test the first hypothesis, we designed two basic interfaces to the Panoptic search engine[2]: one which presented the results of a query as a ranked list of titles of documents, twenty at a time; the other which presented the results of a query as the texts of four documents at a time, each in a scrollable window, ranked in the same order as the first interface. The second interface was intended to reduce user interaction with the system by virtue of not requiring the searcher to follow links from the search results to the actual documents and then back again to the results list, as in the first interface. It was also thought that being able to see the documents immediately would make it easier and faster to evaluate their potential relevance to the search topic, than having first to evaluate on the basis of a title plus snippet surrogate, and then do a second evaluation based on the page itself.

To test the second hypothesis, we designed two different query elicitation methods, that were used in both interfaces. One method had just the word "query" above the box in which the query was to be entered. When this version of either interface was demonstrated to the subjects in the experiment, the experimenter would enter the query as a list of words and phrases. The second method had, above the query entry box, the following: "Information problem description (the more you say, the better the results are likely to be)". When this version of the interfaces was demonstrated, the experimenter entered one or more complete sentences or questions descriptive of the topic and desired results. The second condition was predicted to lead to longer queries, both in terms of all of the words entered, and in terms of the words that were finally interpreted by the Panoptic engine, which used a stop list.

Of course, the treatments which we designed were themselves only predicted to have the desired results. Therefore, in order to investigate the hypotheses, it was first necessary to determine whether these different treatments did in fact lead to the desired results, i.e. less interaction and longer queries. In a sense, then, the specific treatments were hypotheses themselves, which we also investigated. This paper therefore presents results with respect to hypotheses 1 and 2, above, and with respect to the following hypotheses:

3. A search interface which directly presents the ranked documents retrieved by a search will lead to less user-system interaction than one which presents ranked titles and requires following links to view documents;
4. A search interface which asks searchers to describe their information problems at length will lead to longer queries than one which asks searchers to simply input a query as a list of words or phrases.

---

[1] http://www.excite.com
[2] http://trec.panopticsearch.com/

In addition, the actual implementations of the interfaces themselves may strongly influence user behavior. Therefore, we also present results with respect to usability of, and satisfaction with the interfaces and their various characteristics.

2 Systems, topics and database

In common with the other participants in the TREC 2002 Interactive Track, we used the Panoptic search engine, and the related TREC 2002 Web Track collection, as the basic retrieval system and database. We performed no modifications to the database or retrieval results. We also used the standard eight Interactive Track search topics to specify the tasks that our subjects would perform. Panoptic is basically a best-match search engine, but for queries of four words or less, it instead ranks documents according to a coordination-level algorithm. We decided that this difference would not affect hypothesis 2, since even for such queries, there should be a fairly close match to the results of the best-match algorithm.

All searches were performed using a Sun UltraSparc-IIi (440Mhz) with 512M memory and a 21 inch monitor. The two basic interfaces were implemented using Swing of Java 2 SDK, version 1.3. The four-document-at-a-time interface, called MDD, is shown in figure 1 (with the "information problem" query elicitation, called QE). The twenty-title interface, which used the standard Panoptic result format, called SDD, is shown in figure 2 (with the "query" query elicitation, called NQE)

