Rutgers' TREC-7 Interactive Track Experience

N.J. Belkin, J. Perez Carballo, C. Cool*, D. Kelly, S. Lin,
S.Y. Park, S.Y. Rieh, P. Savage-Knepshield, and C. Sikora

School of Communication, Information & Library Studies
Rutgers University
4 Huntington Street
New Brunswick, NJ 08901-1071
*GSLIS, Queens College, CUNY

Abstract

We present results of a study comparing two different interactive information retrieval systems: one which supports positive relevance feedback as a term-suggestion device; the other which supports both positive and negative relevance feedback in this same context. The purpose of the study was to investigate the effectiveness and usability of a specific implementation of negative relevance feedback in interactive information retrieval. A second purpose was to investigate the effectiveness and usability of relevance feedback implemented as a term-suggestion device. The results suggest that, although there was no benefit in terms of performance for the system with negative and positive relevance feedback, this might be due to specific implementation issues.

1.0 Introduction

As in TREC-7, we continued the work begun in our TREC-6 experiments (Belkin, et al., 1998), investigating the effectiveness and usability of negative relevance feedback (RF) in interactive information retrieval (IR). One reason for considering negative RF was that subjects in our previous TREC experiments had expressed the desire to be able to make negative judgments on retrieved documents which would subsequently affect retrieval and ranking. Another was our belief that a particular way of implementing negative RF would lead to identifying documents in which "good" query terms appear in inappropriate contexts (Belkin, et al., 1998).

Our TREC-6 results seemed rather inconclusive, primarily because of the small number of subjects taking part in the experiment, the small number of searches that they conducted, and because of what seemed to be problems with our interface design. The TREC-7 experimental protocol gave us the opportunity to compare directly, rather than indirectly, our two conditions (positive RF only, versus positive plus negative RF), thereby also increasing the number of subjects in each condition, as well as the number of searches by each subject. In addition, we redesigned our interface to take account of problems that were made evident in TREC-6.

Following the results of Koenemann (1996), we implemented RF as a term-suggestion device for user-controlled query expansion. In TREC-7, we attempted to investigate the effectiveness and usability of this implementation of RF, or at least of its effect on our main questions concerning negative and positive RF.

Below, we first discuss the systems that we used to investigate our research questions, and the methods that we applied. We follow with an overview of the results, divided into three sections: characteristics of the subjects; *effectiveness* of the positive RF system (RUINQ-G) versus the positive plus negative RF system (RUINQ-R), and of term suggestion; and, *usability* of the two systems and of term suggestion. We then discuss some relationships amongst these results, and some interpretations of them, and conclude with some suggestions about where to go next.

2.0 Methods

This section describes our subjects, experimental IR systems, and the procedures that we followed while conducting our TREC-7 Interactive Track experiment.

2.1 Searchers

Sixteen volunteer searchers were recruited to participate in this study from the population of students in the School of Communication, Information and Library Studies at Rutgers University. None of our subjects had taken part in previous TREC studies and none had prior experience with our RU-INQUERY system. Demographic characteristics of the searchers and their experiences with IR systems are described in Section 3.1.

2.2 Experimental IR Systems

We used Inquery 3.1p1 including its default values for indexing, retrieval, and RF. The major difference between our implementation and the standard version

of Inquery, apart from the interface, was in RF implementation. We modified Inquery's RF function so that it produced a list of 50 terms, for both positively and negatively judged documents. As users made RF judgments about documents, the top *n* terms were presented in a term suggestion window. At the user's discretion, these terms could be added to the existing query. The term ranking algorithm was default Inquery.

Positive and negative RF were both implemented using the standard Inquery method for positive relevance judgments. However, terms that co-occurred in highly-ranked negatively judged documents and in either the original query or in positively judged documents were excluded from the suggested "bad" terms list. The number of terms suggested was determined by the formula: $n = 5i + 5$ in which i is the number of judged documents, and n is no greater than 25.

