

Query Length in Interactive Information Retrieval

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, New Brunswick NJ USA

*Graduate School of Library and Information Studies, Queens College, CUNY, Flushing NY USA

+1 732 932 7500 *+1 718 997-3790

[belkin diane gkim jaykim hyukjinl muresan muhchyun xjyuan]@scils.rutgers.edu *ccool@qc.edu

ABSTRACT

Query length in best-match information retrieval (IR) systems is well known to be positively related to effectiveness in the IR task, when measured in experimental, non-interactive environments. However, in operational, interactive IR systems, query length is quite typically very short, on the order of two to three words. We report on a study which tested the effectiveness of a particular query elicitation technique in increasing initial searcher query length, and which tested the effectiveness of queries elicited using this technique, and the relationship in general between query length and search effectiveness in interactive IR. Results show that the specific technique results in longer queries than a standard query elicitation technique, that this technique is indeed usable, that the technique results in increased user satisfaction with the search, and that query length is positively correlated with user satisfaction with the search.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – *Query formulation, Search process*

General Terms

Performance, Design, Experimentation, Human Factors.

Keywords

Interactive information retrieval, query length, query effectiveness

1. INTRODUCTION

It is well understood and documented that, in best-match information retrieval (IR) systems, increased query length leads to increased performance, as measured by recall and precision in batch-mode experimental IR systems, and it is well documented that query length in typical operational interactive IR systems (in particular Web search engines, whether best- or exact-match) is rather short, typically between two and three words long (e.g. Jansen, Spink & Saracevic, 2000). This evident mismatch has led to a significant research effort at increasing query length. The most popular approach has been to automatically expand short queries without user intervention by various techniques, most often by some type of pseudo-relevance feedback (e.g.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Sigir'03, July 28–August 1, 2003, Toronto, Canada.

Copyright 2003 ACM 1-58113-646-3/03/0007...\$5.00.

Walker, et al., 1998), sometimes by thesaural or other type expansion (e.g. Milic-Frayling, *et al.*, 1998; Xu & Croft, 1996). Such studies have shown that these sorts of query expansion lead to increased performance, again as measured in experimental, batch-mode evaluation. There have also been a few attempts at increasing query length by encouraging searchers to enter longer queries in the first instance (e.g. Karlgren & Franzén, 1997; Belkin, et al., 2002). Although such studies have indicated that there seem to be relatively simple techniques that can encourage longer queries, they have not systematically investigated whether these queries actually result in better performance in interactive IR. Thus, the current situation is that, although there is evidence that longer queries perform better in non-interactive best-match IR systems, we can only infer that automatically-enhanced queries may perform better in interactive IR, and we have almost no evidence that longer queries obtained through searcher encouragement will lead to better performance in interactive IR. In this paper, we report on a study which explicitly investigated one technique designed to elicit longer than usual queries from searchers, and the effect of the technique, and of query length in general, on performance in interactive IR.

Belkin, et al. (2002) compared two query-elicitation modes, a query-entry line and a scrollable query-entry box, to determine whether the box mode (which allowed searchers to enter and see a complete query of five 40 character-long lines at a time) would result in longer queries than the line mode (which allowed searchers to see 50 characters of a query). They also compared two query types: queries constructed as lists of key-words and/or phrases, versus queries constructed as full sentences or questions. They found that the box mode led to somewhat longer queries, and that the full sentence or question type led to significantly longer queries. Although their study was not designed explicitly to evaluate the effect of query length on performance, they did find a consistent relationship between query length and one measure of performance in the task, completeness of answer. This was, however, only a descriptive finding, and not inferential.

Based on the results of Belkin, et al. (2002), we designed a study within the TREC 2002 Interactive Track (cf. Hersh & Over, 2003) to explicitly investigate the effect of method of query elicitation on query length, and the effect of query length on various measures of interactive IR performance. In particular, with respect to query length, we investigated two major research questions:

RQ1: What can make searchers' query length in interactive IR longer, and will searchers find such techniques acceptable and usable?

Within this research question, we posed the following hypothesis:

H1: 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.

