New Tools and Old Habits: The Interactive Searching Behavior of Expert Online Searchers using INQUERY

Jürgen Koenemann* Richard Quatrain Colleen Cool

Nicholas J. Belkin†

School of Communication, Information, and Library Studies

Rutgers University

New Brunswick, New Jersey 08903

{koeneman,ccool,belkin}@cs.rutgers.edu

Abstract

We present data that describe the interactive searching behavior of ten searchers using the INQUERY retrieval engine in the context of the TREC-3 routing task. We discuss how these searchers with a strong background in the use of traditional online retrieval mechanisms adapted, after very limited training, to the use of a best-match, ranked-output, full-text retrieval mechanism.

1 Introduction

The development of IR system technology from slow, batch-mode, query submission to online query submission with rapid response generation by the IR system now allows for a highly interactive, iterative query development process. From experimental studies, it has been known for some time that best-match, ranked-output retrieval techniques are in general superior to exact-match systems in terms of recall and precision performance measures (e.g. [Belkin & Croft1987]). Furthermore, it seems that systems which allow queries to be put in unstructured form allow easier query formulation than those which require boolean structure, and are, at least for end-users, more effective [Frei & Qiu1993, Turtle1994]. Finally, it is quite clear that automatic relevance feedback, using term weighting and query expansion, either singly or in combination, significantly improves retrieval performance [Salton & Buckley1990].

The first users of these new retrieval mechanisms are likely to be those that have had access to and experience with traditional, boolean, set-based systems. We are interested in how these users will adapt to the new retrieval mechanisms. On the one hand, these users, typically professional online searchers, will have acquired an extensive repertoire of searching strategies. For example, searchers may be able to transfer their knowledge about the effects of adding terms to boolean queries in the form of conjunctions or disjunctions to a situation where they add terms within or outside the scope of proximity operators. On the other hand, these searchers are novices with respect to the retrieval mechanism and some of its interactive features. Furthermore, whereas some experiences with information retrieval mechanisms may be helpful because they can be directly transferred to the new situation, others may be not applicable or not effective in the new situation. We were interested to see the degree to which searchers attempted to transfer their existing search strategies such as a component-approach or use of boolean operators to the new environment and to what degree they would experiment with new tools such as relevance feedback made available to them.

We used the TREC-3 routing task context as a testbed to investigate these questions. We provided

*Department of Psychology

†to whom all correspondance should be addressed

ten experienced online searchers with a training manual that guided them through a sequence of hands-on exercises that familiarized them with the various features of the interface. After searchers completed the first part of the manual they performed one routing task (that is not included in our analysis). During the training phase, we provided some additional help when serious misconceptions persisted or when searchers experienced breakdowns from which they could not recover on their own. Each searcher then performed interactive searches for five TREC-3 topics (each lasting up to 20 minutes) without further help.

This paper focuses on the analysis of these expert online searchers with no prior INQUERY searching experience (the results reported as rutir1 in this volume). The comparative analysis of their performance and behavior with the experimenters' performance on the same task (reported as rutir2 in this volume) will be the focus of a forthcoming report.

The remainder of the paper is organized as follows. We first provide details on the searchers, experimental materials, the interface and retrieval mechanism, and the experimental procedure. Next we discuss summary performance and interaction data that demonstrate common trends among searchers. We continue with a discussion of some particular interactive searches that highlight some of the successes and failures users experienced. We conclude with some observations on the ways in which our searchers used the interactive facilities which were made available to them, and with some discussion of problems of evaluation of highly interactive IR systems.

2 A Case Study

2.1 Searchers

We recruited 10 professional searchers to participate in our study. The searchers had between 1 and 15 years experience with online search with traditional, boolean, set-based information retrieval mechanisms ($\overline{X}$ = 6.1 years) and their current position required the large majority to perform online searches at least a couple of times each week. All searchers had a Master's degree in Library Science and used computers and graphical user interfaces on a regular basis. Searchers had at best some experience with full-text search but almost no experience with ranked-output retrieval mechanisms (8 out 10 had no experience at all). Two out of ten searchers had previously seen a cursory, ten-minute demonstration of the INQUERY system with an interface similar to the one used in this study in the context of a classroom demonstration. None of the searchers had any prior hands-on experience with the INQUERY retrieval engine nor with the specific interface employed. Searchers volunteered their time and were not compensated for their efforts.

2.2 Materials

The Retrieval Mechanism We used the INQUERY retrieval engine (version 1.6.3), developed at the University of Massachusetts [Callan et al.1992]. The underlying mechanism of INQUERY is a Bayesian probabilistic inference network [Pearl1988] which provides strict rules for the computation of probabilistic belief values for each document in which the degree of belief that a document is relevant is based on the terms that are shared by the query and the document and the operators in the query formulation used to combine these terms. For a detailed description of the INQUERY system see [Callan et al.1992]. The standard stop-word and stemming algorithms were used. We restricted the query language to the use of the two proximity operators #uw and #n and the synonym operator #syn, i.e. none of the soft or strict boolean operators of INQUERY were available to the searchers. See the System Description and Timing of the system for a detailed description.

The Interface The interface to INQUERY was a modified version of the INQUERY TK user interface developed at the University of Massachusetts [Callan et al.1992] implemented in Tcl/Tk [Ousterhout1994] a widget-style interface development language for the X windowing system running under the UNIX operating system. INQUERY and the interface ran on a remote SUN workstation; we used a networked SPARC 4 with greyscale monitor, mouse, and keyboard for the searcher's interaction. See the appendix for a detailed description of the interface and all its features.

2.3 Procedure

The 50 TREC-3 routing topics (101–150) were categorized by hardness (median R-precision from TREC-2 results) into five groups. We created ten topic sets consisting of five topics each with about equal overall hardness by randomly selecting one topic from each hardness pool for each of the ten sets. Given these ten topic sets, we randomly assigned each searcher to one of the ten sets. The final queries generated by our ten searchers together constitute our **rutir1** submission for TREC-3.

Searchers were given a short general description of the TREC project and were given a consent form that outlined the experiment and the data collection. Upon signing of the form they were led to a screened-off area that was set up with the computer equipment and a video recording device. The experimenter gave a 2-minute introduction to INQUERY by listing its major features, namely ranked output (rather than set based retrieval), full text search (rather than title, abstract, or keyword search), the availability of automatic relevance feedback functionality, and the ability to process free form, unstructured queries. Searchers were next given a 14 page, ring-bound tutorial, used for a hands-on training session of about 90 minutes. The tutorial concluded with an example of their searching task for the experiment (see appendix for a detailed description of the tutorial and the searching task).

After a short break each of the ten searchers performed a series of five searches. The topics were determined by the random selection procedure described above and within one topic set topics were sorted by increasing hardness. That is, participants searched easier topics first and topics became harder with each search. Even though this fixed order for all searchers conflated hardness with learning, we nevertheless felt that searchers had to perform easier routing tasks first to enable at least some initial retrieval successes that motivated participants to continue with the experiment.

At the beginning of each search participants were handed the TREC topic description and a description of the routing task (see appendix). They were free to start interaction with the system at any time. During searches participants were allowed to refer back to the tutorial if they desired to do so but experimenters provided no further help during the trials except for restarting the system if a system failure had occurred. A search was completed when the searcher was satisfied with the result and saved a query as "final". If searchers had not completed their search after about 16 minutes they were reminded that each search was restricted to 20 minutes. After twenty minutes searchers were asked to complete the current iteration and to save what they felt was their best query as the final query.

Each search was followed by a short break during which the searcher filled out a short questionnaire regarding the particular search that had been completed (their familiarity with the topic, the ease of query construction, their confidence in the quality of the final query, and their rating of the severity of the 20 minute time constraint for this search). Searchers did not receive any feedback from the experimenter regarding the quality of the queries or regarding the strategies they employed except for the general encouragement that they were doing fine (independent of their actual performance).

After the last search and search questionnaire searchers filled out a background questionnaire and were debriefed about the goals of the study.

3 Results and Discussion

This section contains the major results from our study. We provide a quantitative and qualitative analysis of the data and present some explanations for the observed behavior. We first discuss the performance of searchers with respect to the TREC-3 routing task. The next section captures the structure and content of all queries issued throughout the interaction. The third section focuses on the interactive process itself and describes how searchers interacted with the various features of INQUERY and the interface and how they iteratively developed their final queries.

3.1 Performance

This section describes the performance of our searchers using the standard performance measures of precision at 100 retrieved documents and R-precision. We also looked at other performance measures like precision at 20 documents, precision at 1000 documents, and average 11-point precision. All measures are highly correlated and using any of them would not change our findings. We first present data on the performance of searchers on the **training** collection, followed by data on the performance on the TREC-3 **test** collection. We compare these performances to each other and compare the test results to the best, median, and worst results of TREC-3.

Training Collection Searchers performed their 5 searches each on the TREC disks 1&2 (the TREC-2 "ad hoc" database). During the interaction, searchers were given the relevance rating for each displayed document title (obtained from TREC-2). Although searchers were instructed to perform a routing task, most searchers focused on developing a good adhoc query for the training collection under the (implicit) assumption that good performance of the query on the training collection was not a sufficient but a necessary condition for a successful routing query. Thus, it makes sense to analyze the performance of searchers on the training collection.

The mean R-precision was just under 30% ($\overline{X} = 0.2905, s_D = .1661$). The overall precision at 100 documents retrieved was $\overline{X} = .3828$ with a standard deviation of $s_D = .2719$. The single worst performance was on topic 101 for which the final query retrieved no documents at all because of an error in the query formulation and topic 124 for which the R-precision was .0114 (prec.@100=.02). The best R-precision results were obtained for topics 134 (R-prec.=.6596, prec.@100=.8500) and 148 (R-prec.=.6443, prec.@100=.7900), the best precision at 100 retrieved documents was obtained for topics 130 (R-prec.=.6216, prec.@100=.90) and 135 (R-prec.=.5478, prec.@100=.93).

Precision on Training Collection				
	avg. R-prec		avg. prec. @100	
n=5	$\overline{X}$	s_D	$\overline{X}$	s_D
S01	0.3328	0.1582	0.4520	0.2899
S02	0.1664	0.1573	0.2360	0.2454
S03	0.1365	0.1159	0.2180	0.1934
S04	0.3070	0.1137	0.3800	0.2411
S05	0.3815	0.2702	0.4500	0.3522
S06	0.2396	0.1532	0.3540	0.2753
S07	0.3393	0.1485	0.4600	0.2941
S08	0.3848	0.1560	0.4720	0.2958
S09	0.3067	0.1354	0.3260	0.2280
S10	0.3100	0.1434	0.4800	0.3648
All	0.2905	0.1661	0.3828	0.2719

Table 1: Searcher differences for precision of final queries on training collection. Given are the average R-precision and average precision at 100 documents for each searcher.

There was considerable variation among searchers and topics. The average R-precision for searchers (see table 1) ranged from .1365 for searcher S03 to .3848 for searcher S08. R-precision correlated highly with precision at 100 retrieved documents, $r = .8926, df = 48, p < .001$.

There was a moderate but not significant negative correlation between years of online search experience and average R-precision, $r = -.44, df = 8, n.s.$, that is, searchers with more experience in traditional online searching tended to perform worse.

As described, topics were grouped by hardness (based on the median TREC-2 performance) and searchers searched for topics in increasing order of hardness. There was a significant trial effect for R-precision, $F(4,36) = 15.52, p < .001$, as well as for precision at 100 documents retrieved $F(4,36) = 32.47, p < .001$. Performance decreased with increasing hardness (see table 2).

This trend is difficult to interpret since the order of topic hardness was not systematically varied. Thus, possibly learning effects and other factors such as fatigue may have contributed to the result. However, learning seems not to be occurring since it should re-

Precision on Training Collection				
	avg. R-prec		avg. prec. @100	
n=10 Trial	$\overline{X}$	s_D	$\overline{X}$	s_D
1	0.4293	0.1601	0.7120	0.2044
2	0.4105	0.1232	0.5880	0.1561
3	0.2758	0.1519	0.3250	0.1613
4	0.1966	0.0711	0.1730	0.0672
5	0.1401	0.0919	0.1160	0.0829
All	0.2915	0.1661	0.3828	0.2719

Table 2: Trial differences for precision of final queries on training collection. Given are the average R-precision and average precision at 100 documents for each trial (Hardness TREC-2 category)

Precision on Test Collection				
	avg. R-prec		avg. prec. @100	
n=5	$\overline{X}$	s_D	$\overline{X}$	s_D
S01	0.3052	0.1812	0.3340	0.3217
S02	0.1304	0.0979	0.1160	0.1124
S03	0.1211	0.1560	0.1740	0.2825
S04	0.1549	0.1525	0.1860	0.2066
S05	0.4007	0.2609	0.3340	0.2723
S06	0.1723	0.1802	0.2300	0.2230
S07	0.3213	0.1835	0.4080	0.2491
S08	0.3825	0.1827	0.4420	0.3072
S09	0.2904	0.1876	0.3120	0.2419
S10	0.2699	0.1410	0.3620	0.3430
All	0.2549	0.1880	0.2898	0.2597

Table 3: Subject differences for precision of final queries on the test collection (TREC disk 3). Given are the average R-precision and average precision at 100 documents for each searcher

sult in better not in worse performance. One may argue that without learning results would have been even worse for harder topics. The data for other TREC-3 participants (see below) and our own rutir2 results suggest that indeed hardness is the major factor since there was no inter-trial learning involved in these systems and orders presumably varied from the one we employed; nevertheless the same pattern of decreasing performance can be found.