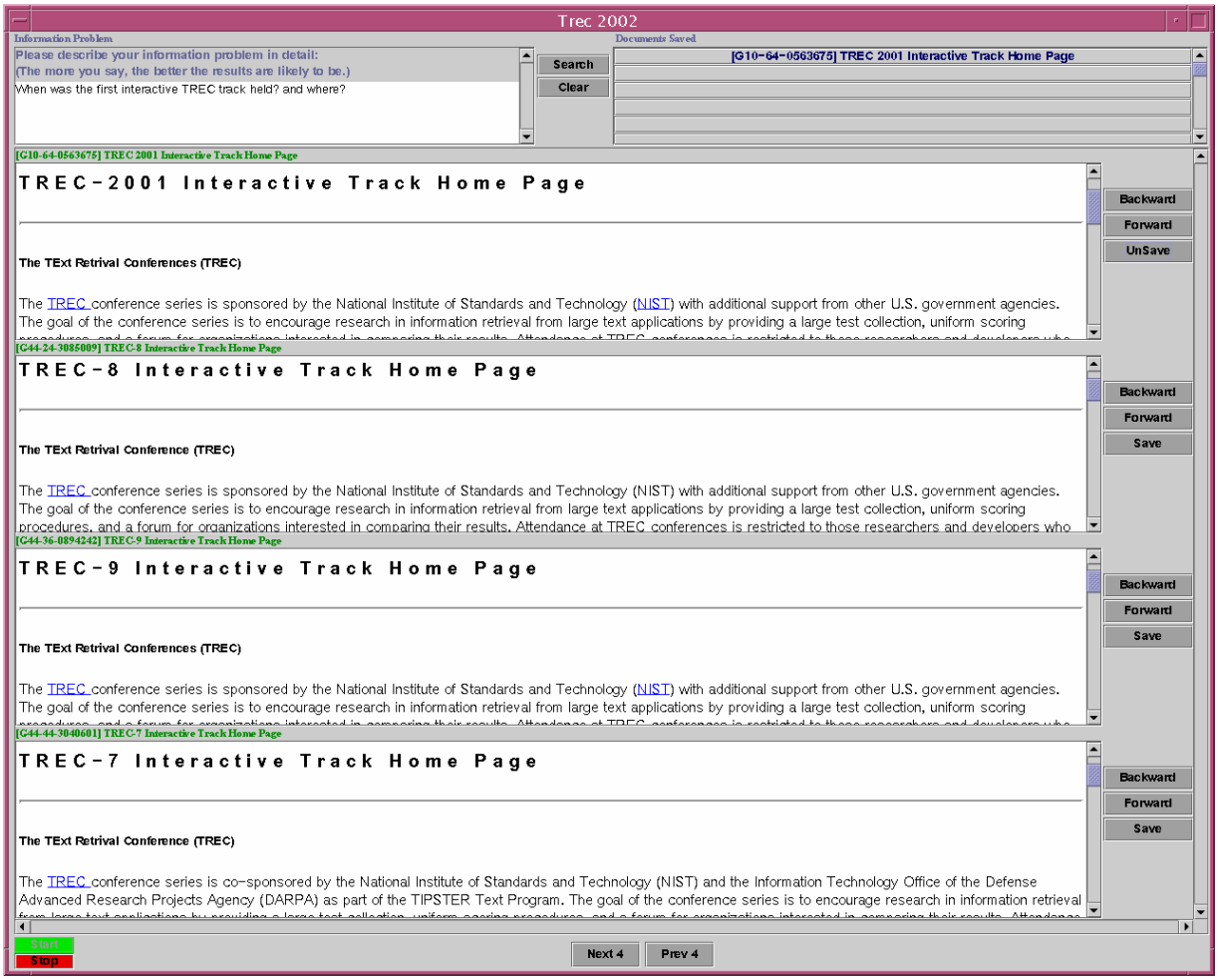

Figure 1. MDD interface, with query enhancement.