This hypothesis was based on the results of Belkin, et al. (2002), on a survey of query elicitation techniques used in digital reference situations, and on the problem statement elicitation techniques described in Belkin, Oddy & Brooks (1982). From the data and ideas in these sources, we designed a query elicitation mode which asked people to describe their information problems, rather than to enter a query. This hypothesis was tested by comparing query length in this mode, with query length in a default, list of keywords and/or phrases elicitation mode.

We investigated the usability of the information problem elicitation mode through one “objective” measure, number of iterations, and one “subjective” measure, ease of starting a search.

Furthermore, we investigated the effect of several factors which might influence query length. One was type of information task that the searcher was engaged in (based on results such as those reported by Byström and Järvelin, 1995); another was the type of topic of the search; a third was familiarity of the searcher with the topic and task (based on results such as those reported by Kelly and Cool, 2002); and the fourth was perceived difficulty of the search.

RQ2: Does query length affect any measures of performance or effectiveness in the search task?

Within this research question we posed two hypotheses:

H2: A system which encourages long queries will lead to better performance in the search task than one which does not.

With this hypothesis, we were explicitly concerned with testing whether the information problem elicitation interface, given that it did encourage long queries, led to better performance than one which elicited keyword and/or phrase queries. Performance was measured by an “objective” measure, correctness of answer for the search task, and a “subjective” measure, searcher’s satisfaction with the search results. In addition, number of query iterations per search was considered an indicator of search effectiveness.

H3: Query length will be positively correlated with performance in the search task.

With this hypothesis, we were concerned with determining whether query length, regardless of query elicitation mode, affected performance in the search task. The same measures were used to test H3 as H2.

2. METHODS

2.1 Study Design

As this study was conducted within the TREC 2002 Interactive Track, we followed that Track’s guidelines.¹ These stipulated

that all participants in the track would test their systems in a comparative mode, using a common set of eight search topics, classified into two search task types, and four search topic types (examples of each task and topic type are in Appendix A). The general study design was within subject, with each subject searching a block of topics in one of two systems, and then a block in the other system. Each participating site was sent a pseudo-randomly generated list of 16 topic-block/system-order combinations by the Track organizers, which were to be applied to the minimum number of 16 subjects. This list guaranteed balance between system order and topic block order.

All participants in the Track used the identical database, the TREC 2002 Web Track database, which was a crawl of the .gov domain (see Craswell & Hawking, 2003 for a complete description), and the identical search engine, Panoptic.² Although Panoptic is basically a best-match search engine, which is the context in which we wished to test query length, for queries of four or fewer words, it changes to a coordination-level ranking algorithm, in which documents containing all n query terms are ranked before those containing $n-1$, and so on, disregarding the effect of term weights. Since this algorithm still results in a ranking close to that of a best-match system, we decided that query length could be investigated using Panoptic.

Our study investigated, in addition to query entry mode, method of display of results. In this paper, we do not report on the latter aspect. But since it affected our overall design, we indicate here how both aspects were studied within the same design. We studied these two aspects by running two complete sets of the basic Interactive Track design, in each of which the within-subject variable was display of results. We tested the effect of query elicitation mode with a between-subjects design, using a different mode for each of the two iterations of the basic study. Thus, query elicitation mode was tested with 32 subjects, 16 in each mode (cf. (Belkin, et al., 2003) for explicit details of the final design).

Our baseline query elicitation mode, called NQE, consisted of a scrollable query entry box of five 40-character lines, with the label above the box: “Query terms”. Subjects in this mode were instructed, during the introductory tutorial for this interface, to enter their queries as a list of key-words or phrases. Our experimental query elicitation mode, called QE, consisted of an identical box, but with the label above it saying: “Information problem description (the more you say, the better the results are likely to be)”. In the tutorial for this interface, subjects were told that they could enter their queries as multiple full sentences or questions.

All searches were performed using a Sun UltraSparc-III (440Mhz) with 512M memory and a 21 inch monitor. We wrote two basic interfaces as applications, connecting to Panoptic. They were implemented using Swing of Java 2 SDK, version 1.3. Screen shots are in Appendix B.

2.2 Conduct of the Study

All searching was done at the experimental site. When subjects arrived, they were asked to sign an informed consent form.³ They then completed a background questionnaire, eliciting

¹ <http://www-nlpir.nist.gov/projects/t11i/guidelines.html>

² <http://trec.panopticsearch.com/>

³ Rutgers IRB approval No. 01-407M

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 (pages) 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, ease of starting 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 compared 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>.

