Rutgers' TREC-6 Interactive Track Experience

N.J. Belkin, J. Perez Carballo, C. Cool[°] , S. Lin, S.Y. Park,
S.Y. Rieh, P. Savage, C. Sikora, H. Xie and J. Allan[*]

School of Communication, Information & Library Studies
Rutgers University
4 Huntington Street
New Brunswick, NJ 08901-1071

Abstract

The goal of the Rutgers TREC-6 Interactive Track study was to compare the performance and usability of a system offering positive relevance feedback with one offering positive and negative relevance feedback. Our hypothesis was that the latter system would better support the aspect identification task than the former. Although aspectual recall was higher for the system supporting both kinds of relevance feedback (0.53 vs. 0.46), the difference was not significant, possibly because of the small number of subjects (four in each condition, each doing three searches). Usability results were also equivocal, perhaps due to the complexity of the system. Compared to ZPRISE, the control system without relevance feedback, both relevance feedback systems were rated more difficult to learn to use, but more effective.

1. Introduction

The focus of the Rutgers TREC-6 Interactive Track study was investigating the effectiveness and usability of negative relevance feedback (RF) in interactive information retrieval (IR). This followed from the results of our TREC-4 and TREC-5 studies, in which our subjects expressed a desire to be able to control retrieval in order to suppress documents they did not like. This led us to hypothesize that supporting negative as well as positive relevance judgments in interactive IR would lead to improvements in performance of various IR tasks. These results, and the manner of implementation of negative RF which we think follow from them, are reported in Cool, Belkin & Koenemann (1996).

Briefly, we suggest that there are basically two ways in which negative relevance judgments can be understood in the context of automatic RF. The first, which we call the "classical" model, assumes that terms which appear in the query and positively judged documents, and also in negatively judged documents are "poor" terms from the point of view of IR, since they are bad discriminators. This model therefore reduces the query-term weight of such terms until they reach zero weight, when they are removed from the query. Experiments in non-interactive environments have shown that using negative weights decreases performance. In this model, there is generally no account taken of terms which appear only in negatively judged documents.

[°] Graduate School of Library & Information Studies, Queens College, CUNY

[*] Center for Intelligent Information Retrieval, Department of Computer Science, University of Massachusetts at Amherst.

In contrast to the classic model, we propose an alternative interpretation, which assumes that terms in the original query, and which are added to that query through positive RF are "good" terms whether or not they also appear in negatively judged documents, since they are indicators of what the searcher is looking for. The meaning of the negative judgment in this model is understood to be that the *context* in which the good terms appear is inappropriate to the searcher's problem, or that the topic which they represent is treated only peripherally, or from an inappropriate point of view. Thus, important terms in the negatively judged documents, which do not appear in positively judged documents, are understood as indicators of the inappropriate context, or the main topic, or the inappropriate point of view. This model thus leads us to a quite different way to implement RF, in which query terms which appear in positively judged documents (irrespective of their appearance in negatively judged documents) have their query-term weights increased, and in which the query is expanded by both the important terms in the positively judged documents (with positive weights) and by the important terms in the negatively judged documents which do not appear in the query or the positively judged documents (with negative weights).

The TREC-6 Interactive Track task of identifying the different aspects of a topic offers an especially good environment to investigate the effectiveness of the type of RF our model suggests. Our hypothesis is that once a searcher has identified some aspect of a topic in a particular document, a negative relevance judgment on that document will depress the retrieval status value (RSV) of other documents on the topic which treat that specific aspect, thus promoting documents which treat different aspects of the topic in the output ranking, making it easier to find these new aspects. On the other hand, positive relevance judgments will tend to increase the RSV of other documents which treat the same aspect of the topic, thus demoting documents treating different aspects of the topic in the output ranking, making it more difficult to find new aspects.

Following the results of Koenemann (1996) and Koenemann & Belkin (1996), which suggest that user control of RF leads to enhanced performance and usability, we implemented RF in both of our experimental systems as a term-suggestion device for query expansion, rather than as an automatic query modification device. Thus, the terms which would be added through automatic RF were displayed to the searcher as each relevance judgment was made, for the searcher to choose from for adding to the query (as either a positive or a negative term). The interface and the details of the implementation are described more fully in section 2.