Appendix A contains a screen dump of the positive and negative RF system (RUINQ-R). The positive only RF system (RUINQ-G) was identical, except that there were no Bad Terms to Avoid window, no Clear Bad Docs button, and no Bad RF radio buttons. Our interface offered the following features and functionality:

- Query terms window - used to input a free-form query, with minimal structure (phrases and negative terms).
- Results Summary window - displayed the titles of ten documents and provided radio buttons for marking documents as good, bad (in the positive and negative RF condition), and saved.
- Document window - displayed text of a selected document.
- Pop-up Instance Labeling window - used to label saved documents according to the "instances" that they represented.
- Documents Saved window - listed the saved document's title and its associated instance label.
- Good Terms to Add window -- displayed suggested terms which could be added to the query by clicking on them.
- Bad Terms to Avoid window -- displayed suggested terms which could be added to the query by clicking on them (this window was only presented in the positive and negative RF condition).
- Search Button - used to retrieve a list of documents.
- Clear Query button - used to remove all terms in the query terms window.
- Clear Good Documents and Clear Bad Documents Buttons -- used to "unmark" previously marked good and bad documents, respectively.
- Show Next Keyword, Show Best Passage, Show Next, and Show Prev buttons - used to quickly navigate through the full text of a document.
- Exit button - used to end a search session.

Both systems ran on a SUN Ultra 140 with 64MG memory and 9GB disk under Solaris 2.5.1 with a 17" color monitor.

2.3 Procedure

Each searcher conducted eight searches in accordance with the TREC-7 Interactive Track experimental guidelines. Searchers were alternately assigned to one of two experimental conditions. In one condition, searchers conducted four searches using RUINQ-R with positive and negative relevance feedback (RF) and then conducted their next four searches using RUINQ-G with positive only relevance feedback. In the second condition, system order was reversed so that searchers used the system with positive only RF followed by the system with positive and negative RF. Within each condition, topic block presentation was counterbalanced so that half of the subjects searched on topic block B1 (365i, 357i, 362i, 352i) first, while the other half searched on block B2 (366i, 392i, 387i, 353i) first.

On arrival, the subjects read and signed a consent form explaining their rights and the potential risks associated with participation in the experiment. Next, they completed a demographic questionnaire that gathered background information and probed their previous searching experience. After completing the Controlled Associations Test FA-1, they were ready to begin their first tutorial. The tutorial familiarized them with the features and usage of the RU-INQUERY system that they would be using for the first half of the experiment. They received written instructions for the task in general and specific instructions for the current search topic. They were allotted 15 minutes to complete each search. As they searched, they specified "instances" of the topic as they identified them and "thought aloud." A videotape recorded the computer monitor during their searches and captured their "thinking aloud" utterances and the entire search interaction was logged. After completing each search, a brief questionnaire was completed in which participants assessed their topic familiarity, search difficulty, search result satisfaction, aspect identification confidence, and satisfaction with the amount of time allotted for the search. This process was completed for three more searches and then participants were given the opportunity to take a break. After the break, the same process was repeated

for the second system that they used to conduct the remaining four searches. After completing all eight searches, the participants completed an exit questionnaire and an exit interview. We added the exit interview and some additional survey questions to the standard Interactive Track data collection instruments, to better understand users' perceptions of the usefulness of both positive and negative RF, of the usefulness of term suggestion, and of the usability of the two systems' interfaces. All subjects were tested individually and required approximately 3-1/4 hours to complete the study.

3.0 Results

3.1 Characteristics of the Subjects

The subject group included 11 females and 5 males, whose ages ranged from 23 to 58. Thirteen of the subjects either had, or were pursuing a graduate degree in library science. Of these 13 subjects, 1 was a Ph.D. student in library science and one had already earned a Ph.D. in a subject area outside of library science. Three of the subjects either had, or were candidates for Master's level degrees in areas outside of library science. One subject had only a high school diploma. The occupations of the subjects did not vary greatly. Eight subjects reported being students, 4 reported being librarians, 2 reported being members of academic faculty and 2 were in neither of these categories. None of the subjects reported having previously participated in any TREC experiments.

The median number of years reported for overall experience doing online searching was 3.00 (M=3.59, SD=3.76). The minimum amount of experience reported was 1 and the maximum was 17. The average amount of previous search experience on different types of systems did not vary by much. This previous search experience was accessed using a five-point scale where 0=no experience and 5=great deal of experience. Subjects reported the most experience using a point and click interface (M=4.75, SD=0.68) and the least experience using CD-ROMs (M=3.06, SD=1.48). The average rating and standard deviations for subjects' search experience with library catalogs, commercial online systems, and the World Wide Web were, respectively: M=4.18, SD=0.83; M=3.46, SD=0.99; M=4.37, SD=0.80. Subjects reported conducting searches an average of 4.43 (SD=0.62) on a five-point scale where 1=never; 2=once or twice a year; 3=once or twice a month; 4=once or twice a week; and 5=once or twice a day. Subjects also reported the extent to which they enjoy carrying out information searches. A different five-point scale was used where 1=strongly disagree; 3=neutral; and 5=strongly agree. The average response was 4.31 (SD=0.70).