2.3 Measures and Definitions

We used the measures and variables defined in Table 1 to investigate the effect of the QE and NQE conditions, and of query length in general.

Table 1. Definition of measures and variables

Measure/Variable	Definition
Iteration	Each instance of issuing a query during a search on a single topic (number of iterations = number of queries in a search)
Mean Query Length	The average of the length in words of all queries in a search (4-word query + 5-word query + 3-word query = mql of 4)
Correctness of result	Whether a search resulted in saved pages which completely and correctly answered the search topic

2.4 Subjects

We had 32 volunteer subjects who were recruited largely from the student population at (deleted for reviewing purposes). Some were given credit for participating in the experiment and writing a brief description of their experience. Twenty-six (81%) were female and 6 (19%) male. Our subjects were most likely (47%) to be between 28-37 years of age, while their ages ranged overall from 18 to 57. 37% had completed a Masters degree, and

47% expected to complete one. All subjects were required to have some experience using Web search engines. 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.

3. RESULTS

3.1 Query Length

3.1.1 QE and query length

The first research question that we explored considered a technique for making searchers' query length in interactive IR longer, and if searchers would find such a technique usable. To test Hypothesis 1, we compared the mean query length per search in the QE mode to that of the NQE mode. Results from a t-test indicate that searchers using the QE interface entered significantly longer queries ($M=6.45$; $SD=3.00$) than those using NQE interface ($M=4.24$; $SD=1.26$), $t(253) = -7.67$, $p < .01$.

3.1.2 Usability of QE

To explore the usability of QE, we examined subjects' response to a Post-Search question and the mean number of iterations per search. We asked the subjects how easy it was to get started on the search on a 7 point scale to measure the "subjective" usability of NQE and QE. Results from a t-test indicate that there was no significant difference in experienced difficulty between NQE 4.83 (1.60) and QE 4.89 (1.52).

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, $t(253) = 2.98$, $p < .01$.

We also considered query length at each number of iterations. Table 2 displays mean query length at each number of iterations, and the numbers of searches with that number of iterations, for both QE and NQE. From these data, we see that, although mean query length is consistently longer for QE than NQE at each iteration number, the number of searches with N iterations decreases much more rapidly for QE than for NQE, as does query length up to iteration 6.

Table 2. Mean query length (standard deviation) at each iteration number.

Iteration		Mean Query Length (SD)		N	
NQE	QE	NQE	QE	NQE	QE
1	1	4.23 (1.42)	6.58 (3.12)	127	128
2	2	4.38 (1.55)	6.19 (2.52)	87	69
3	3	4.13 (1.62)	5.18 (2.02)	60	40
4	4	3.97 (1.59)	5.41 (2.27)	34	17
5	5	3.75 (1.88)	4.13 (1.73)	16	8
6	6	4.29 (1.50)	5.40 (1.67)	7	5
7	7	3.50 (1.00)	5.00 (0.00)	4	1
8		3.00 (0.00)		2	
Totals		4.19 (1.52)	6.09 (2.77)	337	268

3.1.3 Factors Affecting Query Length

As mentioned in Section 2, the information problems used in our study were classified along two dimensions: task type and topic type. We wished to see whether query length varied according to these types. Table 3 displays the mean query length and standard deviations according to task type and topic type. There are no significant differences between these means.

Table 3. Mean query length for task type and topic type.

	Mean	Standard Deviation
Task Type		
Website	5.45	2.58
Quantity	5.38	2.61
Topic Type		
Government Regulation	5.62	2.28
Health	5.91	2.67
Project	4.81	2.59
Travel	5.06	2.57

Subjects' knowledge of the information problems was determined by questions from the Pre- and Post-Search Questionnaires. Topic familiarity was assessed by a Pre-Search Questionnaire which asked subjects to indicate their familiarity with a topic on a 7-point scale, spanning from *Not at All* (1) to *Somewhat* (4) to *Extremely* (7). Perceived difficulty of the information problem was assessed by two Post-Search Questions which asked subjects to indicate how easy it was to do a search on the topic and if they had enough time to do an effective search. Subjects responded to these two questions using 7-points scales identical to the familiarity scale. Because of low numbers of observed values in some cells, we grouped subject responses to each of these three questions into three groups: Low (1-3), Medium (4) and High (5-7).