Test Collection The performance of final queries on the TREC-3 test collection (TREC disk 3) is the official performance evaluation criterion for TREC-3. It evaluates whether the queries developed by searchers in the context of the training collection were indeed suitable routing queries that produced relevant documents when applied to a different albeit similarly structured collection. In summary, the results from the training collection were replicated for test collection with somewhat lower overall results as one would expect.

The mean R-precision was $\overline{X} = 0.2549$ with a standard deviation of $s_D = 0.1880$. The mean precision at 100 documents retrieved was $\overline{X} = .2898$ with a standard deviation of $s_D = .2597$. Again, the single worst performance was on topic 101 for which the final query retrieved no documents at all because of a query error. Topics 104, 105, 106, and 116 also retrieved zero relevant documents. Similar results were obtained using precision at 100 retrieved documents as a measure of performance: in addition to the above topics, queries performed very poorly on topics 102, 103, 116, 127, 131, 144, and 149 (all with prec.@100 <.04. With the exception of topics 127 and 149 all these topics were characterized by a very small number of relevant documents. In evaluating these low numbers one needs to take into account that the theoretical maximum performance on these topics was very low as well since 5 topics had under 10 relevant documents in the collection and 23 topics had under 100 relevant documents.

The best R-precision results were obtained for topics 148 (R-prec.=.6952, prec.@100=.7700) and topic 135 (R-prec.=.5961, prec.@100=.84), the best precision at 100 retrieved documents was obtained for topics 150 (R-prec.=.4119, prec.@100=.88) and topic 135.

There was considerable variation among searchers and topics. The average R-precision for searchers (see table 3) ranged from .1211 for searcher S03 to .4007 for searcher S05. R-precision correlated highly with precision at 100 retrieved documents, $r = .826, df = 48, p < .0001$.

Precision on Test Collection				
	avg. R-prec		avg. prec. @100	
n=10 Trial	$\overline{X}$	s_D	$\overline{X}$	s_D
1	0.4503	0.1784	0.6460	0.2504
2	0.3579	0.1211	0.3530	0.1870
3	0.2015	0.1855	0.2410	0.1727
4	0.1409	0.1124	0.0940	0.0885
5	0.1238	0.0895	0.1150	0.0848
All	0.2549	0.1880	0.2898	0.2597

Table 4: Trial differences for precision of final queries on test collection (TREC disk 3). Given are the average R-precision and average precision at 100 documents for each trial (Hardness TREC-2 category)

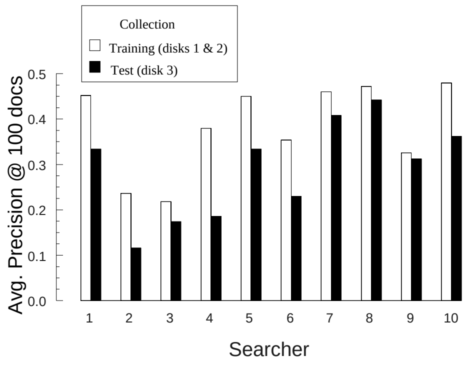

Figure 1: Searcher Averages

There was a significant effect of trial (hardness based on Trec-2 median R-Precision) on R-precision, $F(4,36) = 17.008, p < .001$, as well as for average precision at 100 retrieved documents, $F(4,36) = 23.975, p < .001$.

Differences between Training and Test Collection The routing task required searchers to develop a routing query in the retrieval context of a training collection (TREC disks 1& 2) but for an unknown test collection (TREC disk 3). It is, thus, interesting to compare the performance on the test collection to the performance on the training collection. The average R-precision dropped from .2905 to .2549 (a difference of .0356, $s_D = .0996$). This drop was significant, $F(1,49) = 6.379, p < .05$. The same holds for average precision at 100 retrieved documents: performance dropped significantly from .3828 to .2898 (diff= .093, $s_D = .171$), $F(1,49) = 14.75, p < .001$.

The correlation between precision at 100 retrieved documents for the test and the training collection was significant, $r = .794, p < .001$. The correlation for the R-precision measure was $r = .849, p < .001$.

The by far worst drop for R-precision (-.468) occurred for topic 104 (to 0.0), a topic with only 3 relevant documents in the test collection, whereas the best case was an improvement in R-precision by 0.1457 for topic 110 (from .4442 to .5899) probably due to the large number of relevant documents in the test collection (534). Overall, the performance dropped for 32 topics, remained the same for 1 topic, and increased for 17 topics. For 40 out of the 50 topics the difference was less than .1.

The worst precision drop at 100 retrieved documents (-.71) was for topic 134 (from .85 to .14) due to a small number (19) of relevant documents in the test collection, whereas the best case was an improvement in precision at 100 retrieved documents by 0.3 for topic 142 (from .37 to .67) probably due to the large number of relevant documents in the test collection (638). Overall, the performance dropped for 33 topics, remained the same for 1 topic, and increased for 16 topics. For 21 out of the 50 topics the difference was less than .1.

Figure 1 shows the average precision at 100 retrieved documents for each of the ten searchers on the training as well as the test collection. One searcher's queries (S09) did only minimally worse (.014 precision difference) on the test collection, and three searchers (S03, s07, and S08) had on average only slightly less good performance ($\leq .05$ precision difference) on the test collection. The remaining 5 searchers had larger drops in performance ($> .1$ precision difference). The worst drop occurred for searcher S04 whose queries showed an average drop in precision of .194.

We also looked at differences between the training and the test collection by trial (hardness category). Average precision at 100 retrieved documents dropped for all trials (hardness categories) with essentially no drop for the hardest topics (-.001) and the biggest drop occurring in trial 2 (-0.235). These differences were barely significant, $F(4, 36) = 2.667, p = .048 < .05$. However differences by trial were not significant for the R-precision measure, $F(4, 36) = 1.575, n.s.$ This is not surprising since the hardness categories were formed based on the TREC-2 performance on the training collection TREC disks 1&2 and thus are less applicable to the test collection TREC disk 3.

Comparing Rutir1 to General TREC-3 Performance The variation in indexing and retrieval mechanisms makes it difficult interpret the differences between our human searchers and the general performance of all systems that participated in TREC-3. Nevertheless, such a comparison allows us to judge the performance of our information system, comprised of INQUERY, the XINQUERY interface, and the human searcher as a unit and to detect general similarities and dissimilarities. Since R-precision data were not available at this point for all systems, we restrict our discussion to a comparison of precision as measured by the best, median, and worst precision at 100 retrieved documents for each topic. One should note as a proviso that the comparison against the best and the worst performance is somewhat problematic if one averages over a (sub)set of topics since one would not expect a single system to perform as well (or as badly) as the level which one obtains by combining the best (worst) performances of a large group of systems. A fairer comparison is probably the comparison with the median performance of all TREC-3 systems.

The average precision at 100 retrieved documents for our human searchers was .2898. The combination of best systems reached an average precision of .5342, a significant difference of .2444, $t(48) = 9.31, p < .001$. The median TREC-3 average precision was .4190, a significant difference of .1292, $t(48) = 12.22, p < .001$, and the combination of worst system performances yielded an average precision of 0.1782, a significant difference of .1116 in favor of our searchers, $t(48) = 10.25, P < .001$.

We correlated the performance of the human searchers (rutir1) with the best, median, and worst results for all TREC-3 systems. A high positive correlation would suggest that the human searcher would perform similarly on each topic (albeit on a different level) whereas a negative correlation would suggest that, e.g. topics that were most difficult for machine-based systems were much easier for human searchers and vice versa. The correlation of precision at 100 retrieved documents between rutir1 and the best TREC-3 results was $r = .802$ ($n = 50$ topics), the correlation of rutir1 precision at 100 retrieved documents with the median TREC-3 result was $r = 0.870$, and the correlation of rutir1 precision at 100 retrieved documents with the worst precision level in TREC-3 was $r = .829$; all significant at $p < .001$. This suggests that the difficulty of topics relative to each other was similar for human searchers and the machine approaches. One should note that the independence assumption for the above statistical test is violated, but given the size of the effects these violations will not change the results.

Given that relative searcher performance was influenced by topic difficulty based on the training collection, we may want to confirm our results for the test collection by comparing searchers not just among themselves (see above), but also against the general TREC-3 performance on the matching topic sets. Table 5 indicates the average difference in precision at 100 documents between the human searchers and the best, median, and worst TREC-3 results.

The searchers are sorted in order of decreasing average precision. Searcher performance is correlated with the best and median TREC-3 performance on the same topic set. Compared to the median performance, searcher S08 performed best (an average of -0.022 under the median), whereas searcher S02 performed worst (an average of -.234 under the median TREC-3 precision). We will return to these differences when we discuss the search processes and query characteristics associated with these searchers in more detail.

We discussed earlier that trial effects could be attributed either to learning or to the hardness of the

Comparison to TREC-3 Precision @100 on Test Collection				
n=5	$\overline{X}_{rutir1}$	$\overline{X}_{Best}$	$\overline{X}_{Median}$	$\overline{X}_{Worst}$
s08	0.4420	0.5840	0.4640	0.2660
s07	0.4080	0.6720	0.5380	0.2140
s10	0.3620	0.6260	0.4760	0.1800
s01	0.3340	0.6060	0.4420	0.1000
s05	0.3340	0.6180	0.4980	0.2480
s09	0.3120	0.5680	0.4400	0.2280
s06	0.2300	0.4360	0.3600	0.1800
s04	0.1860	0.3920	0.3040	0.1420
s03	0.1740	0.3780	0.3180	0.1600
s02	0.1160	0.4620	0.3500	0.0640
All	0.2898	0.5342	0.4190	0.1782

Table 5: Subject differences for precision of final queries on test collection. Given are the average precision at 100 documents for each searcher and the average performance for the same topics with the best, median, and worst TREC-3 results

topic. It is, thus, interesting to compare the average precision of our human searchers for the various hardness categories to the best, median, and worst precision values for the same topic sets achieved by the TREC-3 systems as a group. Figure 2 shows the average precision at 100 retrieved documents for the human searchers on the test and on the training collection and compares it to the best, median, and worst TREC-3 results on the test collection. One can see that there is a steady linear decline of precision with increasing hardness for all systems. Thus, it seems appropriate to attribute the drop in performance among the human searchers to the hardness of the topic and not to fatigue, negative transfer, or other problems. A by-hardness analysis of differences between the various systems shows no significant effect, except for the somewhat larger drop in performance in hardness category 4 for the human searchers. There is no trend that the difference between human searchers and automatic retrieval systems is larger or smaller based on topic difficulty.

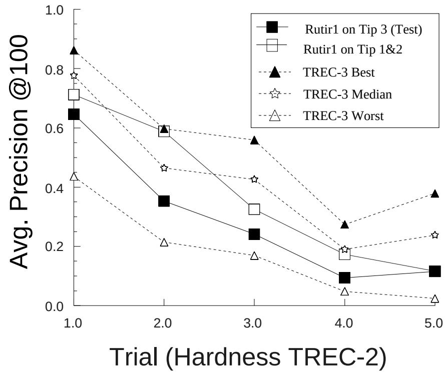

Figure 2: Performance Comparison

3.2 Query Characteristics

In our study we were interested in the nature of interactive query construction. This section describes the queries searchers entered throughout the interaction as well as the final queries saved as the routing queries that constituted our official TREC-3 submission. Sections 3.3.1 and 3.3.2 discuss characteristics of the interactions in which query construction took place.

Searchers entered a total of 338 queries, i.e. an average of $\overline{X} = 6.76$ queries per topic ($s_D = 4.00$). There were large individual differences: the average number of queries per topics ranged from 10 for searchers S02 and S04 to 3.8 queries per topic for searchers S06 and S10 but within searcher variation was also substantial.