3.2 Effectiveness

3.2.1 Effectiveness of Positive versus Positive Plus Negative Relevance Feedback

The precision and instance recall for all subjects in both systems were 0.64 and 0.37 respectively. Overall, the subjects saved 6.16 documents in average within 892.22 seconds (14.74 minutes). Table 1 presents these results, comparing mean performance in RUINQ-G (positive RF only) and RUINQ-R (positive plus negative RF). The differences in peformance between the two systems are insignificant on all four measures; for the basic performance measure of *instance recall* the relevant figure is $t(125) = 0.925$.

In order to check for interaction effects, we compared performance by system order and by topic block order. The results showed that the subjects who used the RUINQ-R system first performed a little bit better in terms of recall (M= .38, SD= .23) than those who used RUINQ-G system first (M= .35, SD=24). However, t-test results revealed that the difference was not statistically significant: $t(125)= -.635$. The means of instance recall between two different block order are almost the same, M= .37 (SD= .25) in Block1 and M= .36 (SD= .22) in Block2.

A *cycle* in our analysis is defined as the number of invocations of the "Search" button plus one. Transaction logs saved during the searches revealed that the subjects engaged in 6.8 cycles per search. Overall, they identified 5.02 instances per search. For

	RUINQ-G Mean (SD)	RUINQ-R Mean (SD)	Total Mean (SD)
Time (seconds)	892.55 (206.47)	891.89 (179.86)	892.22 (192.75)
Documents saved	6.46 (3.78)	5.87 (3.86)	6.16 (3.81)
Precision	.64 (.27)	.63 (.30)	.64 (.29)
Instance Recall	.39 (.24)	.35 (.23)	.37 (.23)

Table 1. Comparison of Performance between RUINQ-G (positive RF) and RUINQ-R (positive & negative RF).

	RUINQ-G Mean (SD)	RUINQ-R Mean (SD)	Total Mean (SD)
Cycles	7.19 (5.30)	6.41 (3.77)	6.79 (4.59)
Instances entered	5.49 (3.74)	4.56 (3.45)	5.02 (3.61)
Terms in the first query	3.52 (3.41)	3.69 (2.98)	3.60 (3.19)
Terms entered by user in the last query	4.02 (4.22)	4.62 (4.16)	4.32 (4.18)
Full documents displayed	27.24 (16.30)	24.80 (12.15)	26.01 (14.35)
Titles shown	304.02 (339.69)	203.33 (250.65)	253.28 (301.24)

Table 2. Comparison of Searching Behavior between RUINQ-G (positive RF) and RUINQ-R (positive & negative RF)

each search, 253.28 unique titles were shown to the subjects, and 26.01 full documents were displayed. Table 2 summarizes the results of transaction logs according to the type of system used.

We analyzed the performance data with respect to the searching behaviors identified in Table 2. The only significant relationship was that greater user terms in the last query resulted in lower performance in terms of instance recall ($r = -.212$, $p<0.05$). The rest of searching behaviors were not correlated with performance results.

We compared the searching behaviors between high-performance subjects and low-performance subjects. A "high performance" subject is one whose mean instance recall is above the mean for all subjects; a "low performance subject is one whose mean instance recall is below that of the mean for all subjects. By this categorization, eight subjects were high-performance, and eight were low-performance.