Table 4 displays the means and standard deviations for each of these measures. The difference between groups for familiarity is statistically significant, $F(2, 254) = 7.33$, $p < .01$, with subjects who are more familiar with the topic of the information problem entering longer queries than those subjects who are less familiar with the topic. For perceived difficulty, there is no significant difference between means for either measure.

Table 4. Mean query length (standard deviation) according to topic familiarity and perceived difficulty.

	Groups		
	Low	Medium	High
Familiarity	5.03 (2.30)	5.64 (2.50)	6.94 (3.63)
Perceived Difficulty			
Easy to search	5.29 (2.53)	5.13 (1.79)	5.47 (2.82)
Enough Time	5.19 (2.46)	4.80 (1.54)	5.61 (2.80)

3.2 Query Length and Performance

3.2.1 QE and Performance

The second research question that we explored asked if query length affects performance. Given that searchers entered significantly longer queries in QE than NQE, we compared directly searcher's task performance in each of these systems.

This was a test of Hypothesis 2. For performance, we considered the correctness of searchers' answers to the information problems and searchers' satisfaction with the system. Table 5 displays the distributions of correctly and incorrectly answered information problems for QE and NQE. Results from a Chi-Square test demonstrate that there is no significant difference in these distributions.

Table 5. Distribution of Correctness for QE and NQE.

		Correctness of Search		Total
		Incorrect	Correct	
System	QE	38	89	127
	NQE	41	87	128
Total		79	176	255

Satisfaction with a search was assessed by a Post-Search Questionnaire that asked searchers to indicate their satisfaction with the search results on a 7-point scale, spanning from *Not at All* (1) to *Somewhat* (4) to *Extremely* (7). 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 significantly so, $t(253) = -1.9$, $p=.058$.

We further considered mean number of iterations per search as an indirect measure of effectiveness, since fewer query iterations for equivalent results would be a more effective search. As reported earlier, the difference between mean number of iterations in QE ($M=2.04$; $SD=1.15$) and NQE ($M=2.46$; $SD=1.21$) is statistically significant, $t(253) = 2.88$, $p < .01$, with searchers using the NQE interface having significantly more iterations than those using the QE interface.

3.2.2 Query Length and Performance: Overall

To test Hypothesis 3, we asked whether mean query length overall was related to better performance, regardless of query elicitation mode. Overall, we found no statistically significant difference in mean query length for correct ($M=5.21$; $SD=2.49$) and incorrect answers ($M=5.34$; $SD=2.62$). To explore the differences in mean query length for each level of satisfaction, it was necessary to group the responses in the same way that responses to Familiarity and Perceived Difficulty were grouped, because of the low number of observations in some cells. The means and standard deviations for mean query length and satisfaction group are displayed in Table 6. Results from an ANOVA indicate that highly satisfied searchers entered significantly longer queries than unsatisfied searchers, $F(2, 254) = 4.17$, $p=.016$.

Table 6. Mean query length according to Satisfaction.

	Satisfaction		
	Low	Medium	High
Mean Query Length	4.68 (1.98)	5.67 (3.04)	5.64 (2.74)

We then considered the mean query length according to searches with N iterations. Table 7 displays these results. Here we note that the more iterations there are in a search (with the exception of 6- and 7-iteration searches), the shorter is the average mean query length. That longer mean query length is associated with fewer iterations suggests that similar performance can be accomplished with less effort with longer initial queries.

Table 7. Mean query length for searches with N iterations.

Iterations	Mean Query Length	Standard Deviation
1	5.90	3.26
2	5.41	2.33
3	5.04	1.73
4	4.80	1.40
5	3.48	1.18
6	4.98	1.62
7	5.14	1.27
8	4.31	0.80

4. DISCUSSION

Because of the well-known variability associated with differences in topics and types of information problems, we tested to see whether these factors influenced query length. There was no difference in query length according to tasks and topic types used in the TREC 2002 Interactive Track; we therefore conclude that the results reported here are general, at least with respect to this limited range of types. However, searchers' familiarity with the topic of the search did, as expected, affect query length (cf. Kelly & Cool, 2002). But since there was no relationship between topic familiarity and the two interface conditions and familiarity was randomly distributed in the subject population, we conclude that this factor does not affect our conclusions.