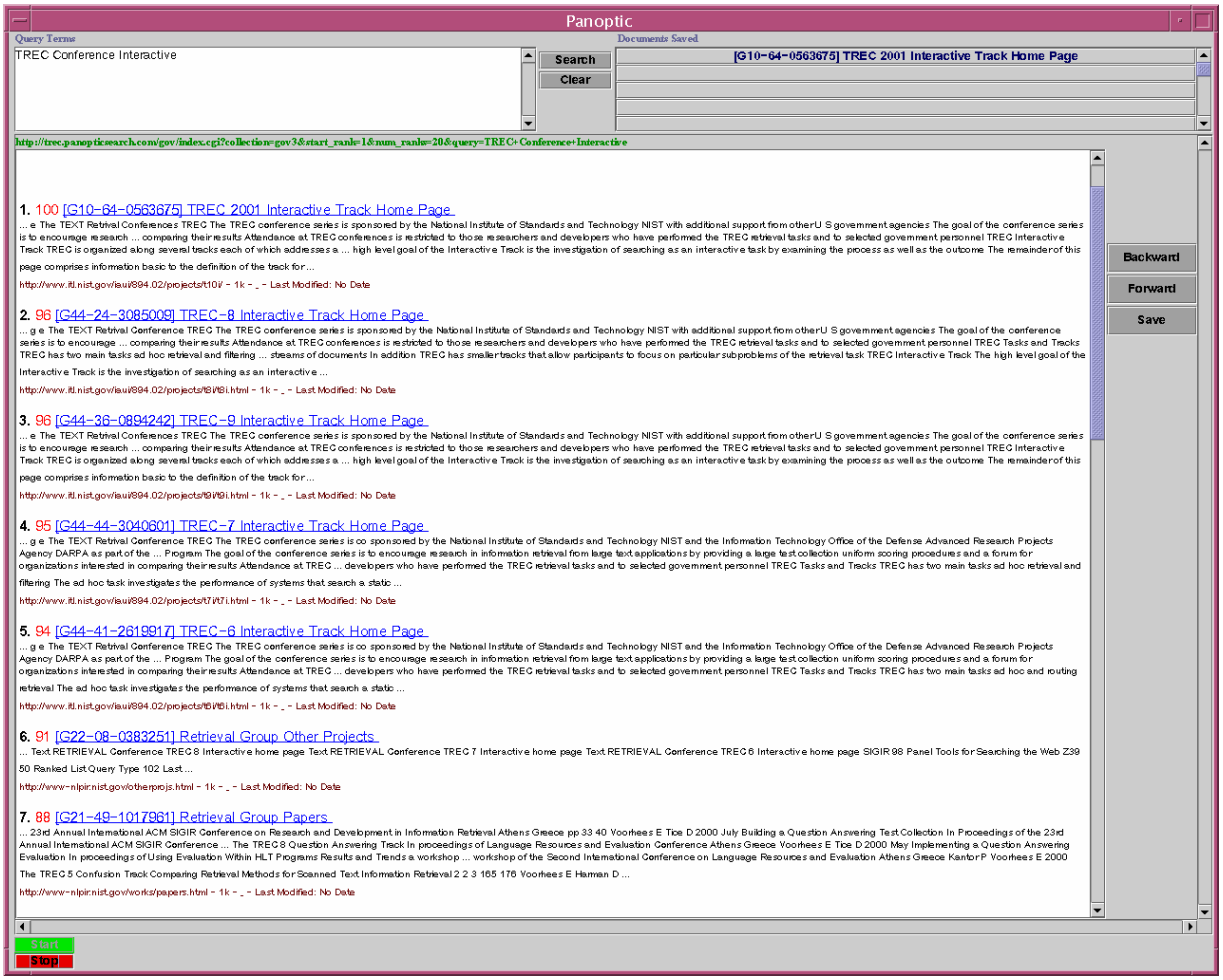

Figure 2. SDD interface, with no query enhancement

As can be seen from the screen shots, both interfaces had identical query entry boxes, and an identical list of saved documents. Saved documents in each interface could be opened for review, and unsaved if so desired. Each interface allowed subjects to follow links from displayed documents, whether the linked documents were in the Web Track collection, or on the live Web outside the collection. In the MDD system, subjects could page through the ranked document list four documents at a time; in the SDD system, subjects could page through the ranked title list twenty documents at a time. In general, seven to eight of the twenty titles were visible on the SDD screen without scrolling in the page; the first fifteen or so lines of a document were visible in each of the four MDD document panes, without scrolling. In MDD, documents could be saved directly by the appropriate button next to the displayed document; in SDD, they could be saved by following the link to the document, and then using the save button next to the saved documents list at the right top interface frame. Documents in SDD could also be saved directly from the results list, without following the link to the whole document, by selecting the relevant title and using the save button. However, this feature was not mentioned in the system demonstration, so it was used only rarely. When links within documents were followed, in either MDD or SDD, the searcher could return to the previous document by using the "Backward" button, and refollow links by using the "Forward" button. In SDD, returning to the search result list from a viewed page required using the "Backward" button.

3 Experiment design and conduct

We followed the basic Interactive Track within-subjects design for investigating the hypotheses related to interaction (1 & 3). With respect to the hypotheses related to query length (2& 4), we iterated the basic Interactive Track design twice, once in the QE condition, and once in the NQE condition, thus using a between subjects design. In both cases, subjects searched for answers to four topics using one interface, and then for four topics using the

other interface. The assignment of subjects to conditions MDD and SDD, and the topics that were searched in each, is shown in table 1. This design was applied to the first sixteen subjects with query elicitation mode NQE, and repeated for the second set of sixteen subjects with query elicitation mode QE.