There was a significant increase in the number of queries with increasing trials, $F(4, 36) = 4.67, p < .01$. Table 6 gives the averages for each trial (hardness category). Due to the confounding of trial and hardness alternative interpretations are possible and verbal protocols and interaction data suggest that both are at play: on the one hand, easier topics often led to satisfactory performance of early queries and searchers simply ended the interaction without having used up the allotted time. Secondly, searchers became more confident in the use of system features

Average Number of Queries		
n=10 Trial	$\overline{X}$	s_D
1	4.40	3.1340
2	5.90	2.9981
3	6.30	4.4485
4	7.20	3.5528
5	10.00	4.0825
all	6.7600	3.9978

Table 6: Trial (Hardness Trec-2) differences for average number of queries.

Query Characteristics				
Querytype	First	Interm.	Last	All
no rf				
no operators	11	30	0	41
w/ operators	39	152	22	213
Sum no rf	50	182	22	254
with rf				
no operators	n/a	10	4	14
w/ operators	n/a	46	24	70
Sum rf	0	56	28	84
Sum	50	238	50	338

Table 7: Query types for first, intermediate, and last (final) queries. Relevance feedback (rf) not possible in first query.

Avg. Number of Query Terms			
n=5	$\overline{X}$ All	$\overline{X}$ First	$\overline{X}$ Final
S01	6.20	3.2	10.2
S02	5.66	3.4	7.0
S03	9.06	5.2	13.2
S04	6.36	3.8	7.2
S05	4.26	2.8	5.2
S06	6.00	4.6	6.2
S07	4.93	3.2	6.6
S08	16.15	11.2	17.0
S09	8.00	8.4	8.6
S10	10.47	6.4	11.4
All	7.19	5.22	9.26

Table 8: Subject differences for the number of terms entered per query. Given are the average number of user terms entered for all queries, the first query, and the last query.

and were able to execute more queries in the constrained time span allowed for each topic.

We can characterize the queries by three factors: the use of terms in the query, the use of operators (synonym operator and proximity operators), and the use of relevance feedback to expand the user's formulated query and to reweight the terms. Searchers had been trained on the use of all three query formulation mechanisms. Users entered terms for all queries. Table 7 provides a breakdown of the use of operators and relevance feedback in the first, last, and intermediate queries.

The average number of terms in a query that was entered by the searcher was $\overline{X} = 7.19$. The majority of these words was taken from the topic descriptions. The minimum number of query terms was one and the maximum was 36 terms. For 42 of the searches users never entered more than 12 terms.

Table 8 shows the average number of user terms entered for all 338 queries, for the first query for each topic, and for the final query for each topic. The 50 first queries contained an average of $\overline{X} = 5.22$ terms whereas the the 50 final queries had an average of $\overline{X} = 9.26$ words entered by the user. One should note that the number of processed terms in the final queries is even higher in cases (28) where relevance feedback led to query expansion. The average number of processed terms in the final queries was $\overline{X} = 23.82$ terms.

The average number of user terms per query also varied with trial (topic hardness). The largest number of terms was found for trial 1 ($\overline{X} = 9.75$) with a steady decrease to trial 4 ($\overline{X} = 5.64$). Trial 5 had an average of $\overline{X} = 7.46$ user terms per query. Further analysis needs to be done to find an explanation for these data. It may be the case that an increased reliance on relevance feedback combined with increasing fatigue led to the decline in user-supplied terms. There was a weak but significant positive correlation between the number of user words and the precision at 100 retrieved documents, $r = .2855, p < .05$, that is, more user words led to better performance. However, the correlation between the actual number of processed terms used for retrieval (i.e. after relevance feedback led to query expansion) was not significant, $r = .209, p > .1$.

Searchers were introduced during the training process to the use of three non-boolean operators: fixed-order proximity (#n), unordered proximity (#uw), and synonym (#syn). The average number of operators per query was $\overline{X} = 2.54$. The average number of operators used in final queries was $\overline{X} = 3.34$, and the average number of operators in first queries was $\overline{X} = 1.94$. Searchers made almost equal use of all three operators (27.9% #n, 35.5% #uw, 36.6% #syn). There was no significant correlation between operator use in the final query and performance. The structure of queries was relatively simple; the vast majority of queries had at most one operator nested inside of another operator (typically a synonym operator inside of a proximity operator).

Table 9 details the use of operators by each searcher. Given are the average number of operators used in each query. Most searchers used on average between one and three operators in each query. Only searchers S03 and S08 had significantly higher use of operators. Searcher S08, the best searcher in our group, used four times as many unordered proximity operators compared to the other searchers and had a higher frequency of synonym use as well. Searcher S03 used many more phrase (fixed proximity) operators than others, and slightly more synonym operators as well.

Table 10 provides a breakdown of operator use by trial (hardness). Searchers used the most operators during their first search, in particular, #n operators. The differences between trials were not significant due to the variation within groups.

Finally, we looked more closely at the use of relevance feedback. We noted already above (see table 7) that over half of the final queries used relevance feedback and that a total of 84 queries were relevance feedback queries. Tables 11 and 12 show the use of relevance feedback on a by-trial (hardness) and by-searcher basis.

There was no systematic increase or decrease in relevance feedback usages over trials. Thus, neither

Operator Use in Query Construction				
n=5	$\overline{X}$all	$\overline{X}$ #n	$\overline{X}$#uw	$\overline{X}$#syn
s01	2.22	0.68	0.93	0.61
s02	1.28	0.02	0.88	0.38
s03	4.89	2.40	0.86	1.63
s04	1.18	0.38	0.00	0.80
s05	1.71	0.90	0.65	0.16
s06	2.53	1.37	0.58	0.58
s07	2.04	0.07	0.98	1.00
s08	6.81	0.27	4.19	2.35
s09	2.24	0.57	0.29	1.38
s10	2.95	1.68	0.11	1.16
All	2.54	0.71	0.90	0.93

Table 9: Searcher differences for the use of operators. Given are the average number of operators per query. The columns contain the averages for the number of all operators used, the use of the proximity operators #n and #uw, and for the use of the synonym operator

Operator Use in Query Construction				
n=10 Trial	$\overline{X}$all	$\overline{X}$ #n	$\overline{X}$#uw	$\overline{X}$#syn
1	3.43	1.39	1.07	0.98
2	3.14	0.51	1.39	1.24
3	1.86	0.35	0.81	0.70
4	2.63	0.81	0.79	1.03
5	2.18	0.69	0.68	0.81
All	2.54	0.71	0.90	0.93

Table 10: Trial (hardness) differences for the use of operators. Given are the average number of operators per query. The columns contain the averages for the number of all operators used and averages for each operator type

Use of Relevance Feedback			
n=10 Trial	rf possible	rf used	%
1	34	7	0.21
2	49	8	0.16
3	53	16	0.30
4	62	27	0.44
5	90	26	0.29
all	288	84	0.29

Table 11: Trial (Hardness) differences for the use of relevance feedback. Given are the number of queries in which rf use was possible (non-first queries), the number of queries in which rf was used, and the percentage of use for each trial.

Use of Relevance Feedback			
n=5	rf possible	rf used	%
S01	36	12	0.33
s02	45	0	0.00
s03	29	9	0.31
s04	45	25	0.56
s05	26	6	0.23
s06	14	6	0.43
s07	41	0	0.00
s08	21	8	0.38
s09	19	8	0.42
s10	14	10	0.71
all	288	84	0.29

Table 12: Subject differences for the use of relevance feedback. Given are the number of queries in which rf use was possible (non-first queries), the number of queries in which rf was used, and the percentage of use for each subject.

experience with the system nor the increased hardness of topics led to a significant increase in the use of relevance feedback. Searcher differences were pronounced: two searchers (S02, S07) did not use relevance feedback for any of their searches, whereas searchers S04 and S10 used relevance feedback for 56% and 71% respectively of the queries where it was possible to do so. A detailed analysis of the protocols will be needed to determine situations that trigger the use of relevance feedback by searchers. There was no correlation between the use of relevance feedback and performance.

3.3 The Interaction Process

The searchers in this project were presented with the task of developing a single best query for each of the topics given to them. Therefore, our discussion of interactive searching behavior is in fact a discussion of query formulation procedures. An interesting question which we address in this study concerns the ways in which experienced online searchers interact with a new retrieval system, INQUERY. The searchers in our study were experienced users of interactive systems, and they came to this experiment with a stock of routine query formulation procedures. However, some of these procedures, such as boolean set construction, do not easily map onto INQUERY. At the same time, INQUERY offered several interactive features, such as automatic relevance feedback, which were completely new to the searchers. The searchers in our study were faced with conceptual problems of understanding the topic itself, and in addition, the conceptual and procedural problems of learning to use the INQUERY system. An interesting research question concerns the extent to which searchers in this situation try to make the new system work to support their typical searching behaviors; or whether the searchers develop new searching behaviors, and new query formulation procedures as they interact with the new retrieval system.

3.3.1 Characteristics of the Interaction

We begin our discussion with some general characterizations of the interactive query formulation process, relating them, where relevant, to performance measures. The extent to which query reformulation actually took place in our searches, and the means and character of the changes, can be indicated by several parameters, including:

- changes between the initial queries and the final queries;
- number of iterations in a search;
- use of relevance feedback;
- use of features supporting manual reformulation;
- kinds of interaction with displayed output

Changes between initial and final queries can be measured relatively easily in terms of differences in numbers of query features. Tables 7 and 8 and the related discussion in section 3.2 show that there were substantial changes in these characteristics, with mean number of searcher-supplied words increasing from 5.22 to 9.26 (to 23.82 when terms supplied by relevance feedback are considered), and number of operators increasing from 1.94 to 3.34.

The number of iterations per search is a strong indicator of extent of interaction. Our definition of iteration, "that part of a search bounded by uses of the "Run Query" button" (first iteration beginning with starting the search, last ending with saving the final query), is equivalent to the concept of "cycle" as used in discussing interaction in traditional boolean retrieval systems (cf. [Harter1986]). The mean number of iterations per search, for all searchers, was 7.760. There was no significant difference in number of iterations by searchers, but the number of iterations increased significantly by trial (that is, by hardness), $F(4,36) = 4.607, p < .01$ (see table 13).

There was a medium but significant negative correlation between the number of iterations and precision at 100 retrieved documents, $r = -.435, p < .05$ and a medium correlation between the number of iterations and trial (hardness), $r = .438, p < .01$. One possible interpretation is that topic difficulty led to increased number of iterations but this increased effort did not pay off. To account for training and hardness, we computed the correlation between the number of iterations and the precision at 100 documents retrieved separately for each trial, i.e. for topics with similar difficulty. None of the correlations was significant (they ranged from .001 to -.498, but the trend was clearly in the same direction as the overall result: the more iterations searchers performed, the poorer was the result. One possible partial explanation is the fact that the submission of a syntactically wrong query ended one iteration. Thus, searchers who made syntax errors and maybe had more difficulties in general had more iterations.

Average Number of Iterations		
n=10 Trial	$\overline{X}$	s_D
1	5.50	3.064
2	6.90	2.998
3	7.30	4.449
4	8.10	3.604
5	11.00	4.083
all	7.7600	3.982

Table 13: Trial (Hardness Trec-2) differences for average number of iteration, per topic.

The methods by which searchers reformulate their queries can be characterized, in our setting, as automatic (using relevance feedback) or manual (adding/deleting terms and/or operators). Automatic relevance feedback was used in 29.2% of all of the queries in which it was possible. Interestingly, it was used in 58.3% of the final queries (28 out of the 48 searches which actually had some reformulation). In fifteen searches relevance feedback was never used. Two searchers accounted for ten of those searches, the remaining five distributed amongst four searchers. Two of the latter instances were searches with no reformulation of any kind. See tables 11 and 12 in section 3.2 for detailed data on use of relevance feedback, and its effect on performance.

Manual query reformulation, that is, addition, deletion or replacement of terms or operators by the user directly, was not used in seven searches. Five of those searches used automatic relevance feedback, and for two there was no reformulation at all.

From our data, it is clear that manual and automatic query reformulation are not mutually exclusive. Two searches involved no reformulation at all, five searches involved only automatic query reformulation, and 13 searches used only manual query reformulation. Thus 28 of the 50 searches used both automatic and manual query reformulation techniques.

The use of relevance feedback in our interface required the user to turn relevance feedback on, and to select at least one document as relevant (negative relevance feedback is not supported in INQUERY). Over all 50 searches, relevance feedback was turned on 54 times, and turned off 19 times (relevance feedback remains on in our interface to INQUERY, unless explicitly turned off, once it is invoked). The mean number of documents selected per iteration, over all searches in which relevance feedback was used, was 3.969, and the mean sum of documents marked as relevant, per search, was 16.120, with a range of 0 - 116. Over all searches, there were only 25 instances of deselection of documents.