We ran our experiment according to the TREC-6 Interactive Track protocol, with four subjects searching on the control system (ZPRISE) and the positive RF system (ruinq1), and four subjects searching on the control system (ZPRISE) and the positive plus negative RF system (ruinq2). Unfortunately, we goofed and did not log the ZPRISE searches in either of these conditions. Thus, although we can compare subjective judgments of the three systems, we cannot compare performance of either of our experimental systems with the experimental systems of the other participants in the Interactive Track, since they can be strictly compared only through *differences* in performance on the control and experimental system(s) at each site, not the absolute performance on any measure.

2.0 Methods

In this section we describe the research methods we used in conducting our experiment, along with a description of the systems themselves.

2.1 Research Methods

Eight volunteer searchers were recruited to participate in the study, from the population of students in the School of Communication, Information and Library Studies at Rutgers University and from the larger community of information professionals in the New Jersey area. As a condition of the study, none of the participants had taken part in previous TREC studies and none had any prior experience with either RU-INQUERY or ZPRISE. The general demographic characteristics of the searchers and their experiences with IR systems are described below in section 3.1.

Each searcher performed six searches on six topics: three of the searches on a control system (ZPRISE) and three on an experimental system (RU-INQUERY). Searchers were alternately assigned to one of two versions of the RU-INQUERY system. Version 1 (E1) offered positive relevance feedback only; while version 2 (E2) offered both positive and negative relevance feedback. Using ZPRISE as a control system (C), the searchers were randomly assigned to one of the following conditions: E1 and C; C and E1; E2 and C; C and E2. We replicated the conditions twice, using a single ordering of topics.

Before conducting their searches, participants completed a self-administered questionnaire which asked about their demographic characteristics and their searching experiences with a variety of IR systems. They then received a 20-minute interactive tutorial for each of the IR systems, prior to searching on them. Searchers were given 20 minutes to conduct each of their six searches. After each search, subjects answered several questions about their familiarity with the search topic, experiences with the searching task, and their satisfaction with the results. Each search was videotaped, and computer logged. Participants were instructed to "think aloud" about what they were doing, and why, as they searched and these verbal protocols were captured on the videotapes. This process was repeated six times, across the two systems. At the end of this entire session, searchers completed an exit interview which focused on their understanding and use of relevance feedback; their perceptions of the utility of RF for the aspect retrieval task; and their experiences with the IR systems.

2.2. Systems

We used InQuery 3.1p1 as the basis for our experimental systems and the ZPRISE Interactive Track Release as the control system. The two versions of InQuery are: 1) the positive relevance feedback only system (ruinq1); and 2) the positive and negative relevance feedback system (ruinq2). Both of these used the default indexing of InQuery 3.1p1, the Porter stemmer, and the default weighting and matching functions. User query formulation was restricted to unstructured queries, plus the phrase operator (instantiated by enclosing the phrase words within double quotes). RF query expansion (for both positive and negative RF) was implemented using the default InQuery 3.1p1 term ranking formula (tf*rdf), with the number of suggested terms determined by the formula:

$5n + 5$, where n = number of judged documents

to a maximum of 25 suggested terms. The query was parsed as a weighted sum, using the default weighting for RF term addition for positive terms, and adding the negative terms under the InQuery "NOT" operator, with 0.6 weight. Appendix A is a screen dump of the ruinq2 interface; the ruinq1 interface is identical, except that the frames in the lower left and upper right of the interface (those having to do with negative term suggestion, and negative term addition, respectively) are removed, and there are no negative RF buttons.