Subject	**Block 1** **System: Topics**	**Block 2** **System: Topics**
1	SDD: 4-7-5-8	MDD: 1-3-2-6
2	MDD: 3-5-7-1	SDD: 8-4-6-2
3	MDD: 1-3-4-6	SDD: 2-8-7-5
4	MDD: 5-2-6-3	SDD: 4-7-1-8
5	SDD: 7-6-2-4	MDD: 3-5-8-1
6	SDD: 8-4-3-2	MDD: 6-1-5-7
7	MDD: 6-1-8-7	SDD: 5-2-4-3
8	SDD: 2-8-1-5	MDD: 7-6-3-4
9	MDD: 4-7-5-8	SDD: 1-3-2-6
10	SDD: 3-5-7-1	MDD: 8-4-6-2
11	SDD: 1-3-4-6	MDD: 2-8-7-5
12	SDD: 5-2-6-3	MDD: 4-7-1-8
13	MDD: 7-6-2-4	SDD: 3-5-8-1
14	MDD: 8-4-3-2	SDD: 6-1-5-7
15	SDD: 6-1-8-7	MDD: 5-2-4-3
16	MDD: 2-8-1-5	SDD: 7-6-3-4

Table 1. Experimental design comparing MDD and SDD. NQE was used for the first 16 subjects, QE for the second set of 16 subjects.

All searching was done at the Information Interaction Laboratory at the School of Communication, Information and Library Studies (SCILS), Rutgers University. When subjects arrived, they were asked first to examine and sign the Informed Consent form[3]. They then completed a background questionnaire, eliciting various demographic data and data concerning searching experience. Next, the experimenter gave a demonstration of the first interface that the subjects would use, which was based on an example topic of the sort that the subjects would be searching on. The subjects were then given a paper form with a description of the first topic that they were to search on, and questions about whether they thought they knew the answer to the topic's question, and their confidence in that knowledge, which they answered at that time. Then, the subjects returned to the computer, were instructed that they would have up to ten minutes to complete the search, that they were to save those documents which helped them to answer the topic's question, and were asked to think aloud during the search. The computer monitor was videotaped during all searches, and the thinking aloud was recorded on the videotape. When the subjects thought they had answered the question, or when they had run out of time, the system was stopped, and the subjects were asked to fill out a questionnaire with respect to their satisfaction with the results of the search, and other characteristics of the search on that particular topic. This procedure was repeated for the next three topics. After the first four topics, subjects were asked to complete a questionnaire regarding their experience searching with that particular interface. They were then given a demonstration of the second interface that they were to use, and then the same procedure was followed for the next four topics. After the second post-system questionnaire, subjects were engaged in a semi-structured exit interview, which was tape recorded. This questionnaire elicited information about common features of the two interfaces, and also comparing the two interfaces. The entire procedure was typically finished in about two hours. All of the data collection instruments, and the scripts for the demonstrations, are available at http:/scils.Rutgers.edu/mongrel/trec2002/instruments

[3] Project approved by Rutgers IRB, number 01-407M.

4 Results

4.1 Subjects

Thirty-two volunteer subjects participated in this experiment. They were recruited largely from the student population at Rutgers SCILS (44% were full-time students), and some were given credit for participating in the experiment and writing a brief description of their experience. Twenty-six (81%) of the participants were female and 6 (19%) were male. Our subjects were most likely (47%) to be between 28-37 years of age, while their ages ranged overall from 18 to 57. Given our sampling strategy, it is unsurprising that the searchers in our study had a high level of education. Thirty-seven % had completed a Master's degree at the time of the experiment and nearly half (47%) said that they hoped to complete a Master's degree. Table 2 presents a descriptive profile of the searchers' level of experience with computers. It should be noted that all of the subjects were required to have some experience using Web search engines.