Support for manual query reformulation was minor in our interface. The major such tool was the ability to save query statements, and subsequently combine them. We can get some indication of the general strategy of successive formulation of parts of a query, with subsequent combination of those parts (the most common query formulation pattern in traditional IR systems) through analysis of the use of the save and load query facilities. The mean use of the "save query" tool was 2.140, but since all searchers were required to use this feature at least once, to save the final query, the use of this tool for purposes other than saving the final query was 1.140. The total number of such uses was 57. The "load query" is perhaps the most direct indicator of this type of query reformulation strategy. 41 searches did not use this tool at all, and there were only 36 total occurrences, of which one searcher accounted for 22.

Finally, we can ask where the searchers found the evidence on which to base their query reformulation strategies, whether manual or automatic. In the automatic relevance feedback strategy, the most obvious source of evidence was the relevance judgments that were displayed with each document title. But for many retrieved documents, there were no relevance judgments, so the presumption is that decisions would be made on the basis of the other information in the summary window (belief value, document id, including the acronym indicating the database, and title), or on the basis of the texts of documents. For support of manual query reformulation, the highlighted terms in the texts of displayed documents could serve as an additional source for query reformulation. Accordingly, we consider here the number of titles, and the number of documents, that were viewed by the searchers.

In our interface, the default display as a result of a search was the summary window with five titles. In order to view a document, the user either had to double-click on the desired title, to double-click on the desired title as indicated in the bar graph (this never happened), to select a document through the scroll-bar next to the document display window, or to invoke the "next document" or "previous document" buttons below the document display window (this technique was effective only if a document was already displayed). The mean number of instances of viewing the full-text of a document, per search, was 6.620. Since our definition of iteration leads to one less chance for documents to be displayed than there are iterations in a search, there was an overall average of about one full-text display per iteration. That is, the total number of displayed documents, for all iterations, was 331, and the total number of iterations in which it was possible to see a document was 338.

There were clear differences between searchers in the display of the full-text of documents, as indicated in table 14. The searcher with the highest average (S09) looked at between 13 and 36 documents during a search, whereas the searcher with the lowest average never looked at more than one full text of a document during a single search. Absolute numbers for the other searchers ranged from 0 (searcher S10) to a maximum of 14 full texts viewed.

We identified three different conditions under which our searchers could view titles: the total number of titles displayed during a search or interaction; the number of unique titles displayed during an interaction; and, the number of unique titles displayed during an entire search. The first gives some indication of the overall extent of interaction in a search;

Viewing of Full Text of Documents		
n=5	$\overline{X}$	s_D
S01	5.6	3.21
S02	0.6	0.55
S03	6.6	5.55
S04	8.0	2.45
S05	7.0	2.00
S06	6.4	4.34
S07	8.8	3.03
S08	2.0	1.73
S09	18.4	9.91
S10	2.8	4.38
All	6.62	6.22

Table 14: Searcher differences for the number of times a full text of a document was displayed during one search (topic) and presumable viewed by a searcher.

the second, in comparison to the first, is an indication of searchers' returning to earlier displays; the second and third show how many titles searchers had available to them as sources for query reformulation; and, the third is an indication of differences in outcome of queries. We make the assumption, for general analysis, that the display of a title is equivalent to the searcher having attended to (viewed or seen) that title, and the further simplifying assumption that the number of documents without titles is relatively small (or, alternatively, that the searchers gain at least some information from the other information associated with the summary display, even when there is no title).

Table 15 shows the means per search, by searcher, of total titles seen for each of these viewing conditions.

Overall means for each condition, respectively, are 136.34 titles seen per search, 110.08 iteration-unique titles seen per search, and 67.2 unique titles seen per search. The first column (Table 15) indicates that there was a fairly high degree of interaction overall. The second column suggests that there was not too much "thrashing", or repetitive viewing within an iteration. The mean percentage of duplicate titles seen in an iteration was 19.26%, but eight out of the ten searchers had percentages below this mean, with one accounting for an average of 95 duplicate titles per search. The third column indicates that the queries were in general reasonably effective in finding new documents. Although the overall percentage of unique documents viewed is about 49%, seven of the searchers performed at between 58% and 64% "effectiveness", with one, with a large number of viewed documents, finding only about 27% new titles. There are clear differences in all three characteristics by searcher (table 15), but no significant differences by trial (hardness). It is of some interest to note that of the 6817 titles that were displayed, to all searchers, in all iterations, only 331 (somewhat less than 5%) were selected to be viewed in full.

Overall, these data indicate that the searchers engaged in substantial interaction in the formulation of their queries, using a wide variety of tools and facilities. Judging the effectiveness of the various forms of interaction, in terms of performance, is not easy, however. The problem is that it is not clear how the initial user queries should be interpreted. If one views them as initial starting points for subsequent interaction, then it makes no sense to compare the effectiveness of the first query with that of the final query.

3.3.2 Interactive Query Formulation

In this section we look at some of the query formulation procedures used by our searchers. We describe three sample searches, in order to illustrate three different query formulation strategies; one in which the searcher makes minimal or no use of new system features; one in which the searcher attempts to learn how to use system features over the course of a search, and one in which the searcher uses new system features right from the start.

Topic 122 - Minimal use of new features We chose to describe the query formulation procedures used by S07 in topic 122 as an example of minimal use of new system features. Topic 122 was chosen because it is familiar to all of the interactive systems participating in TREC-3; that is, all interactive sys-

Viewing of Document Titles						
	All Instances of viewing		Unique in Iteration		Unique in Search	
n=5	$\overline{X}$	s_D	$\overline{X}$	s_D	$\overline{X}$	s_D
S01	70.2	17.0	55.0	13.9	42.2	15.3
S02	53.2	24.7	45.2	18.5	31.0	11.4
S03	82.0	55.8	69.6	45.2	37.4	23.5
S04	319.0	94.4	224.0	47.6	85.4	32.9
S05	120.0	30.2	98.0	24.9	69.2	24.2
S06	58.0	30.5	48.0	23.6	33.8	14.5
S07	59.0	27.0	50.0	26.9	36.2	15.8
S08	319.2	129.6	267.8	127.9	157.0	91.5
S09	78.8	42.6	68.8	33.1	49.8	22.3
S10	204.0	102.4	174.4	86.0	130.0	63.6
All	136.3	118.2	110.08	93.4	67.2	55.6

Table 15: Searcher differences for the number of titles displayed in the summary window throughout the course of a single search (topic) and presumably seen by the searcher.

tems have been asked to provide a detailed description of the entire search process for this topic.

S07 began t122 with the following query formulation:

```
#uw50(cancer #syn(drug chemotherapy
treatment))
```

After 8 iterations, the final query was the following:

```
#uw100 (cancer #syn( research testing
evaluation))
#syn(cancer leukemia)
#uw50 (cancer #syn (drug chemotherapy
treatment))
```

In this example, the searcher constructed three separate queries, evaluated them individually and then together as one set. Table 16 gives the steps Searcher 07 went through in developing a final query for topic 122.

This searcher made no use of automatic relevance feedback, and minimal use of full text. For the most part, searcher 07 looked through the full text of documents to make sure that she hadn't missed anything, rather than using them to search for new concepts. Throughout the search, query construction centered around changes in scope rather than on changes in conceptual content of the queries. A total of seven words was used to construct the combined query set.

Searcher 07 used a building blocks approach to query formulation, in which three components were identified, terms selected to represent the components, and finally, the three components merged into one search statement. This is one of the most commonly used approaches to online searching, and it seems reasonable to think that searcher 07 was applying routine query formulation procedures to the INQUERY environment. We characterize the strategy used by searcher 7 as one of "fitting new tools to old habits."

Topic 116 - Partial use of new system features
The strategy employed by searcher 03 for topic 116 is characterized by a combination of manual and automatic query reformulation. S03 began with the following initial query: `#syn(drug generic)`

After 13 iterations, the final query entered by the user in the query window was the following:

```
#uw10( #syn(brand generic) #syn(drug
pharmaceutical))
```

The searcher decided to use relevance feedback, so the final query as produced through automatic rele-

Iteration	Step	Description
1	1	Formulate query 1 (see text)
	2	Run query
2	3	Display full text of the one relevant document in ranked set
	4	Save query 1 without modification
	5	Reset system and construct query #2:
		#uw100(cancer#syn(treatment research testing evaluation)
	6	Run query
3	7	Fixes syntax error (missing closing parenthesis)
	8	Run corrected query
4	9	Look at titles and relevance judgements of top five documents
		(One relevant document in list)
	10	Save query #2 without modification
	11	Reset system and construct query #3:
		#syn(cancer leukemia)
	12	Run query
5	13	Look at titles and relevance judgements of top five documents
		(One relevant document in list)
	14	Save query #3 without modification
	15	Load query #3; load query 1; load query 2
	16	Run combined query set
6	17	Look at titles and relevance judgements of top five documents
		(All have "?"as relevance judgement)
	18	Display full text of document ranked #2 in set
	19	Reset system
	20	Load query #2; delete term "treatment"
	21	Run modified query #2
7	22	Display full text of document 1 in ranked set (relevant)
	23	Display full text of document 4 in ranked set (?)
	24	Display full text of document 5 in ranked set (?)
	25	Save new query #2:
		#uw100(cancer#syn(research testing evaluation))
	26	Load query #3; load query #1
	27	Run combined query set
8	28	Scroll through retrieved documents,
		and look at 4 full-text documents
	29	Save combined query set as final

Table 16: Query Formulation Process for Topic 122 (Searcher 7)

vance feedback was the following:

```
#WSUM(1.000000 20.072113 #UW10( #SYN(
brand gener) #SYN( drug pharmaceut))
9.267074 vitarin 0.763216 bolar 5.181479
fda 0.489698 maxzid 0.420534 hemant
2.163789 dyazid 0.317706 hypertens 0.306859
shah 0.458595 pharmac)
```

In contrast to searcher 07, S03 made some use of new system features. Relevance feedback was used in three iterations; 14 full text documents were looked at in seven of the 13 iterations. A total of 12 search terms were selected by the searcher, and an additional four were added by the system, through relevance feedback. Searcher 03 went through the query formulation moves described in table 17

This example search illustrates some of the ways in which searcher 3 combined manual query expansion techniques (steps 4,8,12) with query expansion through automatic relevance feedback. In these steps, S03 repeatedly used a strategy of running a query, looking at retrieved texts, and then modifying the query. In step 21 the searcher uses relevance feedback to run a new query, and then manually adds a new term. The pattern of interaction in this search suggests that Searcher 03 did not have a complete understanding of how to use the new system, and in particular, how to use relevance feedback.

The example of topic 116 illustrates a search strategy which begins with routine manual query formulation procedures, and then attempts to incorporate some of the automatic query expansion techniques found in INQUERY. We characterize this strategy as an attempt to combine old and new models of systems and search strategies.

Topic 107 – Effective Use of New Features Topic 107 has been selected as an example of a one searcher's ability to effectively use features of INQUERY. Searcher 08 began this search with the following query formulation:

```
#uw300(#uw3(insider trading)japan)
```

After five iterations, the final query as entered in the query window was the following:

```
#uw300(#uw(insider trading) japan#syn(law
regulations guidelines legislation
#uw3(self regulation)))
```

The searcher decided to use relevance feedback, so the final query as produced through automatic relevance feedback was the following 120 term query:

```
#WSUM(1.000000 2.658688 #UW300( #UW3(
insid trade) japan #SYN( law regul guidelin
legisl #UW3( self regul))) 2.621457 profit
2.841300 corpor 2.116411 osaka 2.873363
loss 9.742378 #foreigncountry 1.359818
unlist 9.114288 exchang 3.300622 disclosur
6.330032 compan 2.571121 client 1.268526
diet 1.321397 shimakura 2.059871 broker
[...91 terms left out...]
1.226791 know 11.650723 ministr 3.683496
foreign 4.168829 aide 11.580430 tokyo
1.515035 kiich 1.579114 question)
```

Searcher 8 used the scroll bar to move through the document list, and looked at a total of 268 unique titles over the course of four iterations. He consulted only one full text document. Relevance feedback was turned on during two iterations, and 85 documents were selected by S08 as relevant. The query formulation process for topic 107 (Searcher 8) is depicted in table 18.

Summary These example interactions illustrate three different ways in which experienced online searchers in our study adapted to a new retrieval system, INQUERY. Two of the searchers were able to adapt to the new retrieval mechanism, one by using new system features in ways which supported her routine searching strategies, and the other by developing new searching behaviors which matched the capabilities of the system. A third searcher tried to use both traditional query formulation strategies and automatic features of the new system, with less success.