Hypothesis 1, that query length can be increased by encouragement in the interface, was strongly supported by our results. The QE interface resulted in significantly longer queries, even after stop word removal, over an interface which only asked for a query, even using a box-style query input facility which has been shown to increase query length in and of itself (Karlgrén & Franzén, 1997). Furthermore, we conclude that the QE interface was more usable than the NQE, because the number of iterations in the QE condition was significantly less than in NQE. We take this to mean that less cognitive effort was required in QE, to reach equivalent performance. Also, we saw that the QE condition led to a steeper decline in the number of iterations per search, and in query length, than NQE, supporting the conclusion that QE requires less effort, and therefore is more usable. Furthermore, there was no difference in ease of starting a search between QE and NQE, indicating that the relatively unfamiliar QE mode did not negatively impact its usability, as compared to the more familiar key-word mode.

Hypothesis 2, that the QE condition would lead to better performance than NQE, is supported by the strong difference in favor of QE in the searchers' satisfaction with the results of each search. Although there is no difference in performance between NQE and QE as measured by correctness of search, hypothesis 2 is also supported by the significantly fewer iterations required to achieve comparable correctness and increased satisfaction with results in QE. That query length decreases more rapidly with number of iterations in QE than in NQE also is supportive of Hypothesis 2.

Hypothesis 3, that longer queries lead to increased performance, regardless of query elicitation mode, is supported by the significant relationship between subjects' satisfaction with search results and mean query length, and also by the strong association of mean query length with number of iterations.

Since we can reasonably assume that subjects found the answers for their searches in their last iterations, it seems that longer initial queries led to comparable performance with decreased effort.

5. CONCLUSIONS

This study investigated the effectiveness and usability of a relatively simple interface technique for eliciting longer queries from searchers in a Web-based best-match IR system, and of the effectiveness of longer queries in general in this system. Our goal was two-fold. We wished to see if it was possible to elicit easily longer than usual queries from searchers; and, we wished to see if such queries were actually useful as input to best-match search engines. The query elicitation technique that was investigated was to have the following statement above a five-line, 40-character, scrollable query entry text box:

Information problem description (the more you say, the better the results are likely to be)

It was compared to the equivalent query entry box that had at the top, simply the words "Query terms".

Some things need to be said about how we measured search effectiveness, and our "objective" performance measure, *correctness* of reply to the search topic. There was no relationship between search query length and this measure, either with respect to query length in general, or with respect to the two query elicitation modes. However, about 69% of the search topics resulted in correct responses, which suggests that most of the topics were easy enough that the measure was not sensitive to differences in query elicitation mode or query length. It is also the case that those searches which resulted in incorrect answers were typically incomplete searches, in whatever query elicitation mode, suggesting that they were difficult enough that the ten minutes allotted for each search was the limiting factor. Thus, our major measures of search effectiveness are the searchers' subjective satisfaction with search results, and, the effort expended in completing a search.

The results of the experiment indicate that:

- this technique does indeed result in significantly longer queries than even a somewhat enhanced baseline technique, and is at least as usable, and perhaps more usable, than the baseline technique;
- the technique results in increased searcher satisfaction with search results, and "objective" performance in the task equivalent to the baseline, but with fewer query iterations;
- longer queries irrespective of query elicitation mode are significantly associated with increased searcher satisfaction with search results, and longer initial queries lead to equivalent "objective" results with fewer query iterations.

Thus, we conclude that our quite simple interface-based query elicitation technique results in significantly longer, and more useful searcher queries in a Web searching task than typical query elicitation, for a best-match information retrieval system. Furthermore, we conclude that longer searcher queries result in increased search effectiveness in general, indicating that more words from the searcher describing the person's information problem results in better interactive IR performance. Taken together, our results mean that getting longer queries from

searchers in a best-match Web searching environment is not only possible, but desirable and useful.

6. 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 authors and do not necessarily reflect the views of the National Science Foundation (NSF).