Experience:	N	Minimum	Maximum	Mean	Std. Deviation
Computers, general	32	4	7	6.28	.772
WWW browsers	32	5	7	6.38	.751
Computers at work	31	1	7	6.48	1.18
Academic computing	32	2	7	6.50	.984
Personal computing	32	2	7	6.66	.971
Entertainment	31	2	7	5.39	1.65
Search engines	32	5	7	6.28	.683
OPACS	32	3	7	5.44	1.16
Indexing Services	31	1	7	3.71	1.736

Table 2 Subject Experience with Computers (Based upon a 7 point scale in which 1= None 4=Some 7=A great deal)

Our subjects reported having an average of 6.2 years of searching experience. Using a 7 point scale to measure experience, in which 1=Novice and 7=Expert, the self-assessed level of expertise with computers was, on average, 5.19. Table 3, below presents the frequency with which the participants in our study engaged in a variety of searching activities. Two things are interesting to note from this table. First, our subjects engaged in these searching activities with a fairly high degree of frequency overall. Secondly, it is interesting that of all the searching activities we asked about, searching for government/policy information ranked last in terms of frequency, while searching for project related activities and for entertainment ranked highest.

Searching for:	N	Minimum	Maximum	Mean	Std. Deviation
Projects	32	2	7	5.84	1.05
Shopping	32	1	6	3.94	1.39
Traveling	32	1	6	3.53	1.52
Medical/health	32	1	6	3.34	1.66
Gov't/policy	32	1	6	2.56	1.48
Entertainment	32	1	7	4.44	1.52

Table 3 Subjects' Frequency of Searching (Based on a scale in which 1=Never 4=Monthly 7=Daily)

4.2 Measures and definitions

The variables used to characterize user searching behavior, and their definitions, are shown in table 4. Performance was measured by number of documents saved per search (cf. Belkin, et al., 2001), by user satisfaction with the search (on a seven-point scale, anchored by *Not at all* and *Extremely*, administered at the conclusion of each search), and by correctness and completeness of answer for the topic. Correctness and completeness were determined by comparing the pages which were saved for a search with judgments performed by experimenters at all of the TREC Interactive Track sites of all of the pages which were saved, at all sites, for each topic. Each page was judged as to whether it contained a correct answer to the topic, and if so, in cases where it was relevant, what aspects of the topic each page addressed. Thus, topics 1 ,2, 4, 5 and 6, which asked searchers to identify some specified number of pages

or aspects could have incorrect, correct but incomplete, and correct and complete answers. Topics 3, 7 and 8, which asked for only one site or page, could have only correct or incorrect answers. In this paper, we consider an answer to be correct only if it is complete as well.

Variable	Definition
Pages seen	The total number of title references to pages displayed to the searcher through the course of the search (valid only for SDD)
Unique pages seen	The number of unique title references to pages displayed to the searcher (removing duplicate occurrences of references)
Pages viewed	The total number of pages whose contents were displayed to the searcher
Unique pages viewed	The number of unique pages whose contents were displayed to the searcher (removing duplicate occurrences of pages)
Number of documents saved	The total of all documents which were saved by the searcher through the course of the search
Number of final saved documents	The number of documents which were marked as saved at the conclusion of the search
Number of iterations	The total number of queries issued by the searcher, through the course of the search
Mean query length	The average length of all queries in a search, in words (both with and without stoplist applied)
Unique query length	The total number of unique words used in all of the queries in a search (both with and without stoplist applied)

Table 4. Variables used to describe search behavior

4.3 Descriptive statistics

Table 5 describes overall behaviors for all searches in both systems. The average number of the total pages seen and the average number of the unique pages seen were 145.16 and 56.38 respectively (relevant in SDD only). Meanwhile, the average number of the total pages viewed and the average number of the unique pages viewed were 13.64 and 10.60, in MDD and SDD together, respectively. On average, almost three documents (2.91) were ever saved by the subjects, and somewhat over 2 (2.33) were kept as finally saved documents. The subjects, on average, used just over two iterations (2.25) for their searching. Finally, subjects spent about 8 minutes and 21 seconds for each topic.

	Mean (Standard Deviation)	N
Total pages seen	145.16 (84.48)	128 (SDD only)
Unique pages seen	56.38 (39.05)	128 (SDD only)
Total pages viewed	13.64 (12.09)	255
Unique pages viewed	10.60 (9.68)	255
Documents ever saved	2.91 (2.18)	255
Final saved documents	2.33 (1.53)	255
Iterations	2.37 (1.52)	255
Time (seconds)	501.02 (195.06)	255

Table 5. Overall search characteristics, MDD and SDD together.