The data in our study do not enable us to say what predicts successful adaptation to a new retrieval system environment. From the descriptive evidence we have presented, it seems as if Searcher 7 understood how to use specific features of INQUERY, such as proximity and synonym operators, in ways which supported her routine searching strategies in boolean systems. This searcher made the new system work for her; or we might say that she applied new tools to

Iteration	Step	Description
1	1	Formulate query #1 (above)
	2	Run query
2	3	Look at top 25 titles and their relevance judgements
	4	Delete query 1; replace with:
		#uw200 (#syn(nomenclature #3(chemical formula)
		#3(brand name) #3(generic name)))
	5	Run query 2
3	6	Look at top five titles and relevance judgements
	7	Display full text of one relevant document
	8	Add one term to query 2, to form new query 3
	9	Run query 3
4	10	Correct spelling "drugnomenclatura" to "drug nomenclatura"
	11	Run corrected query
5	12	Look at top five titles and their relevance judgements
	13	Display full text of one relevant document
	14	Delete query 3 and replace with:
		#uw10(#syn(brand generic)#syn(drug or pharmaceutic))
	15	Run query
6	16	Look at top fifteen documents
	17	Display four full text documents
	18	Automatic relevance turned on; one document selected
	19	Run query with relevance feedback
7	20	Look at top five titles and their relevance judgements
	21	Display 2 full text documents
	22	Reformulate query:
		#uw10(#syn(brand generic chem)#syn(drug or pharmaceutic)
	23	Run Query
8	24	Relevance feedback on, add one document
	25	Run query
9	26	Relevance feedback on, delete 1 document, add 1 document
	27	Relevance feedback off
	28	Fifteen titles and relevance judgements looked at
	29	One full text document displayed
	30	Run Query
10	31	Relevance feedback off, 1 document selected
	32	Run query
11	33	Relevance feedback off, 1 additional document selected
12	34	Relevance feedback turned on, query run
13	35	Save last query as final query

Table 17: Query Formulation Process for Topic 116 (Searcher 3)

Iteration	Step	Description
1	1	Formulate query (above)
	2	Run Query
2	3	Relevance feedback turned on
	4	Scroll through the document list
		look at 90 titles (50 unique)
	5	Mark 31 documents relevant for relevance feedback
	6	Retrieve 1 full text document
	7	Run query
3	8	Relevance feedback turned off; save query
	9	Look at 65 titles; 41 of which are unique
	10	Modify query by adding 6 new terms. New query:
		#uw300(#uw(insider trading) japan#syn(law regulations
		guidelines legislation #uw3(self regulation)))
	11	Run query
4	12	Relevance feedback turned on; 23 documents added
	13	Run query (step 10) with relevance feedback on
5	13	save query as final query

Table 18: Query Formulation process for topic 107 (Searcher 8

old habits. Searcher 8, on the other hand, appears to have learned new searching behaviors through his interaction with INQUERY. Rather than adapting the new system to his old behaviors, Searcher 8 changed behaviors to fit the new system environment. S08 appears to have understood the concept of automatic relevance feedback, and the procedures for using it in INQUERY. Our last searcher, S03, did not change routine searching behaviors to fit the new system environment, and she did not use features of INQUERY to support her old habits. It appears as if searcher 3 did not have a well formed model of the new system, and her attempts to use it were based upon trial and error experiences.

Based upon this descriptive evidence, we suggest that a future research goal might be the investigation of factors which constrain or facilitate the adoption of effective new system models.

4 Conclusions

One conclusion that we can draw immediately from our data is that searchers in this study have been able to use the interactive features, including relevance feedback, offered by the INQUERY IR model and the interface which we provided. Furthermore, it appears that the combination of automatic and manual query reformulation can be quite effective, if used appropriately. From our few extended examples, we have some reason to believe that appropriate use of the reformulation and interaction facilities or capabilities might depend upon the ability of the searcher to develop an appropriate mental model of the IR system.

One somewhat problematic result was the seemingly negative effect of extent of interaction, as measured by number of iterations, on retrieval performance. Although this effect is conflated, in our study, with the hardness of the topic being searched, there still seems to be some possibly significant effect. This remains for us an open question, but one which might possibly be answered by analysis of the thinking aloud protocols associated with the searches, which we are now beginning.

Perhaps the most immediately striking result of our study is the consistently poor performance of the Rutir1 final queries, with respect to the median performance of all TREC-3 systems. Since this is also a result for all of the other interactive systems in TREC-3, it seems worthy of some comment.

Our searchers basically took words from the topic descriptions as the source for their queries. This is also precisely what the automatic systems did. Perhaps because of a lack of familiarity with the INQUERY system, or with the relationships between language in full-text documents and query effectiveness, or because of lack of good tools to support reformulation, our searchers tended not to exhibit much of what one might expect to be the potential advantage of human searchers in the interactive environment, namely, exploration leading to new ways to formulate the information problem.

The automatic systems, on the other hand, started with very rich queries (that is, basically, all of the terms in the long topic descriptions), and used them precisely in ways that were tailored to their various retrieval techniques. Since the TREC topics are unusually long, in terms of what one might ordinarily expect as input to a system from a user, and in terms of our searchers' starting queries, this alone might explain some of the difference in performance between automatic and interactive systems. It might indeed be possible to interpret the original TREC topics as the result of some substantial user interaction in an IR system (indeed, at least some topics were produced in precisely this way), and thus the automatic systems are working on top of a previous interaction. But this reasoning is all highly speculative, and the issue clearly needs to be addressed explicitly in further studies.

Another issue that needs to be more directly addressed is the relationship of the searcher to the information problem for which the search is being conducted. In an automatic system, this may not have a terribly strong effect on performance, but it seems reasonable to suppose that searching proceeds differently for persons who are doing searching for their own problems, than for persons who are doing delegated searches (e.g. see [Shenouda1990] and [Spink1994]). Thus, the entire experimental framework of TREC may be working against appropriate evaluation of end-user interactive searching.

This brings us to our final comment. That is, the experience of doing a study of an interactive system with human searchers in the TREC context leads us ever more strongly to the conclusion that we need to develop new evaluation measures and methods for interactive IR for end users. One such change might be to have searchers doing ad hoc style searches for their own information problems, or problems which are salient to them in other ways. Another might be to address in some way the idea that queries in such systems are being progressively formulated, rather than progressively better specified. And something that in our case can be done immediately, is to make use of the searchers' comments in the thinking aloud task in order to understand better their problems and their behaviors. This last point is one that we are now taking up, in order to address many of the open issues that have been raised in these concluding comments.

Whatever changes might eventually result in methods for studying interactive IR, it seems clear to us that the experience of our interactive systems studies in TREC-3 has been highly beneficial in several ways. It has given us an early, and important look into how human searchers interact in best-match IR systems with relevance feedback, the context in which we can expect most IR to take place in the future. It has in particular suggested several issues that will need to be addressed in order to make these systems more relevant to this context, and in order to make them more supportive of users. And, it gives us some hope that we will be able to design IR systems which directly support effective human interaction.

Acknowledgements

We owe special thanks to the searchers who volunteered to participate in this study, and to Jamie Callan, Bruce Croft and Steve Harding of the Center for Intelligent Information Retrieval at the University of Massachusetts for their unstinting support of our use of INQUERY, and for running the final queries on the test collection.

References

[Belkin & Croft1987] Belkin, N. J. & Croft, W. B. (1987). Retrieval techniques. In Williams, M. E., editor, *ARIST*, chapter 4, pages 109–145. Elsevier.

[Callan et al.1992] Callan, J. P., Croft, W. B., & Harding, S. M. (1992). The inquery retrieval system. In *DEXA 3: Proceedings of the Third International Conference on Database and Expert Systems Applications*, pages 83–87. Berlin, Springer Verlag.

[Frei & Qiu1993] Frei, H. & Qiu, Y. (1993). Effectiveness of weighted searching in an operational ir environment. In *Information retrieval 93: von der Modellierung zur Anwendung; Proceedings der 1. Tagung Information Retrieval '93*, pages 41–54. Konstanz, Universitaetsverlag Konstanz.

[Harter1986] Harter, S. P. (1986). *Online Information Retrieval.* San Diego, CA, Academic Press.

[Ousterhout1994] Ousterhout, J. K. (1994). *Tcl and the Tk Toolkit.* Reading, MA, Addison-Wesley.

[Pearl1988] Pearl, J. (1988). *Probabilistic reasoning in Intelligent Systems: Networks of plausible Inference.* San Mateo, CA, Morgan Kaufmann.

[Salton & Buckley1990] Salton, G. & Buckley, C. (1990). Improving retrieval performance by relevance feedback. *JASIS*, 41(4):288–297.

[Shenouda1990] Shenouda, W. (1990). Online bibliographic searching: how end-users modify their search strategies. In *Information in the year 2000: from research to applications. Proceedings of the 53rd Annual Meeting of the American Society for Information Science*, pages 117–128. Learned Information, Inc.

[Spink1994] Spink, A. (1994). Term relevance feedback and query expansion: relation to design. In *SIGIR '94. Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 81–90. London, Springer Verlag.

[Turtle1994] Turtle, H. (1994). Natural language vs. boolean query evaluation: A comparison of retrieval performance. In *SIGIR'94. Proc. of the Seventeenth Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 212–220. London, Springer Verlag.

APPENDIX: rutir1 INTERATIVE SYSTEM DESCRIPTION

I. System description

I.1 Screen dump of "typical" screen

Figure 3 shows a typical state of the Rutgers XINQUERY interface during a search. Individual features are discussed below.

I.2 Usable features of the interface

I.2.a General

In our experiments we used the INQUERY retrieval engine (version 1.6.3), developed at the University of Massachusetts (Turtle Croft, 1991). We used the INQUERY TK user interface, XINQUERY, developed at the University of Massachusetts with a number of modifications. Most actions are carried out through keyboard-based and mouse-based interaction with the main Tkinq window (I.1) which is comprised of four subwindows and a number of interaction devices. Below we describe the various funcionalities of the interface, and specify the particular feature whose use we have monitored, in **UPPERCASE.**

I.2.b Query formulation features:

On the top right is a query window with limited text editor functionality. Searchers use this window to enter and edit their queries. Queries can extend over multiple lines and a scrollbar allows scrolling through complex queries that extend beyond the size of the query window.

SAVE QUERY, LOAD QUERY Buttons above the query window can be used to save a query to file or to load a previously entered query into the query window. In either case, a dialog window with a list of current filenames and default names will appear. The user can select an existing name or enter a new name. The text of a loaded file will be appended to the current contents of the query window, allowing for the combination of previously stored query segments. Clicking on the "Reset ALL" button will clear the query window as well as all other information from previously executed queries.