The functions that are offered by the systems are:

1. Unstructured query input plus phrases in the query formulation window (top center frame);
2. Saving, clearing and loading queries;
3. Display of rank, date and title of ten retrieved documents at a time (center frame);
4. Scrolling the title display ten documents at a time;
5. Saving a document to indicate one or more aspects - unsaving by clicking on saved document button (right hand button on the title line);
6. Marking a document relevant or not relevant to get term suggestions - unmarking by clicking on relevant or nonrelevant document button (two left buttons on the title line). Unmarking removes the document from the RF pool and thus changes the appropriate term suggestion display, but does not affect the selected terms;
7. Display of suggested RF query expansion terms (positive terms displayed in upper leftmost frame; negative terms displayed in lower leftmost frame);
8. User selection of suggested terms to be added to the query by clicking on the desired term (displayed in the top rightmost frame for negative terms, the immediately adjacent frame for positive terms);
9. User deselection of RF terms by clicking on the desired term in the appropriate selected term frame (deselected terms returned to the appropriate term suggestion frame);
10. Clearing all relevance markings (removes all term suggestions, but not term selections);
11. Displaying the full text of a document by double clicking on the title line (displayed in the bottom center frame);
12. Scrolling through the full text of the document;
13. Highlighting query terms in the full text display;
14. Scrolling directly to the next query term in full text display (Show Next Keyword);
15. Showing the best (next best, previous best) passage in the full text display, according to default InQuery 3.1p1 method;
16. Displaying the full text of the next document or the previous document in the retrieved list.

Marking a document saved (unsaved) and relevant or not relevant (or unmarking) is indicated by toggling change in color of the relevant button. Relevant was indicated by green, not relevant by red, and the terms in the term suggestion and selected terms frames were in the same colors.

All three systems ran on a SUN Ultra 140 with 64MB memory and 9GB disk under Solaris 2.5.1, using a 17" color monitor.

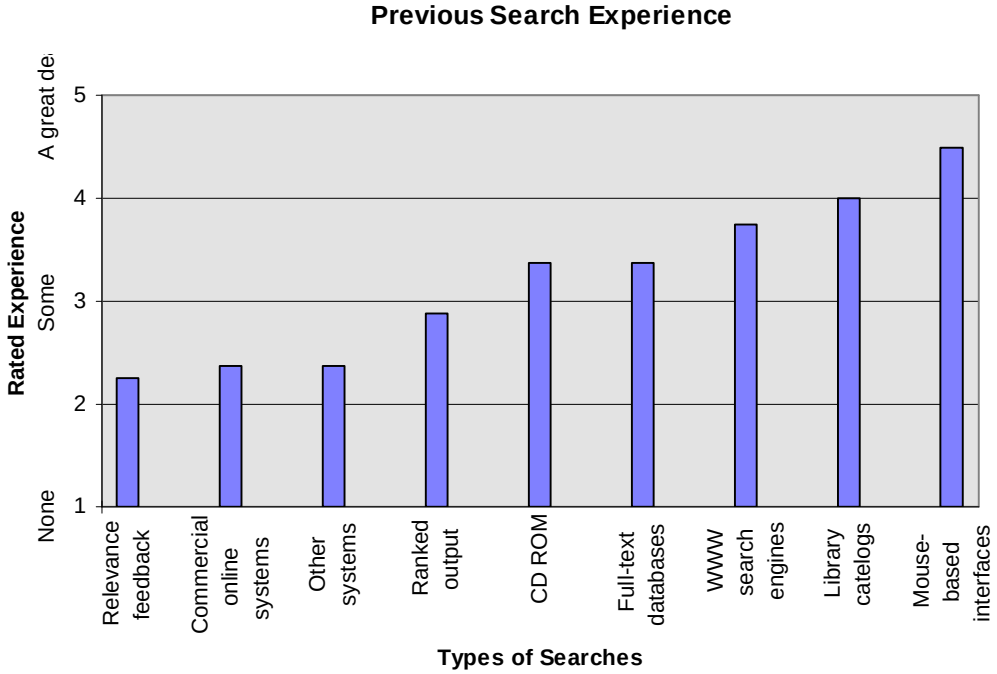
3.0 Results

In the following sections we report on our analyses of the questionnaire and interview data, followed by a discussion of our performance results.