Search behavior ranged widely according to the topic (see table 6). First, the average total pages seen ranged from 116 to 185, and the average number of unique pages seen ranged from 37 to 84. All topics, except topic 2 (about 19 pages), had similar average total pages viewed, between 11 and 14. Also, the average numbers of unique pages viewed ranged from 9 to 11 except topic 2 (about 20 pages). Topic 7, with the smallest average number of final saved documents (1.66) had the largest average number of iterations (3.03) and unique pages seen (84.71). Conversely, topic 5, with the largest average number of final saved documents (2.97) showed the smallest average

number of iterations (1.66) and unique pages seen (37.33). Subject used the least searching time for topic 5 (6 minutes and 33 seconds); the average was 8 minutes and 21 seconds.

	Total	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
Total pages seen	145.16	161.00	134.67	116.82	150.40	116.00	147.06	150.59	185.40
Unique pages seen	56.38	63.53	56.00	44.94	49.33	37.33	43.53	84.71	70.20
Total pages viewed	13.64	12.53	19.53	12.38	10.81	12.88	13.84	13.97	13.13
Unique pages viewed	10.60	10.06	15.22	9.50	8.59	9.66	11.31	11.19	9.23
Number of document ever saved	2.91	3.03	2.56	2.87	3.03	3.28	3.69	1.78	3.00
Number of final saved documents	2.33	2.06	2.38	2.25	2.50	2.97	2.75	1.66	2.06
Number of iteration (queries)	2.37	2.28	2.78	2.03	1.66	1.66	2.13	3.03	2.45
Number of seconds taken	501.02	500.84	578.31	528.53	463.22	392.63	510.75	546.62	486.84

Table 6. Search characteristics by topic.

4.4 Interaction

Hypothesis 3 asked whether the MDD interface resulted in less user interaction than the SDD interface. Table 7 displays and compares the amount of interaction in each system, according to iterations, time (seconds), number of seen and viewed documents (total and unique), and ratios of unique to total seen and viewed, and in the case of SDD, ratios of seen to viewed, total and unique. For MDD, there were no seen documents, as the full texts of documents were always displayed to the subject. From table 7, we see that MDD had significantly more viewed documents than SDD, within a similar amount of time and number of iterations. Also, MDD subjects viewed far fewer documents than SDD subjects saw. While we did not log the amount of scrolling within particular documents or the number of times subjects paged to the next display of twenty titles in SDD or four documents in MDD, from the total number seen in SDD and total number viewed in MDD, we see that subjects paged less frequently in MDD (5) than in SDD (7). Although there was not a significant difference between the two systems in iterations or time, the differences are in the expected direction. On the basis of these data, we conclude that Hypothesis 3 is supported.

Hypothesis 1 asked whether a system designed to reduce the amount of interaction that a searcher has to engage in (i.e. make interaction more effective) will lead to increased satisfaction with the search results, and increased performance, as compared to a system not so designed. Table 8 displays and compares subjects' satisfaction with the search results and subjects' performance according to number of documents saved and number of correct answers. From table 8, we see that subjects were significantly more satisfied with their search results when searching with MDD than when searching with SDD. In terms of performance, subjects saved significantly more documents when searching with MDD than when searching with SDD, but the number of complete and correct answers to the topics did not vary significantly between interfaces.

Interaction Measure	MDD	SDD
Iterations	2.33 (1.52)	2.41 (1.53)
Time (seconds)	481.87 (199.74)	520.03 (189.05)
Number Seen (total)	N/A	145.16 (84.48)
Number Seen (unique)	N/A	56.38 (39.05)
Number Viewed (total)*	20.00 (13.76)	7.32 (4.90)
Number Viewed (unique)*	16.32 (10.73)	4.92 (2.83)
Ratio of unique seen to total seen	N/A	.44 (.24)
Ratio of unique viewed to total viewed*	.86 (.16)	.76 (.22)
Ratio of unique seen to unique viewed	N/A	.12 (.11)

Table 7. Interaction measures for MDD and SDD, mean and (standard deviation) (*p<.01)

Satisfaction or Performance Measure	MDD	SDD
Satisfaction with search results*	4.65 (2.00)	3.95 (2.09)
Number of documents saved*	2.77 (1.75)	1.91 (1.20)
Number of correct answers	93/127 = 73%	84/128 = 66%

Table 8. Satisfaction and Performance Measures for MDD and SDD, mean and (standard deviation) (*p<.01)

When combined with the interaction data above, the performance data provides additional evidence that MDD not only decreased interaction, but made interaction more effective. In the same number of iterations and in the same amount of time, subjects using MDD viewed significantly more documents and saved significantly more documents than those subjects using SDD. Subjects using MDD saved approximately 13% of the documents that they viewed, while subjects using SDD saved approximately 26% of the documents that they viewed, but only 1% of the documents that they saw. Of the documents that subjects using SDD saw, only 5% were viewed.