Low-performance subjects entered more terms in their first query (M=4.21, SD=3.92), and entered more terms in the last query (M=5.25, SD=5.15) than high-performance subjects (M=2.98, SD=2.07 in the first query and M=3.38, SD=2.61 in the last query). These differences were statistically significant with $t(125)=2.21$, $p< .05$ in the number of terms entered in the first query, and $t(125)=2.57$, $p< .05$ in the number of terms entered by users (not chosen from term suggestion features) in the last query respectively. On the other hand, high-performance subjects saved more instances, (M=5.81, SD=4.35), and displayed more full documents (M=29.53, SD=15.40) than low-performance subjects (4.25 (SD=2.50) instances, 22.53 (SD=12.40) full documents). These differences were significant according to the t-test results at the level of 0.05: number of instances, $t(125)= -2.48$; number of full documents displayed, $t(125)= -2.82$. In addition, high-performance subjects saw more document titles (M=321.27, SD=369.14) than low-performance subjects (M=186.34, SD=195.50). This result was also significant, with $t(125) =-2.58$, $p<.01$.

3.2.2 Effectiveness of Term Suggestion

In RUINQ-G, which has positive RF only, the subjects chose 2.31 (SD=3.34) terms per search on average from the positive terms suggested by system. In approximately half (31 searches, 49.2%) of the total 63 searches using RUINQ-G, the subjects didn't choose any term from the positive terms suggested.

In RUINQ-R, which has positive and negative RF, the subjects chose 1.76 terms on average from the positive terms suggested and 1.97 terms from the negative terms suggested. In 30 searches (46.9%) out of 64 searches, no positive terms were chosen. In 39 searches (60.9%) out of 64 searches, the subjects didn't choose any negative terms for their queries.

Neither number of positive terms selected nor number of negative terms selected was signifIcantly correlated with instance recall. In the RUINQ-G system, high-performance subjects chose 2.06 terms, and low-performance subjects chose 2.56 terms. On the RUINQ-R system, the high-performance subjects selected 1.72 positive terms and 2.03 negative terms while low-performance subjects selected 1.81 positive terms and 1.91 negative terms. In the number of terms chosen, either negative terms or positive terms, there was no significant difference between high-performance subjects and low-performance subjects.

The effectiveness of the term suggestion feature was also investigated by analyzing data from subjects' self reports during the exit interview. The subjects were asked, "To what extent did you find the term suggestion feature useful during your searches? Why is that?" and "To what extent did the term suggestion feature improve your ability to identify different aspects of the topics? Why is that?" The usefulness of the term suggestion feature yielded an overall mean rating of 3.19, where 1 represented not at all useful

and 5 indicated completely useful. The extent to which the term suggestion feature was indicated to improve their ability to identify different aspects of the topics had a mean rating of 3.25, where 1 represented no improvement of ability and 5 indicated complete improvement of ability. Overall, the term suggestion feature was not viewed as highly useful.

The subjects who had the positive only system first seemed to be less positive about the term suggestion feature's usefulness than those subjects having the system with both positive and negative term suggestions first. The positive only first subjects complained about the type of words provided and the cognitive overhead in analyzing the terms. Although the positive only second group was more positive about the usefulness in general, it was mostly based on unintended advantages. These included not having to type the word in and providing a summary of the document. Most felt they could think of better words than those provided.

The comments from the positive only first group regarding the feature's contribution to improving their ability to identify aspects were generally high or low with few in between. The subjects either did not like the feature at all or thought it was really useful. The ones who discussed it as useful generally mentioned the bad term suggestion advantages or the summary overview the good terms provide. The positive only second group were generally more positive, but several just did not want to use it for the intended purpose. As with the other group, it helped some get more items, but others used it as summary information.

The comparison of subjects based on performance indicated few differences in effectiveness of the term suggestion feature. The effectiveness of the system in terms of being useful during the searches had a mean ratings of 3.13 and 3.25 for the high performers and low performers respectively (on a 5-point scale where 5 is highest). The effectiveness of the system was also not different for the two groups on the question of the feature improving their ability to identify aspects. Both high performers and low performers had a mean ratings of 3.25 (on a 5-point scale where 5 is highest).

3.3 Usability

3.3.1 Usability of the systems in general

Subjects were asked to consider their search experience following four searches with each system. Using a five-point scale where 1=not at all; 3=somewhat; and 5=extremely, subjects were asked to answer 3 questions: how easy it was to learn to use the system; how easy it was to use the system; and how well they understood how to use the system. The results yielded no significant difference between subjects' mean ratings on these questions for RUINQ-G and for RUINQ-R. These ratings are displayed in Table 3.