3.1 Characteristics of the Searchers

Our subject group included 6 females and 2 males. The subjects were distributed fairly evenly across age categories with the youngest being under 21 and the oldest between 51 and 60. Five of the eight subjects had, or were pursuing, a graduate degree in library science. The other three subjects indicated no education in library science and had, or were pursuing, Bachelor degrees in other fields. As mentioned above, none of the subjects reported participating in any previous TREC experiments or having any previous experience with the ZPRISE or RU-INQUERY information retrieval systems. The median number of years reported for overall experience doing online searching was three and a half (M = 4.6, SD = 3.04). The minimum amount of experience reported was one and a half years, while the maximum amount was 10 years.

Figure 1: Mean previous search experience on different systems reported by subjects.



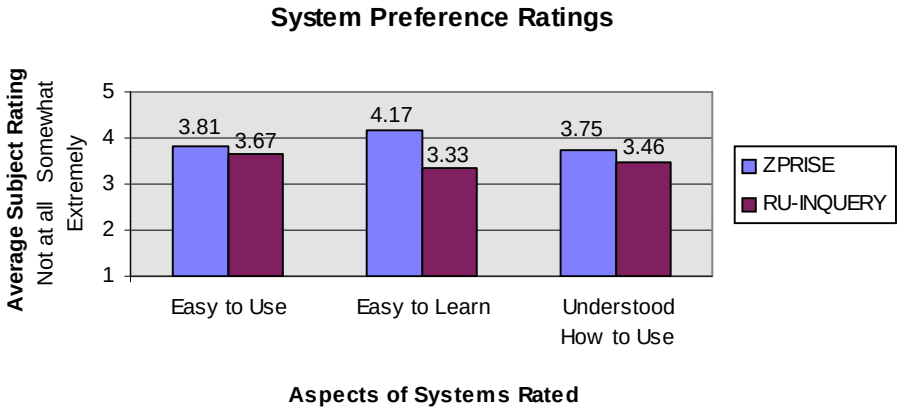
As can be seen in Figure 1, the average amount of previous search experience on different types of systems varied widely. Using a five point scale ranging from none to a great deal of experience, the most experience was reported using mouse-based interfaces (M = 4.5, SD = 1.07). Relevance feedback had the lowest reported experience ratings (M = 2.25, SD = 1.28). Surprisingly, the average rating for experience on commercial online systems, such as Dialog, Lexis and BRS Afterdark, was fairly low (M = 2.38. SD = 1.19). Otherwise, the subjects reported having a fair amount of experience on each of the different system types (ranked output systems, M = 2.87, SD = 1.25; CD ROM, M = 3.38, SD = .52; full-text databases, M = 3.38, SD = 1.06; WWW search engines, M = 3.75, SD = 1.28; library catalogs, M = 4.0, SD = .76).

3.2 Subjective Ratings of Searchers

After each search the subjects provided subjective ratings related to the specific search and to the system they used. These ratings were made based on a 5-point scale where 1 was "not at all," 3 was "somewhat" and 5 was "extremely." Collapsing across topics and systems, the 8 subjects felt they were mildly to moderately familiar with the topics (M = 2.46, SD = .56), that the searches on the topics were moderately easy (M = 3.42, SD = .70), that they were somewhat satisfied with the results (M = 3.06, SD = .88) and somewhat confident that they identified all the possible aspects for the topic (M = 2.92, SD = .94). There was more variability on responses to whether subjects felt they had sufficient time to do an effective search, although it was moderately high (M = 3.60, SD = 1.27).

Although there were significant correlations between each pair of subjective performance measures, there was no significant correlation between rated familiarity with the search topic and confidence with the search, ease of searching, satisfaction with the search nor with sufficiency of time for the search (r_{pb} = -.04, ns; r_{pb} = .08, ns; r_{pb} = -.007, ns; r_{pb} = .002, ns, respectively). This supports the assumption that variability in subject familiarity with the search topics should not strongly impact the findings of the study. System order was also evaluated by comparing the average responses on the subjective performance measures from the first system used to the second system used. No significant differences were identified (ease of search, t (7)= -.92, ns; confidence in search results, t (7) = -.50, ns; satisfaction with search, t (7) = -.92, ns; sufficient time for search, t (7) = -1.52, ns). The order in which the subjects used the systems did not significantly influence their subjective ratings of their search performance.