4.5 Query length

Hypothesis 4 asked whether the QE query elicitation mode resulted in longer query length than the NQE mode. Table 9 shows, for all searches in each condition, the mean query length, both with and without applying a stoplist. These figures are the mean of the number of words in each query in a search. The unique query length is the mean number of word types used in all queries in a search. Thus, the mean query length for a single search which used the two queries below is four (four terms in each query), while the unique query length is six (six unique words in the two queries).

Q1: usa congress privacy legislation Q2: usa congress electronic information

We interpret mean query length as a valid measure of the length of queries entered by the searcher. Unique query length, however, is interpreted as a measure of search effort, rather than of query length, since it measures the number of different words that the searcher had to think of over the course of the entire search.

	Mean iterations per search (SD)	Mean Query Length, stoplist (SD)	Mean Query Length, no stoplist (SD)	Unique Query Length, stoplist (SD)	Unique Query Length, no stoplist (SD)
NQE	2.64 (1.63)	4.24 (1.26)	4.85 (1.52)	5.97 (2.32)	6.85 (2.80)
QE	2.09 (1.35)	6.45 (3.00)	10.90 (7.30)	7.84 (3.34)	12.98 (7.33)

Table 9. Query statistics for NQE and QE modes, mean and (standard deviation).

Results from a t-test comparing QE and NQE on the basis of mean query length with stoplist indicate that searchers using the QE interface entered significantly longer queries (M=6.45; SD=3.00) than those using NQE interface (M=4.24; SD=1.26), $t(253) = -7.67$, $p<.01$. Thus, hypothesis 4 is strongly supported.

Hypothesis 2 asked whether a system which encouraged longer queries led to increased performance. Given the results with respect to hypothesis 4, we can investigate this hypothesis directly by comparing NQE with QE. As with

interaction, we evaluate performance with three measures: searcher satisfaction; number of documents saved; and correctness of answer. There was no significant difference between NQE and QE in terms number of documents saved, or correctness of answer. However for satisfaction with search results, searchers were found to be more satisfied with their search results in QE (M=4.54; SD=1.96) than NQE (M=4.05; SD=2.15), although not quite significantly so, t(253) = -1.9, p=.058. So, we found only weak support for Hypothesis 2. Therefore, we investigated directly the relationship between query length and performance. In this analysis, significant correlations were found between satisfaction and mean query length, whether it is with (.137, p <.05) or without stop list (.136, p<.05). This seems to confirm a weaker version of hypothesis 2, that query length leads to better search outcome.

However, the data in table 9 show a negative significant relation between unique query length with stoplist and satisfaction (-.142, p <.05). This result is supported by analysis of correctness of response (table 10). Table 10 shows the relationship between correctness and unique query length, with stoplist and without. In both cases, more words in a search is significantly associated more strongly with incorrect answers than with correct answers ((253) = 2.78, p =.006; t(253) = 2.64, p=.009, respectively). These results support our interpretation of unique words in a search as a measure of search effort.

	correctness of a search	N	Mean (Standard Deviation)
Unique words in a search with stoplist	No	79	7.68 (3.13)
	Yes	176	6.56 (2.91)
Unique words in a search without stoplist	No	79	11.47 (7.15)
	Yes	176	9.23 (5.83)

Table 10. Unique words in query related to search correctness.

5 Discussion

We found support for hypothesis 3, that the MDD interface reduced interaction, and for hypothesis 1, that a system designed to make interaction more effective would lead to increased user satisfaction and increased performance. While subjects viewed significantly more full text documents in MDD than in SDD, they viewed significantly fewer documents in MDD than were seen in SDD. In terms of satisfaction and performance, subjects were significantly more satisfied in MDD than SDD, and saved significantly more documents in MDD than SDD. However, there was no difference in correctness of answers between the two treatments. Given that there were no differences in time and iterations between the two, these results indicate that because subjects were required to engage in more interaction in SDD than MDD, they had lower satisfaction, and decreased search effectiveness by one of two measures.