These same questions were also looked at according to system order, block order and performance status. Again, there was no significant difference for any of these conditions. However, for system order, it does appear that subjects in system order RUINQ-G/RUINQ-R consistently prefer system RUINQ-G to system RUINQ-R while subjects in system order

	RUINQ-G Mean (SD)	RUINQ-R Mean (SD)
Easy to learn to use	4.00 (0.65)	3.81 (0.98)
Easy to use	3.86 (0.74)	3.56 (1.09)
Understand how to use	3.66 (0.72)	3.37 (1.08)

Table 3. Usability of RUINQ-G (positive RF) versus RUINQ-R (positive & negative RF)

	System Order			
	RUINQ-G/ RUINQ-R		RUINQ-R/ RUINQ-G	
	RUINQ-G Mean (SD)	RUINQ-R Mean (SD)	RUINQ-G Mean (SD)	RUINQ-R Mean (SD)
Easy to learn to use	4.0 (.53)	3.37 (1.18)	4.0 (.81)	4.25 (.46)
Easy to use	4.0 (.53)	3.25 (1.28)	3.7 (.95)	3.87 (.83)
Understand how to use	4.0 (.53)	3.12 (1.13)	3.2 (.76)	3.62 (1.06)

Table 4. Usability of RUINQ-G (positive RF) versus RUINQ-R (positive & negative RF) according to System Order

Preferred System:	RUINQ-G N (%)	RUINQ-R N (%)	Total[a] N
Easier to learn to use	12 (80%)	3 (20%)	15
Easier to use	10 (71%)	4 (29%)	14[b]
Liked best	7 (50%)	7 (50%)	14[b]

Table 5. Number (percentage) of subjects preferring one system over the other
Note: [a] one missing data point in each case; [b] one subject rated systems equal

RUINQ-R/RUINQ-G consistently prefer system RUINQ-R. Unfortunately, the low number of subjects did not permit us to do further analysis on this data. These results are displayed in Table 4.

fter subjects had completed all searching, they were asked to compare RUINQ-G and RUINQ-R in an exit questionnaire. Questions included: how easy they were to learn to use; how easy they were to use; and which system the subjects liked best. Subjects were asked to place a "1" next to the easier/liked best system and a "2" next to the more difficult/liked least system. The results are displayed in Table 5. There is a significant difference between which system was easier to learn to use. RUINQ-G was rated as significantly easier to learn to use χ^2 (1, N=15)=5.4, $p < .05$. Although RUINQ-G also appears to be the easier system to use, there was no significant difference in subject ratings. For system preference, or which system subjects *liked best*, RUINQ-G and RUINQ-R each received an equal number of preferences. Although not part of our instructions, one subject rated the two systems as equally likable and equally easy to use.

Subject ratings from the exit questionnaire were then analyzed in terms of system order, block order and high/low status. The results of these analyses are displayed in Tables 6, 7 and 8, respectively. The results for system order indicate that RUINQ-G was easier to learn to use and easier to use. Subjects' rankings of which system they liked best with respect to system order revealed that the majority of subjects in system order RUINQ-G/RUINQ-R ranked RUINQ-R as the most likable system and the majority of subjects in system order RUINQ-R/RUINQ-G ranked RUINQ-G as the most likable system. These results indicate that subjects may have a preference for whichever system that they used last.

When subjects' rankings of which system was easier to learn to use were examined in regard to block order, the results indicate that RUINQ-G was easier to learn to use regardless of block order. For ease of use rankings and block order, the results show that the majority of subjects in both block orders ranked RUINQ-G as the easier system to use. The likeness rankings for RUINQ-G and RUINQ-R revealed that while the majority of the subjects in Block Order 1 ranked RUINQ-G as being the most likable system, the majority of subjects in Block Order 2 ranked RUINQ-R as being the most likable system. The equal and missing data in this category prevented us from concluding that this finding was significant.

When the subjects' rankings of which system was easier to learn to use were examined in regard to performance status, the results indicate that RUINQ-G was easier to learn to use regardless of performance status. The ease of use ratings were similar across both high and low performers: 5 of the high performers ranked RUINQ-G as being the easier system to use and 5 of low performers ranked RUINQ-G as the easier system to use. However, the majority of the high performers liked RUINQ-R best, while the majority of the low performers liked RUING-G best.