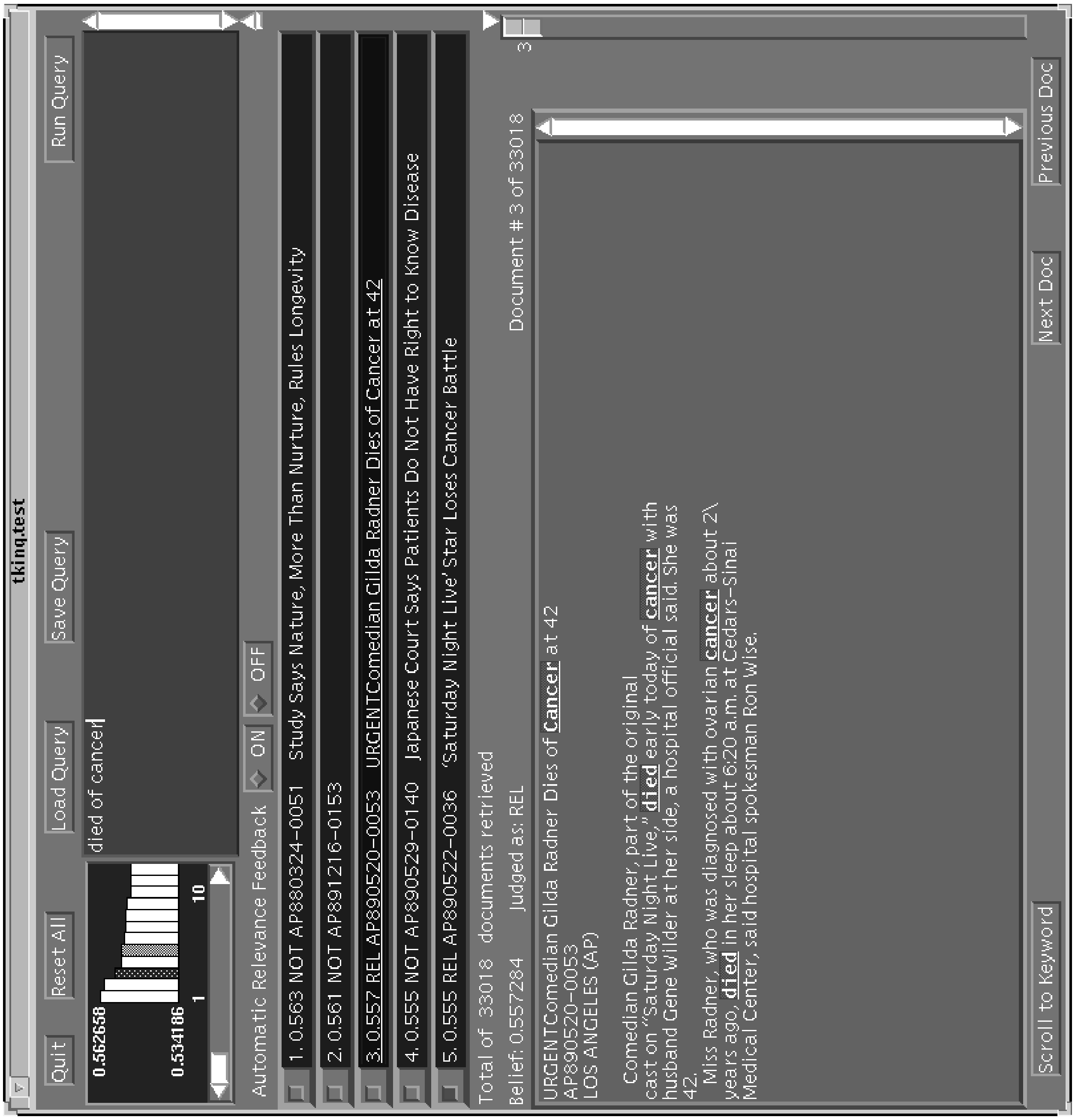

Figure 3: The Rutgers XINQUERY user interface

To execute a query the searcher enters (or loads) a query and clicks on the "Run Query" button above the query window. If the query is syntactically incorrect a warning message and a separate error window will pop-up. The error window will give extremely limited feedback on the type and location of a syntax error (if at all). These windows remain on the screen until the users dismisses them explicitly. If the query is syntactically correct it will be processed further by the INQUERY system. Status messages will appear that indicate that a query is processed, the number of terms that are evaluated, and that the resulting list of retrieved documents is sorted. Upon completion, the summary window (see below) will be updated with the results from the current query.

I.2.c. Search results

The window in the middle is a summary window that gives summary information on five documents retrieved by the system in response to a query. It displays the rank of a document, the belief value assigned by the system, a relevance rating of the document as assigned by the TREC raters (if available), its external document ID, and the title of a document.

SCROLLBAR IN SUMMARY Users can scroll through the ranked list of retrieved documents by means of the scrollbar to the right of the summary window.

SCROLLBAR IN DOCUMENT WINDOW Searchers can also use the scrollbar on the lower right to change the summary window. As the user is dragging the scroller, feedback about the rank of the about-to-be-selected document (that will appear on the top of the summary window) is provided. Both scrollbars are very sensitive if large numbers of documents have been retrieved and direct scrolling to a particular rank in the list is not (easily) possible; triangular buttons at the respective ends of the summary window scrollbar need to be used to change the ranks displayed in the summary window by increments/decrements of five. An indicator under the summary window gives the total number of documents retrieved by the current query.

DOCUMENT SELECTED/UNSELECTED (AS RELEVANT) To the left of the titles are toggle buttons that allow a searcher to mark individual (or groups of) documents.

RELEVANCE FEEDBACK ON/OFF Directly above the summary window is the automatic relevance feedback toggle. If the toggle is "off" (the default), markers only serve the orientation of the user. If automatic relevance feedback is turned "on", every marked document is used as a source in the computation of a relevance feedback query.

DOUBLE-CLICK ON TITLE The bottom half of the window consists of the document window. Users can select a document by double-clicking on its title. As a result, the full text of the document is displayed in the window. A scrollbar can be used to scroll through the document if the document exceeds the given height of the window. Each word in the full text of the selected document who's stem matches a stem of a term in the query is highlighted in the document.

KEYWORD The "scroll to Keyword" button under the document allows a searcher to scroll to the next highlighted term in the document (if any).

NEXT DOCUMENT, PREVIOUS DOCUMENT A searcher can directly display the full text of the next or previous document by clicking on the respective buttons under the document window.

BARGRAPH On the top left is a bar graph window that depicts the belief values assigned to the retrieved documents in graphical form. Each graph depicts a range of 100 documents (by default the top 100 documents). Searchers can use the scrollbar under the bargraph window to bring other portions of these 100 documents into view. The bargraph will be updated to depict a different range of the ranked list if a document outside the current range is selected. The two numeric values represent the belief values assigned to the highest and the lowest ranked documents in the current graph respectively. Each bar represents one document and the height of the bar corresponds to its belief value; heights are scaled relative to the belief value assigned to the document ranked lowest in the current graph, i.e. the rightmost document in the bar graph has a bar height of 0. The bar associated with the currently selected document is hatched. If a document is marked by the searcher (with automatic relevance feedback being either on or off) will appear grayed. A searcher can bring a document into full text view by double-clicking on the associated bar in the bar graph.

A user quits the inquery system by clicking on the top left "Quit" button.

I.3 Style of Interface:

Graphical User Interface (GUI) (see also section I.2)

II Experimental conditions

II.1 Searcher Characteristics

a. **Number of searchers in experiment**: 10

b. **Number of searchers per topic**: 1

c. **Age/age group of searchers**.

Age	n (searchers)
Under 21	0
21-30	3
31-40	3
41-50	4
Over 50	0

d. **IR searching experience of searchers**

We measured four dimensions of searching experience in this study:

- **Years of searching experience**: Mean: 6.1 years, Range: 1–15

- **Frequency of online searching**

Searching Frequency	n (searchers)
More than twice a day	3
Once or twice a day	1
More than twice a week	0
One or two times a week	4
More than twice a month	0
One or two times a month	1
Less than monthly	1

- **Experience searching full-text databases**

Experience searching full-text databases was measured by responses to a five point Likert scale, in which 1= None at all, 3= Some and 5= A great deal.

Full-text Experience	n(searchers)
1 (None at all)	2
2	2
3 (Some)	5
4	0
5 (A great deal)	1

- **Experience searching in ranked-output information retrieval systems**

The same five point Likert scale was used to measure searchers' experience using ranked-output retrieval systems, as full-text systems.

Ranked-output Experience	n (searchers)
1 (None at all)	8
2	1
3 (Some)	1
4	0
5 (A great deal)	0

e. **Educational level of searchers**

Searchers were asked to list all of their college or university degrees.

All of the searchers had, or expected, the Masters degree in Library Science. Four of our searchers had graduate degrees in addition to the MLS. Two searchers had a JD; two had, or expected to obtain, a PhD; and two searchers had an additional masters degree.

f. **Undergraduate major of searchers**

Major	**n**
English	1
History	1
Psychology	2
Political Sci	1
Zoology	1
Math	1
Chemistry	1

Two of the searchers did not list their undergraduate major. One of these searchers had a JD; the other searcher had a PhD in Oceanography.

g. **Experience/familiarity with subject of topic**

After completing each search topic, subjects were asked about their familiarity or experience with the subject matter covered in that topic. A five point Likert scale was used to record responses to the question, "Is this topic related to things you normally search on?" where 1=Not at all; 3= Somewhat; and 5=Very much. Overall, the Trec topics were not routine subject matter for the subjects who searched them. The mean familiarity value for all 50 topics, across all searchers is 2.0. The distribution of topics across the scale is the following:

Familiarity	**n (topics)**
1 (Not at all)	20
2	13
3 (Somewhat)	12
4	3
5 (Very much)	2

Work affiliation:	**n (searchers)**
Academic Library	6
Industrial/Research Lib	2
Research Lab	1
Full time MLS student	1

II.2 Task description

Below are the verbatim instructions for the task set our searchers. The cut-off level for each topic was written on the topic description.

THE DOCUMENT ROUTING TASK

Task and context

Imagine a situation in which you are an intermediary using the INQUERY system to satisfy the information needs of a user. Rather than being a one-time affair, the user wants you to develop a query he or she can use repeatedly in the future to retrieve documents of interest. For example, the user may want to reissue the same query every month to stay up-do-date on developments in his or her field. Thus, your goal is to develop a query that not only retrieves relevant documents now (and ranks them reasonably high) but that will work equally well on future document collections with different documents.

Since ranking systems like INQUERY may retrieve large numbers of documents, the user will not be able to look at all documents the system retrieves. The number of documents that a user might be willing to look at will vary by each information problem. The user will tell you for each topic how many documents the user is willing to read. The user will not consider any document ranked below this number.

Topic Description

The user has prepared a statement that describes in detail the topic he or she is interested in. This statement also contains a description of the kinds of documents the user considers to be relevant and those s/he considers not to be relevant. Assume that the user has left you with this written statement, i.e. you cannot get clarifications and you cannot negotiate changes.

For the purposes of training we are using a somewhat outdated document collection. Because of this, some topics will appear outdated as well. Please act as if you are doing this search at a time in which the topic would have been relevant.

User Judgments

To help you with your task, the user has looked through parts of the currently available document collection and has noted which of the documents are relevant and which are not relevant. These ratings have been made available to you in the INQUERY system. The summary for each document has a notation after the belief value that states the judgement of the user: documents that the user found to be relevant are marked "REL" and documents the user judged not- relevant are marked "NOT". If the user has not judged a document it is marked as "?". The judgements are also given above the full display of a document. Again, since the user is not currently present, you should accept ratings as they are.

- Reset the system
- Develop a query for the topic description handed to you.

  Your query should retrieve as many documents as possible among the top number of documents that the user is willing to look at which satisfy the specified information needs. The query should also be likely to do will on future document collections.

  NB: Each topic had a routing cut-off written on the topic description.

  You may use any of the system features (including automatic relevance feedback) you have learned about, but you are not required to do so.

  You have up to 20 minutes to complete this task.

  We are interested in how people interact with the INQUERY system. Please think-aloud throughout this session, i.e. we ask you to verbalize any thoughts, comments, plans, or goals related to what you are doing and why you are doing it as they come to mind. Act as if you were talking to yourself (rather than talking to us). Occasionally we may remind you to keep talking.
- Once you are satisfied with your result, save your final query under the name "final".

II.3 Training

II.3.a Description of the training process

Subjects were given a short general description of the TREC project and were given a consent form that outlined the experiment and the data collection. Upon signing of the form they were led to a screened-off area that was set up with the computer equipment and a video recording device. The experimenter gave a 2-minute introduction to INQUERY by listing its major features, namely ranked output (rather than set based retrieval), full text search (rather than title, abstract, or keyword search), the availability of automatic relevance feedback functionality, and the ability to process free form, unstructured queries. Subjects were next given a 14 page, ring-bound tutorial. The tutorial contained a sequence of interaction exercises that demonstrated the various interface features, the use of operators, and the use of automatic relevance feedback. Subjects set their own pace working through the tutorial. One experimenter was seated next to the subject and provided additional help and guidance if the subject got stuck or had questions that were not answered by the tutorial. The experimenter did not carry out any interactions with the system except for occasional restarts of the system necessitated by system failures; most interventions by the experimenter were short interjections that made sure that the intended learning experience took place. The document collection used in the tutorial was disk 3 of the TREC test collection and all examples were taken from the implicit context of topic 77 (poaching of wildlife). Subjects completed the tutorial's main portion (the first eleven pages) in 50-70 minutes.

The last part of the tutorial (after a short break, if desired) consisted of a test trial of an actual search. Subjects received the same task description used later in the actual trials (see II.2) and practised on a topic that was not part of the TREC-3 topic set, namely topic 84 (Alternative/Renewable Energy Plant Equipment Installation). The document collection was again the TREC test collection disk 3. Subjects were also asked to "think-aloud" during this phase as this was a requirement during the later search trials. An experimenter operating under the same model described above was available during this trial search as well. No feedback was given by the experimenter regarding the quality of individual queries nor regarding the quality of the overall search session. Subjects were asked after 15 minutes to wrap up their search and to save a final query formulation.

II.3.b Time for training

Time for training in minutes (times are approximate whole minute): Mean:80, Range: 65 - 90

III Search Process

1. Clock time (i.e. real, elapsed time) per search, from the time the searcher was given the topic, until the final query was saved, in seconds. Note that users were restricted to 20 minutes (1200 seconds).