7. REFERENCES

- [1] Belkin, N.J., Cool, C., 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 & D.M. Harman (Eds.) *The tenth text retrieval conference, TREC 2001* (pp.465-472). Washington, D.C.: GPO.
- [2] Belkin, N.J., Cool, C., Kelly, D., Kim, G., Kim, J.-Y., Lee, H.-J., Muresan, G., Tang, M.-C., & Yuan, X.-J. (2003). Rutgers Interactive Track at TREC 2002. In E.M. Voorhees & D. M. Harman (Eds.). *The 2002 Text REtrieval Conference (TREC 2002)* (in press). Washington, D.C.: GPO.
- [3] Belkin, N.J., Oddy, R.N. & Brooks, H.M. (1982) ASK for information retrieval. Part 2. *Journal of Documentation*, 38(3), 145-164
- [4] Byström, K. & Järvelin, K. (1995). Task complexity affects information seeking and use. *Information Processing & Management*, 31(2), 191-213
- [5] Craswell, N. & Hawking, D. (2003) Overview of the TREC 2002 Web track. In E.M. Voorhees & D. M. Harman (Eds.). *The 2002 Text REtrieval Conference (TREC 2002)* (in press). Washington, D.C.: GPO.
- [6] Hersch, W. & Over, P. (2003) TREC 2002 interactive track report. In: E.M. Voorhees & D.K. Harman (Eds.) *The 2002 Text REtrieval Conference (TREC 2002)* (in press). Washington, D.C.: GPO.
- [7] Jansen, B.J., Spink, A. & Saracevic, T. (2000). Real life, real users and real needs: A study and analysis of users' queries on the Web. *Information Processing and Management*, 36(2), 207-227.
- [8] Karlgren, J. & Franzén, K. (1997) Verbosity and interface design. Retrieved on 17 January 2002 at: <http://www.ling.su.se/staff/franzen/irinterface.html>
- [9] Kelly, D. & Cool, C. (2002) The effects of topic familiarity on information search behavior. In: G. Marchionini & W. Hersch (Eds.) *JCDL 2002. Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries* (74-75). New York: ACM.
- [10] Milic-Frayling, N., Zhai, C., Tong, X., Jansen, P. & Evans, A (1998). Experiments in query optimization, the CLARIT system TREC-6 report. In: Voorhees, E.M. & Harman, D.K (Eds.) *The Sixth Text REtrieval Conference (TREC-6)* (pp. 415-454). Washington, D.C.: GPO.
- [11] Walker, S., Robertson, S.E., Boughanem, M., Jones, G.J.F., Sparck Jones, K. (1998) Okapi at TREC-6 Automatic ad hoc, VLC, routing, filtering and QSDR In: Voorhees, E.M. & Harman, D.K (Eds.) *The Sixth Text REtrieval Conference (TREC-6)* (pp 125-136). Washington, D.C.: GPO.
- [12] Xu, J. & Croft, W.B. (1996) Query expansion using local and global document analysis. In: H.P. Frei, D. Harman, P. Schäuble & R. Wilkinson (Eds.). *SIGIR '96. Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (4-11). New York: ACM.

APPENDIX A: Interactive Track Tasks and Topics

Task Types:

1. Find any N short answers to a question, to which there are multiple answers of the same type.
2. Find any N websites that meet the need specified in the task statement

Topic Types:

1. Looking for personal health information
2. Seeking guidance on US government laws, regulations, guidelines, policy
3. Making travel plans
4. Gathering material for a report on a given subject

Topics:

1. You are traveling from the Netherlands, and want to bring some typical food products as gifts for your friends. What are three kinds of food products from the Netherlands that you are not allowed to bring into the US? [Task 1; Government Regulation]

2. You are concerned with privacy issues related to electronic information and would like to know what laws have been passed by the US Congress regarding these issues. Identify three such laws. [Task 1; Government Regulation]

3. A friend has a private well which is the family's only source of drinking water. Locate a US publication, which contains guidelines for the maintenance of safe water standards for private well use. [Task 2; Health]

4. You are not sure about the safety of genetically engineered foods, and would like to find more information and research on this topic. Name four potential types of safety problems that have been raised. [Task 1; Health or Project]

5. You are interested in learning more about what measures the US government has taken since 2001 to prevent Mad-Cow Disease. Identify three such measures. [Task 1; Health or Project]

6. Name/find three research programs/projects that investigate the treatment/causes of dwarfism. [Task 1; Project]

7. You are planning a cycling expedition along the Silk Road in Central Asia. Find a website that is a good source information about health precautions should you take. [Task 2; Travel]

8. You are planning to travel to the northeast territories of India and wonder if there are any problems/restrictions for tourists. Find a website that is a good source of information about such problems/restrictions. [Task 2; Travel]

APPENDIX B. Screen Shots of the Two Interfaces.