System Order:	RUINQ-G/RUINQ-R			RUINQ-R/RUINQ-G		
Preferred System:	RUINQ-G N=	RUINQ-R N=	Total	RUINQ-G N=	RUINQ-R N=	Total[a]
Easier to learn to use	6 (75%)	2 (25%)	8	6 (86%)	1 (14%)	7
Easier to use	5 (63%)	3 (37%)	8	5 (83%)	1 (17%)	6[b]
Liked best	3 (37%)	5 (63%)	8	4 (67%)	2 (33%)	6[b]

Table 6. Number (percentage) of system preferences according to System Order
Note: [a] one missing data point in each case; [b] one subject rated systems equal

Block Order:	Block Order 1			Block Order 2		
Preferred System:	RUINQ-G N=	RUINQ-R N=	Total	RUINQ-G N=	RUINQ-R N=	Total[a]
Easier to learn to use	7 (88%)	1 (12%)	8	5 (71%)	2 (29%)	7
Easier to use	6 (75%)	2 (25%)	8	4 (67%)	2 (33%)	6[b]
Liked best	6 (75%)	2 (25%)	8	1 (17%)	5 (83%)	6[b]

Table 7. Number (percentage) of system preferences according to Block Order
Note: [a] one missing data point in each case; [b] one subject rated systems equal

Performance Status:	High Performers			Low Performers		
Preferred System:	RUINQ-G N=	RUINQ-R N=	Total	RUINQ-G N=	RUINQ-R N=	Total[a]
Easier to learn to use	7 (88%)	1 (12%)	8	5 (71%)	2 (29%)	7
Easier to use	5 (71%)	2 (29%)	7[b]	5 (71%)	2 (29%)	7
Liked best	2 (29%)	5 (71%)	7[b]	5 (71%)	2 (29%)	7

Table 8. Number (percentage) of system preference according to Performance Status
Note: [a] one missing data point in each case; [b] one subject rated systems equal

3.3.2 Use and Usability of Term Suggestion

Subjects were asked to rate their understanding of the term suggestion feature and their use of the term suggestion feature to modify their searches using a five-point scale where 1=not at all; 3=somewhat; and 5=extremely. The results from these questions are displayed in Table 9. Overall, subjects rated their understanding of term suggestion with mean = 3.78. Subjects in system order RUINQ-G/RUINQ-R rated their understanding of term suggestion significantly higher (M=4.25) than those subjects in system order RUINQ-R/RUINQ-G (M=3.17), with $t(12)=2.16$, $p<0.05$. Block order and performance status had no significant effect with respect to understanding of term suggestion.

Overall, subjects rated their use of the term suggestion feature to modify their searches with a mean of 3.32. Subjects in system order RUINQ-G/RUINQ-R did not respond much differently than those in system order RUINQ-R/RUINQ-G (G: M=3.5; R: M=3.1). Subjects in Block Order 1 rated this question significantly lower (M=2.81) than those subjects in Block Order 2 (M=4.0), with $t(12)=.019$, $p<0.05$. The ratings on this question for high and low performers were very similar (M=3.42; M=3.21).

The usability of the term suggestion feature was specifically addressed in the exit interview. The subjects were asked, "To what extent did you understand how to use the term suggestion feature? Why is that?" and "To what extent did you use the term suggestion feature to modify your searches? Why is that?" Overall, the mean rating for understanding the feature was 3.84 where 1 represented no understanding and 5 indicated complete understanding. The overall mean rating for using the feature was 3.25, where 1 represented no use and 5 indicated complete use. Generally, subjects described their understanding of the feature in terms of a synonym suggestion tool. The use of the feature varied in the way it was used. The negative term suggestion feature was generally used to constrain the documents retrieved, however the positive term suggestion feature was often used to get an overview of the document or to find synonyms.

The subjects who used the feature with positive only term suggestion before using it with both positive and negative, had a higher mean rating for their understanding of the system than those who used the systems in the reverse order, 4.25 and 3.44 respectively. The subjects in the positive only first system order who gave the lowest ratings indicated they had a tendency to mark documents as good or bad randomly looking for certain words or ideas, which were not found. The higher ratings within this subject group were given by subjects who discussed the term suggestion feature as synonyms or words to refine the query. The subjects using the positive only system first had a higher mean rating for the amount of feature use than those using that system second, 3.5 and 2.9 respectively. The lower ratings from this group came from subjects who expressed a distrust of the bad term suggestion feature, because they suggested it was unclear how it worked. Higher ratings came from subjects who indicated that they used negative terms to eliminate documents already seen.