Mean: 928.6, Median: 981.0, SD: 245.24, Range: 421 - 1328

2. Number of documents "viewed" in during the search.

a. Definitions of viewing:

We report on three categories of viewing, as defined below:

1. "All titles" defined as the total number of document titles displayed during the course of the entire search (the default display is five titles at a time). This can be thought of as title "tokens".

2. "Unique titles", defined as the number of unique titles seen by the user during the course of the entire search. This can be thought of as title "types".

3. "Text", defined as the number of documents for which the user had the text of a document displayed.

b. Number of items viewed per search (repeat for each viewing category).

Viewing Behavior				
Viewing Type	Mean	Median	SD	Range
All titles	136.3	94.5	118.2	15-515
Unique titles	67.2	51.0	55.6	8-310
Text	6.6	5.5	6.2	0 - 36

3. Number of iterations per search.

a. Definition of iteration:

An iteration is all that occurs between one invocation of the "run query" button and the next invocation of that button. The first iteration begins when the user is given the topic for searching, and the last iteration ends when the user invokes the "save query" button, saving the query as "final". The query run by an invocation of the "run query" button is part of the iteration defined by that invocation as the end point.

b. Number of iterations per search.

Mean: 7.76, Median: 7.0 SD:3.98, Range: 2 - 21

4. Number of terms used in queries.

a. Number of terms in the first query of a search, per search.

Mean: 5.38, Median: 4.0, SD: 3.90, Range: 1 - 21

b. Number of terms in the final query of a search, per search.

Mean: 23.82, Median:18.0, SD: 22.50, Range: 3 - 120

5. Use of system features.

The following table provides summary statistics for the use of system features, $N = 50$ topics (10 searchers).

Use of System Features				
Feature	$\overline{X}$	M_d	SD	Range
bargraph double-click (to view fulltext)	0	0	0	0
scrollbar in summary	19.34	10.5	21.9	0–95
scrollbar in document window	2.72	2.0	3.6	0–20
double-click on title (to view fulltext)	4.82	4.0	4.0	0–16
next document (to view fulltext)	1.14	0.0	3.1	0–20
previous document (to view fulltext) 5 uses by 4 searchers	0.1	0	0	0–2
keyword	6.48	0.5	11.4	0–41
save query	2.14	1.0	1.8	1–7
load query	0.72	0.0	1.8	0–8
relevance feedback on	1.08	1.0	1.0	0–5
relevance feedback off	0.38	0.0	0.8	0–4
relevance feedback used (number of iterations/search)	1.68	1.0	1.9	0–7
documents selected (as relevant)	8.26	5.0	9.9	0–54
documents unselected	0.50	0.0	1.2	0–7

6. Number of user errors made per search

a. Definition of an error

Error is defined as the use of syntax in a query which caused the query to be uninterpretable by the system query parser (the most common syntax error was to not include a required closed parenthesis).