We found strong support for hypothesis 4, that the QE mode would lead to significantly longer queries than NQE. However, we found only weak support for hypothesis 2, that searchers in the QE mode would perform better than in the NQE mode. The somewhat weaker, related hypothesis, that longer queries would be associated with better performance, was only supported in part (with respect to mean query length being significantly associated with greater satisfaction with the search). However, in these circumstances, the number of iterations in a search might be considered an indirect measure of performance, if number of iterations is interpreted as effort needed to accomplish the task. The mean number of iterations per search (and standard deviation) for QE was 2.09 (1.35); for NQE, 2.64 (1.63). Results from a t-test indicate that subjects using QE had significantly fewer iterations than subjects using NQE, $t(253) = 2.98$, $p<.01$. Since there was no difference between correctness in the QE and NQE modes, we find further support for hypothesis 2, in that comparable results were achieved with less effort in QE than in NQE.

We found a significant negative relationship between unique query length and correctness of answer, as well as with satisfaction with a search. We speculate that these results might be explained by an interaction effect between unique query length and degree of interaction. As the number of iterations increases, the unique query length also increases. There is a strong correlation between iterations and unique query length (.44, p<.01). And the number of iterations is negatively correlated with satisfaction (-.53, p<.001). In other words, number of iterations might be the common cause variable that leads to both longer unique query length and dissatisfaction. Therefore, when query length is averaged, instead of uniquely counted, the positive correlation between query length and satisfaction is revealed, because the iteration variable is held more-or-less constant. An alternative explanation is that both number of iterations and unique query length are indicators of difficulty of the search topic. These issues deserve further investigation.

6 Conclusions

Our results support the idea that reducing the amount of interaction required of a searcher, therefore making interaction more effective, leads to a better experience for searchers, and that the MDD interface, which displays documents directly for judgment and use, rather than requiring users to judge on the basis of a surrogate and then follow links to the documents, does make interaction more effective in just this way. If our speculations about interaction effects between iterations and query length are correct, such a result would tend to support the general interaction hypothesis. This leads us to conclude that alternatives to the web browser-based paradigm of displaying search results as lists of links which need to be traversed to get to actual documents need to be further investigated, and that displays which afford direct access to documents are likely to be preferable in several ways to lists of links.

We found also that query length in a Web searching environment can be substantially and significantly enhanced by using a rather simple interface technique. Enhancing queries length in this way led to some increase in users' satisfaction with search results, and to significant increase in effectiveness of searches, considered as degree of effort required to achieve specific level of performance in the search task. Thus, we see support both for the possibility of increasing search length in interactive IR, and for the utility of doing so, at least in best-match search systems. These results suggest that IR systems need not be bound by the finding that queries presented to current systems are short, especially since most interfaces in current systems are designed to elicit short queries. In particular, the results suggest that much more thought should be given to how to elicit information problem descriptions in interactive IR systems. And, they suggest an alternative to pseudo-relevance feedback and similar techniques for enhancing query length, that may be more closely related to searcher needs than those techniques.

7 Acknowledgements

We wish to thank our colleagues who worked so hard with us in designing, scheduling and running the experiments: Amymarie Keller, Yuelin Li and William Voon, and our wonderful volunteer subjects. The work reported in this paper was supported in part by NSF Grant Number IIS 9911942. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation (NSF).

8 References

Belkin, N.J., Keller, A., Kelly, D., Perez-Carballo, J., Sikora, C., Sun, Y. (2001) Support for Question-Answering in Interactive Information Retrieval: Rutgers' TREC-9 Interactive Track Experience. In E.M. Voorhees and D.K. Harman (eds.) *The Ninth Text Retrieval Conference, TREC 9*. (463-475).

Belkin, N.J., Jeng, J., Keller, A., Kelly, D., Kim, J., Lee, H.-J., Tang, M.-C., Yuan, X.-J. (2002) Rutgers' TREC 2001 Interactive Track Experience. In E.M. Voorhees and D.K. Harman (eds.). *The Tenth Text Retrieval Conference, TREC 2001* (465-472). Washington, D.C.: GPO.