	Overall	System Order		Block Order		Performance Status	
		GR	G	Block 1	Block 2	High	Low
Understanding of term suggestion feature	3.78	4.25	3.17	3.5	4.17	3.42	4.14
Use of term suggestion feature to modify searches	3.32	3.5	3.1	2.81	4.0	3.42	3.21

Table 9. Usability/Use of Term Suggestion Feature

The subject group using the positive only system after the system with both positive and negative term suggestions, mostly complained about the use of the feature rather than addressing how well they understood it. When the use of the feature was discussed it was as providing alternate terms and further direction for honing the search. The group indicated a moderate use of the feature. They discussed the bad term suggestion feature as being used in the way intended. However, the good term suggestion was not viewed as particularly helpful, but the highest rater used it to scan the content of the documents.

The usability of the system in terms of being able to understand how to use the terms suggestion feature had a mean rating of 3.38 for the high performers and 4.13 for the low performers (on a 5-point scale where 5 is highest). The better performers purported to understand the feature less well. However, there was no difference in usability ratings in terms of how much they indicated using the term suggestion feature. High performers had a mean rating of 3.25 and low performers had a mean rating of 3.25 (on a 5-point scale where 5 is highest). Interestingly, high performers generally had more comments than low performers.

3.4 Subject Characteristics and Effectiveness and Usability

None of the demographic characteristics that were observed, including performance on the FA-1 Controlled Associations Test, was significantly related to any of the explicit performance or usability measures discussed above. But we should note that this was a relatively homogeneous population.

4.0 Discussion

Our initial hypothesis that a system providing both positive and negative RF would perform better than one that offered positive RF only was not supported by our experiment. In our study, the two systems performed no differently on instance recall. We might conclude, therefore, that there appears to be no benefit of negative RF, on the task presented to our subjects.

However, the self-report and interview data provide us with a broader picture of subjects' uses and understandings of the systems and the features offered, which helps to explain the performance results we observed. These data suggest that subjects had difficulty conceptualizing how to use the *term – suggestion* feature, in particular. Because of their unfamiliarity with negative RF, many subjects distrusted the "bad term" suggestion feature and hesitated to use it. Furthermore, when subjects directly compared the two systems, they rated the system with positive only RF as being easier to learn, and easier to use. But at the same time, the subjects also rated both systems evenly in terms of overall preference, with respect to the task.

A significant finding of our study is the inverse relationship between number of query terms and performance. At the moment we have no explanation of this result, nor do we have any idea about causal direction. Clearly, further investigation of the actual interactive behavior of the searchers is needed in order to understand this result.

5.0 Conclusions

Once again, it appears that we have demonstrated that incorporation of negative RF in a system which offers RF as a term-suggestion device, does no harm, with respect to performance on the "instances" task. On the face of it, this is not a terribly exciting result. However, this result, taken in combination with the usability results, does suggest that making negative RF more learnable and more usable than in our current system could lead to a performance advantage over positive RF only systems. This further suggests that much more research on appropriate conceptual models of RF, and on interfaces to support interaction with RF needs to be done before the more general issue can be resolved.