NQE Query Elicitation Mode, 20 Retrieved Title Results Display Mode

The screenshot shows the Panoptic search interface. At the top, the title bar reads "Panoptic". Below it, the "Query Terms" field contains "TREC Conference Interactive". To the right, the "Documents Saved" list shows "[G10-64-0563675] TREC 2001 Interactive Track Home Page". The main search area displays a list of 20 results. The first result is highlighted in blue and reads: "1. 100 [G10-64-0563675] TREC 2001 Interactive Track Home Page". The description for this result states: "... e The TEXT Retrieval Conference TREC The TREC conference series is sponsored by the National Institute of Standards and Technology NIST with additional support from other U S government agencies The goal of the conference series is to encourage research ... comparing their results Attendance at TREC conferences is restricted to those researchers and developers who have performed the TREC retrieval tasks and to selected government personnel TREC Interactive Track TREC is organized along several tracks each of which addresses a ... high level goal of the Interactive Track is the investigation of searching as an interactive task by examining the process as well as the outcome The remainder of this page comprises information basic to the definition of the track for...". Below the description is the URL: "http://www.itl.nist.gov/iad94.02/projects/t101v - 1k - - Last Modified: No Date". The interface also features navigation buttons on the right side: "Backward", "Forward", and "Save". At the bottom left, there are "Start" and "Stop" buttons.

QE Query Elicitation Mode, Four Scrollable Document Results Display Mode

The screenshot displays a web-based interface for the TREC 2002 experiment. At the top, the title bar reads "Trec 2002". Below it, there is an "Information Problem" section with a text input area and "Search" and "Clear" buttons. To the right, a "Documents Saved" list shows "[G10-64-0563675] TREC 2001 Interactive Track Home Page".

The main content area displays four scrollable document results, each with a title and a brief description:

- TREC-2001 Interactive Track Home Page**
The TREC 2001 conference series is sponsored by the National Institute of Standards and Technology (NIST) with additional support from other U.S. government agencies. The goal of the conference series is to encourage research in information retrieval from large text applications by providing a large test collection, uniform scoring procedures, and a forum for organizations interested in comparing their results. Attendance at TREC conferences is restricted to those researchers and developers who...
- TREC-8 Interactive Track Home Page**
The TREC 8 conference series is sponsored by the National Institute of Standards and Technology (NIST) with additional support from other U.S. government agencies. The goal of the conference series is to encourage research in information retrieval from large text applications by providing a large test collection, uniform scoring procedures, and a forum for organizations interested in comparing their results. Attendance at TREC conferences is restricted to those researchers and developers who...
- TREC-9 Interactive Track Home Page**
The TREC 9 conference series is sponsored by the National Institute of Standards and Technology (NIST) with additional support from other U.S. government agencies. The goal of the conference series is to encourage research in information retrieval from large text applications by providing a large test collection, uniform scoring procedures, and a forum for organizations interested in comparing their results. Attendance at TREC conferences is restricted to those researchers and developers who...
- TREC-7 Interactive Track Home Page**
The TREC 7 conference series is co-sponsored by the National Institute of Standards and Technology (NIST) and the Information Technology Office of the Defense Advanced Research Projects Agency (DARPA) as part of the TIPSTER Text Program. The goal of the conference series is to encourage research in information retrieval from large text applications by providing a large test collection, uniform scoring procedures, and a forum for organizations interested in comparing their results. Attendance at TREC conferences is restricted to those researchers and developers who...

Navigation controls are located on the right side of the document list, including "Backward", "Forward", and "UnSave" buttons for the first document, and "Backward", "Forward", and "Save" buttons for the others. At the bottom left, there are "Start" and "Stop" buttons. At the bottom center, there are "Next 4" and "Prev 4" buttons.