b. Number of user errors per search

Mean: 0.7, Median: 0, SD:1.18, Range: 0 - 5

7. Transcript for Topic 122 (Searcher 7)

What follows is the log of the interaction for topic 122 (searcher S07, trial 4. Given are the elapsed time, the real time, a description of the searcher's actions, system responses and experimenter comments, Statements enclosed in "" constitute the complete verbal protocol of the searcher. The full text of viewed documents is reproduced for short documents, for long documents only the approximate amount that fit on the screen and was visible without scrolling is given.

```
Iteration 1:
00:00:00 11:46:34  START topic from VIDEO
00:00:15 11:46:49  starts entering query, enters "uw50" (misses # in uw)
00:00:26 11:47:00  backspaces over uw50 and replaces it with #uw50(
00:00:54 11:47:28  adds term #syn and term - pauses: #uw50(cancer #syn (drug

                   "That is something I don't understand about this system..."
                   "chemotherapy ... treatment ..."
                   enters complete query:
                   #uw50 (cancer #syn (drug chemotherapy treatment))
00:01:28 11:48:02  hit run query button

Iteration 2:
                   *** parsed query ***
                   #UW50( cancer #SYN( drug chemotherap treatment))
                   *** Query processing: 4 words
00:01:34 11:48:08  ***  new summary  ***
                   1. 0.653723 NOT AP880524-0052 Ex-President's Brother Enters Cancer Institute for Treatment
                   2. 0.653723  ?  AP880322-0156 Experimental Treatment Avoids Surgery for Bladder Cancer
                   3. 0.653723 REL WSJ870415-0071 Cetus Venture Wins Approval to Market Anti-Cancer Drug -
                   4. 0.653723  ?  DOE2-42-0592
                   5. 0.653723  ?  DOE2-38-0875
                   *** 2443 documents retrieved ***
00:01:50 11:48:24  *** Double_Click_Title ***
                   Reads document, voice not audible
                   3. 0.653723 REL WSJ870415-0071

                   Cetus Venture Wins Approval to Market Anti-Cancer Drug ---
                   Partnership With Ben Venue Gets FDA's Clearance To Sell
                   Generic Product --- By Marilyn Chase Sta

                     Cetus Corp. said its joint venture with Ben Venue
                   Laboratories Inc. received
                   approval from the U.S. Food and Drug Administration to begin
                   marketing its second generic anti-cancer product.

                     Cetus, based in Emeryville, Calif., and Ben Venue, Bedford,
                   Ohio, will soon begin marketing vinblastine sulfate, a
                   standard chemotherapy agent extracted from periwinkle
                   flowers and used in the treatment of Hodgkin's disease,
                   lymphomas, testicular cancer, unresponsive breast cancer and
                   other malignant tumors.
                   [...]

00:02:12 11:48:45  hits save query button
                   enters file name "cancerdrug''; queries read:
                      #uw50 (cancer #syn (drug chemotherapy treatment))
                   RF: #UW50( cancer #SYN( drug chemotherap treatment))
00:02:27 11:49:00  clicks on "OK"
00:02:32 11:49:05  chooses "Reset All"
00:02:35 11:49:08  enters "uw" backspaces over it
00:02:37 11:49:10  enters "#uw"  pauses
00:02:48 11:49:21  adds to query to read "#uw100 (cancer"  and pauses
00:03:08 11:49:41  adds #syn( for query to read " #uw100 (cancer #syn(" pauses
00:03:10 11:49:43  completes query to read:
                   #uw100 (cancer #syn(treatment research testing evaluation)
00:03:27 11:50:00  clicks on "Run Query" button
```

```
Iteration 3:
00:03:29 11:50:02  Gets syntax error msg.(small window)
00:03:30 11:50:03  Gets error parsed query window [missed closing parenthesis]
00:03:31 11:50:04  Dismisses  error message
00:03:32 11:50:05  Dismisses large window error msg
00:03:33 11:50:06  "I have an about 50% hitrate on putting the parenthesis"
00:03:36 11:50:09  "I'm also trying to run a demo search system..."
                   adds the closing parenthesis for query to read:
                   #uw100 (cancer #syn(treatment research testing evaluation))
00:03:39 11:50:12  clicks on "Run Query" button

Iteration 4:
                   *** parsed query ***
                   #UW100( cancer #SYN( treatment research test evalu))
                   *** Query processing: 5 words
00:03:46 11:50:19  gets ***  new summary  ***
                   1. 0.632350  ?  AP880322-0156 Experimental Treatment Avoids Surgery for Bladder Cancer
                   2. 0.632350 REL WSJ870519-0014 Tests to Begin on a Toxin That Attacks Cancer Cells
                   3. 0.632350  ?  FR891011-0077
                   4. 0.632350  ?  FR89322-0189
                   5. 0.632350  ?  DOE2-42-0592
                   *** 3954 documents retrieved ***
00:03:48 11:50:21  "A lot of question marks, not a lot of no's (?)"
00:03:53 11:50:26  clicks on "Save Query" button
                   enters file name "cancertreat"; queries read:
                       #uw100 (cancer #syn(treatment research testing evaluation))
                   RF: #UW100( cancer #SYN( treatment research test evalu))
00:04:04 11:50:37  clicks on "OK"
00:04:14 11:50:47  clicks on "Reset All"
00:04:17 11:50:50  starts entering query  "#syn( cancer leukemia )"
00:04:31 11:51:04  "Hmm actually"  picks up topic sheet reads description
                   "these [inaudible]  clients actually...
                   he or she is what I mean is actually the client
                   is actually interested only in the specifics of it [??]...
                   but there are a lot of different forms of cancer ...
                   [inaudible] ... it doesn't matter what ... [sighs]
                   it doesn't matter what kind of cancer it is... Hmm...
                   I don't know what kind of cancer ... [???inaudible]
                   #syn (cancer leukemia )
00:05:20 11:51:53  clicks on "run Query" button

Iteration 5:
                   *** parsed query ***
                   #SYN( cancer leukemia)
                   *** Query processing: 2 words
00:05:24 11:51:57  gets ***  new summary  ***
                   1. 0.593854 REL WSJ920318-0054 Medicine:    Scientists Say Progress Is Made
                   2. 0.593854 NOT WSJ920305-0102 Technology &amp; Health: Colon Cancer Test Cuts Chance
                   3. 0.593854 NOT WSJ920302-0204 Technology &amp; Medicine:    ICI Cancer Drug's
                   4. 0.593854 NOT WSJ920226-0129 Technology:    Medarex Drug Trials Approved
                   5. 0.593854 NOT WSJ920219-0135 Technology &amp; Medicine: Combination of Acne Drug...
                   *** 9412 documents retrieved ***
00:05:26 11:51:59  click on "Save Query" button
                   enter filename "cancerany"; queries saved:
                       #syn (cancer leukemia )
                   RF: #SYN( cancer leukemia)
00:05:32 11:52:05  click on "OK"
00:05:35 11:52:08  click on "Reset All"
00:05:35 11:52:08  click on "Reset All"
00:05:36 11:52:09  "I'm gonna see what I'm getting (?)
00:05:39 11:52:12  click on "Load Query" button
                   selects file "cancerany" which reads:
                   #syn (cancer leukemia )
00:05:43 11:52:16  clicks on "OK"
00:05:49 11:52:22  clicks on "Load Query" button
                   selects file "cancerdrug" which reads:
                   #uw50 (cancer #syn (drug chemotherapy treatment))
00:05:53 11:52:26  clicks "OK"
```

```
00:06:02 11:52:35  clicks on "Load Query" button
00:06:03 11:52:36  move cursor over files, seems unsure which one to load next
00:06:11 11:52:44  selects file "cancertreat" which reads:
                   #uw100 (cancer #syn(treatment research testing evaluation))
00:06:14 11:52:47  clicks on "OK"
                   current content of query window now:
                   #syn (cancer leukemia )
                   #uw50 (cancer #syn (drug chemotherapy treatment))
                   #uw100 (cancer #syn(treatment research testing evaluation))
00:06:19 11:52:52  clicks on "Run Query" button

Iteration 6:
                   *** parsed query ***
                   #SUM( #SYN( cancer leukemia) #UW50( cancer #SYN( drug
                   chemotherap treatment)) #UW100( cancer #SYN( treatment
                   research test evalu)))
                   *** Query processing: 11 words
00:06:23 11:52:56  "Let's see what happens .. comes back ... up [??]
                   ..'cause the ...ah.. "
00:06:29 11:53:02  gets ***  new summary  ***
                   1. 0.626642  ?  AP880322-0156 Experimental Treatment Avoids Surgery for Bladder Cancer
                   2. 0.626642  ?  DOE2-42-0592
                   3. 0.626642  ?  DOE2-38-0875
                   4. 0.626642  ?  DOE2-32-1236
                   5. 0.626642  ?  DOE2-10-1097
                   *** 9412 documents retrieved ***
00:06:31 11:53:04  "[laughs] well comes up with a lot of questions...oh well"
00:06:35 11:53:08  double-click on title 2. full text for doc 2. displayed
                   2. 0.626642  ?  DOE2-42-0592

                   Electron beam therapy has become more widely available in the radiation
                   oncology community. It is a unique modality, offering important contributions
                   to the management of cancer. It is, however, necessary to optimize dose
                   distribution, patent selection, and treatment techniques.

00:06:40 11:53:13  ''at least it didn't come up with instant no's'' [??]
00:06:46 11:53:19  looks at document (evaluates its relevance)
                   "well  [inaudible]  research and development"
00:06:58 11:53:31  clicks on *** Next_Doc *** button, new full text displayed
                   3. 0.626642  ?  DOE2-38-0875

                   Indications and contraindications for radiation treatment of esophagus
                   cancer are presented. The role of chemoradiation among esophagus cancer
                   treatment methods is determined. Thechnical,dosimetric and clinical data
                   are sequently delivered. Preparation of a patient for chemoradiation
                   is described. Recommendations on their most efficient use are given.
                   95 refs.; 15 figs.; 19 tabs.

00:07:11 11:53:44  looks at document full text
                   "treatment is getting some ...
00:07:20 11:53:53  clicks on "Reset All" button
00:07:23 11:53:56  clicks on "Load Query" button
                   selects file "cancertreat" which reads:
                   #uw100 (cancer #syn(treatment research testing evaluation))
00:07:27 11:54:00  clicks on "OK"
00:07:28 11:54:01  "I better modify this"  (deletes the word "treatment"), now:
                   #UW100( cancer #SYN( research test evalu))
00:07:37 11:54:10  clicks on "Run Query" button

Iteration 7:
                   *** parsed query ***
                   #uw100 (cancer #syn( research testing evaluation))
                   *** Query processing: 4 words
00:07:40 11:54:14  "I'm leaving out treatment because it was... hmmm"
00:07:43 11:54:17  gets ***  new summary  ***
                   1. 0.644194 REL WSJ870519-0014 Tests to Begin on a Toxin That Attacks Cancer Cells
                   2. 0.644194  ?  FR891011-0077
                   3. 0.644194  ?  FR89322-0189
```

```
                  4. 0.644194  ?  DOE2-20-0076
                  5. 0.644194  ?  DOE1-87-0742
                  *** 3028 documents retrieved ***
00:07:47 11:54:21 double-clicks on title of document 1. Full text gets displayed:
                  1. 0.644194 REL WSJ870519-0014

                  Tests to Begin on a Toxin That Attacks Cancer Cells

                    Cetus Corp. said the first clinical tests on humans will soon begin involving
                  a breast cancer treatment that " targets"  diseased cells and attacks them with
                  a toxintipped monoclonal antibody.

                    Under the treatment, a toxin will be bonded to antibodies that recently were
                  proven capable of locating and attaching themselves to cancer cells in humans.
                  Scientists hope the toxin will destroy the targeted cells without the damage to
                  healthy tissue that accompanies conventional chemotherapy  cancer treatments.
                  [...]

                  An official of Cetus, a biotechnology concern that makes both the toxin and the
                  antibody, stressed that the initial test will involve only a few patients.
                  If it is successful, he said, the tests will be expanded, but he added it
                  will probably be " several years"  before the company would be prepared to
                  seek Food and Drug Administration approval to market the product.
                  The treatment has proven successful in tests on animals, he added.

                  Cetus, which said it expects to begin similar tests on ovarian cancer in humans
                  later this year, noted that other companies are testing the process on other
                  cancers,including melanoma and colon cancer.

00:07:49 11:54:23 reads through document, moves cursor over full text.
00:08:04 11:54:38 "I'm looking for  any  added term  I could conceivably use to
00:08:07 11:54:41 clicks on "Scroll to Keyword" button
00:08:10 11:54:44 clicks on "scroll to Keyword" button (no change in display (?))
00:08:15 11:54:49 moves cursor into bargraph, onto scroll in bar graph
                  but reconsiders and moves cursor back into summary window
00:08:18 11:54:52 double-clicks on title for doc 4. Full text displayed:
                  4. 0.644194  ?  DOE2-20-0076

                  In this paper, the emphases are put on the description of design principle
                  and calculation method for radiation protection in the new radiotherapy
                  department of Cancer Research Institute and Hospital, Chinese Academy
                  of Medical Science, as well as the evaluation of meusaring results. In
                  addition, the problem of photo-neutron contamination among 16 MV X-ray
                  from SL 75-20 Philips Linear Accelerator and its relavent shielding measures
                  have been discussed.

00:08:23 11:54:57 "I'm looking at question mark documents to see...
                  "radiation protection [from doc text] that's not relevant"
00:08:36 11:55:10 double-clicks on title for doc 5. Full text displayed:
00:08:37 11:55:11 reads doc full text (cursor in full text window)
                  5. 0.644194  ?  DOE1-87-0742

                  High Performance Gel Permeation Chromatography (GPC) was evaluated as
                  an alternative to the more expensive Nuclear Magnetic Resonance (NMR)
                  spectroscopy technique for cancer detection using human plasma. These
                  two techniques show a biphasic relationship which can be explained on
                  the basis of the relative amounts of the lipoprotein levels present in
                  the plasma and a good correlation with total triglyceride concentrations
                  obtained from standard blood tests. The major difference in the GPC elution
                  nprofiles (254 nm) of plasma from normal individuals and that from cancer
                  patients occurred in the peak eluting at the void volume. This peak has
                  a retention time consistent with very low density lipoprotein (VLDL)
                  and is elevated in most cancer patients and in normal patients with triglyceride
                  levels greater than 200 mg/ml. The use of these techniques as a screening
                  test for cancer in an asymptomatic population needs further evaluation.

00:08:54 11:55:28 clicks on "Save Query" button
                  edits default file name (add 2) to "cancertreat2 ", which reads:
```

```
                    #uw100 (cancer #syn( research testing evaluation))
                    RF: #UW100( cancer #SYN( research test evalu))
00:09:01 11:55:35  clicks on "OK"
00:09:06 11:55:40  "So I'm gonna to try again having left out treatment"
00:09:07 11:55:41  clicks on "Load Query" button
                    selects file "cancerany" which reads:
                    #syn (cancer leukemia )
00:09:12 11:55:46  clicks on "OK"
00:09:15 11:55:49  "Lets see what I get"
00:09:16 11:55:50  clicks on "Load Query" button
                    selects file "cancerdrug" which reads:
                    #uw50 (cancer #syn (drug chemotherapy treatment))
00:09:21 11:55:55  clicks on "OK", contents of query window now:
                    #uw100 (cancer #syn( research testing evaluation))
                    #syn (cancer leukemia )#uw50 (cancer #syn (drug chemotherapy treatment))
00:09:25 11:55:59  "The more I use this the more confused I get"
00:09:30 11:56:04  "Put a return "
                    adds a return that puts last loaded query on newline:
                    #uw100 (cancer #syn( research testing evaluation))
                    #syn (cancer leukemia )
                    #uw50 (cancer #syn (drug chemotherapy treatment))
00:09:31 11:56:06  *** Run_Query ***

Iteration 8:
                    *** parsed query ***
                    #SUM( #UW100( cancer #SYN( research test evalu)) #SYN( cancer
                    leukemia) #UW50( cancer #SYN( drug chemotherap treatment)))
                    *** Query processing: 10 words
00:09:40 11:56:15  gets ***  new summary  ***
                    1. 0.625523 REL WSJ870519-0014 Tests to Begin on a Toxin That Attacks Cancer Cells
                    2. 0.614311 REL WSJ910412-0102 Technology &amp; Medicine: Cancer Drug Taxol May Wor
                    3. 0.614101 NOT AP881023-0047 New Cancer Treatment Tested On Two Patient
                    4. 0.613794 REL WSJ900419-0016 Technology &amp; Medicine:    New Drug Shows Promise Tr
                    5. 0.612908 NOT WSJ870424-0155 Whose Life Is It Anyway? --- By Robert K. Oldham
                    *** 9412 documents retrieved ***
00:09:45 11:56:20  "Hmm... a little better"
00:09:46 11:56:21  Double-click on title for document 4. Full text displayed:
                    4. 0.613794 REL WSJ900419-0016

                    Technology &amp; Medicine:    New Drug Shows    Promise
                    Treating    Type of Leukemia    ----    By Ron Winslow
                    Staff Reporter of The Wall Street Jour

                    A new cancer drug proved strikingly successful with very
                    few side effects in treating a relatively rare form of
                    leukemia, researchers said.
                       The drug, called 2-CdA, was nearly 100% effective treating
                    12 patients with hairy-cell leukemia. The patients underwent
                    just one seven-day treatment and experienced almost none of
                    the debilitating side effects such as vomiting, hair-loss and
                    kidney and liver problems associated with other chemotherapy.

                       The results need to be replicated by other researchers
                    before definitive conclusions can be made, and the long-term
                    effect of the treatment remains unknown. But the initial
                    experience with the compound is especially promising, and an
                    apparent improvement over other therapy for the disease,
                    scientists said.
                    [...]
00:09:56 11:56:31  "Okay It gets as good as I'm going to be able to get"
                    "It won't   [inaudible]''  [laughs]
00:10:02 11:56:37  scrolls outside doc counter to top, new full text display:
                    1. 0.625523 REL WSJ870519-0014

                    Tests to Begin on a Toxin That Attacks Cancer Cells

                        Cetus Corp. said the first clinical tests on humans will soon begin involving
                    a breast cancer treatment that " targets"  diseased cells and attacks them with
```

```
                    a toxintipped monoclonal antibody.

                        Under the treatment, a toxin will be bonded to antibodies that recently were
                    proven capable of locating and attaching themselves to cancer cells in humans.
                    Scientists hope the toxin will destroy the targeted cells without the damage to
                    healthy tissue that accompanies conventional chemotherapy cancer treatments.
                    [...]
00:10:10 11:56:45   scroll outside doc counter: new summary and full text displayed:
00:10:11 11:56:46   gets ***  new summary  ***
                    79. 0.587702 NOT DOE1-29-0593
                    80. 0.587600 NOT WSJ911226-0005 Technology &amp; Health: Many Women With Breast Cancer
                    81. 0.587433 NOT WSJ910405-0005 Technology:    Genentech Inc. Cancer Therapy
                    82. 0.586518 NOT AP880525-0051 Signs That Chemotherapy May Help Control Lung Cancer
                    83. 0.586406 NOT WSJ911107-0058 Business Brief -- Schering-Plough Corp.: Company Reite


                    79. 0.587702 NOT DOE1-29-0593

                    The combination of radiotherapy (RT) and chemotherapy (CT) has markedly
                    improved the therapeutic results for those tumors which are both chemosensitive
                    and radiosensitive, such as lymphomas, embryonal tumors, small cell lung
                    carcinomas, breast cancers, etc. Despite some spectacular results reported
                    following non-controlled studies, a significant increase in the total
                    survival or the relapse free survival has never been documented in controlled
                    trials in head and neck, anal, ovarian carcinomas. However, in these
                    tumors, a combination of RT and CT may reduce the mutilations and sequellae
                    caused by the treatment and may induce an increase in the survival in
                    some subsets of patients. Further clinical research is needed along these
                    lines. Cross resistance between ionizing radiation and
                    [...]
00:10:13 11:56:48   "[still laughing] oh you know .. there are some there .. there
                    are not a 150 there"
00:10:19 11:56:54   scrolls outside doc counter to top: new summary and full text:
00:10:20 11:56:55   gets ***  new summary  ***
                    1. 0.625523 REL WSJ870519-0014 Tests to Begin on a Toxin That Attacks Cancer Cells
                    2. 0.614311 REL WSJ910412-0102 Technology &amp; Medicine:    Cancer Drug Taxol
                    3. 0.614101 NOT AP881023-0047 New Cancer Treatment Tested On Two Patients
                    4. 0.613794 REL WSJ900419-0016 Technology &amp; Medicine:    New Drug Shows
                    5. 0.612908 NOT WSJ870424-0155 Whose Life Is It Anyway? --- By Robert K. Oldham

                    1. 0.625523 REL WSJ870519-0014
                    Tests to Begin on a Toxin That Attacks Cancer Cells

                       Cetus Corp. said the first clinical tests on humans will soon begin involving
                    a breast cancer treatment that " targets"  diseased cells and attacks them with
                    a toxintipped monoclonal antibody.

                       Under the treatment, a toxin will be bonded to antibodies that recently were
                    proven capable of locating and attaching themselves to cancer cells in humans.
                    Scientists hope the toxin will destroy the targeted cells without the damage to
                    healthy tissue that accompanies conventional chemotherapy cancer treatments.
                    [...]
00:10:21 11:56:56   "I used chemotherapy"
00:10:26 11:57:01   "I really, I really think this is, I think this is the final one.''
00:10:29 11:57:04   clicks on "Save Query" button
                    enters name "final", query reads:
                        #uw100 (cancer #syn( research testing evaluation))
                        #syn (cancer leukemia )
                        #uw50 (cancer #syn (drug chemotherapy treatment))
                    RF: #SUM( #UW100( cancer #SYN( research test evalu)) #SYN( cancer
                        leukemia) #UW50( cancer #SYN( drug chemotherap treatment)))
00:10:36 11:57:11   Clicks on "OK"
00:10:40 11:57:15   "[inaudible] in my particular performance"[laughs]
```