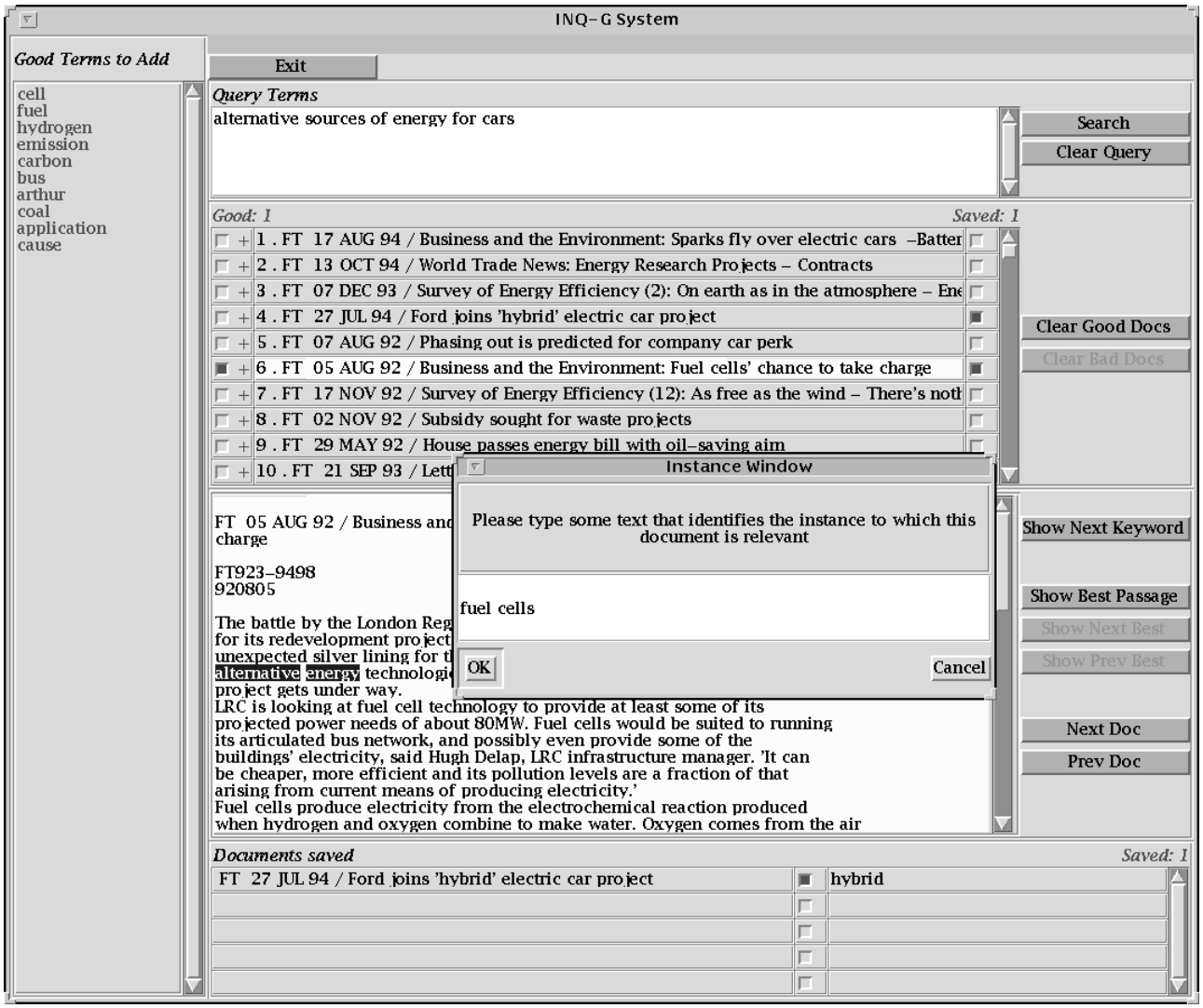
RF as term suggestion appeared also not to be well

understood by our subjects, nor effectively used, in this implementation. Although some subjects understood it, and used it, as a query enhancement (i.e. new term finding) device, the fact that many understood it as a synonym device is unfortunate. This result could be explained by the nature of our subject pool, and also (more whimsically) by the speculation that the FA-1 test might have conditioned them to think in terms of synonyms, it is clear that a better conceptual model of RF as term suggestion needs to be developed.

Finally, the counter-intuitive result that fewer query terms led to better performance needs to be investigated in much more detail. The result could be an artifact of the task itself, which would in itself be of considerable interest. But to understand this result will require detailed analysis of the interactions themselves, from a variety of points of view. An intriguing possibility that this result suggests is that interactive IR may work in much different ways than automatic, and that we may need to reconceptualize our understandings of what constitute "good" ways to do IR. We hope that our further analyses of these data, and related experiments, will shed light on this question.

6.0 References

Belkin, N.J. et al. (1998) Rutgers' TREC-6 interactive track experience. In D. Harman and E. Voorhees, eds. TREC-6. Proceedings of the sixth Text Retrieval Conference. Washington, D.C.: GPO.

Koenemann, J. (1996) *Relevance feedback: Usage, usability, utility*. Unpublished PhD dissertation, Department of Psychology, Rutgers University.

APPENDIX A. Screen Dump of RUINQ-R