Figure 2: Subject ratings of ZPRISE and RU-INQUERY across search topic on ease of use, learning and understandability. (Note: N = 8)



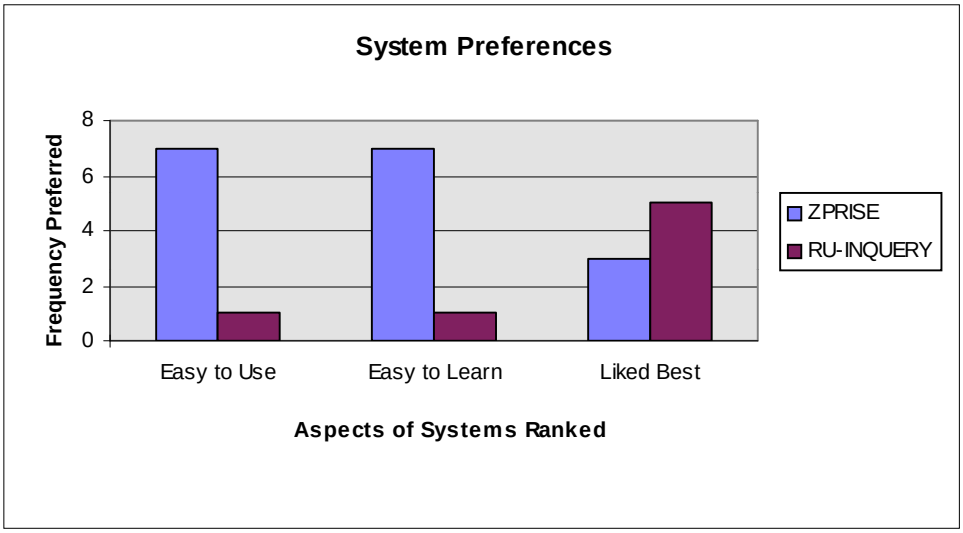
1= Not at all
3= Somewhat
5= Completely

Overall, the subjects rated both systems above average on ease of use, learning and ability to understand how to use it (see Figure 2). The five-point scale used "not at all," "somewhat" and

"extremely" as anchors. The average rating on each of these areas was higher for ZPRISE than for RU-INQUERY. Indeed, the highest average rating for RU-INQUERY is still lower than the lowest average rating for ZPRISE. When ratings of the RU-INQUERY systems are evaluated separately, it is clear that the 4 subjects using both positive and negative relevance feedback rated the system higher and more consistently (easy to use, M = 4.0, SD = .72; easy to learn, M = 3.67, SD = .27; understand how to use, M = 3.67, SD = .98) than those only receiving positive relevance feedback (easy to use, M = 3.33, SD = 1.19; easy to learn, M = 3.0, SD = 1.27; understand how to use, M = 3.25, SD = 1.28). When comparing the average ratings of the 4 subjects using the higher ranking RU-INQUERY to the average ratings of the 8 subjects on ZPRISE, RU-INQUERY has a slightly higher average rating for ease of use, but remains lower on ease of learning and understanding how to use it.

Subjects provided additional subjective ratings, relative to their overall experience, after doing all 6 searches on the two systems. On a five point scale where 1 is "not at all", 3 is "somewhat" and 5 is "completely", on average, subjects rated their understanding of the task very highly (M = 4.0, SD = 1.07). They rated the search tasks in the study moderately similar to searching tasks they typically perform (M = 3.5, SD = .93). They rated ZPRISE as somewhat different compared to the RU-INQUERY system that they worked on in the study (M = 3.38, SD = .74). When comparing ZPRISE to the RU-INQUERY system on ease of use, 7 of the 8 subjects identified ZPRISE as easier to use. The one subject choosing RU-INQUERY as the easier system was using the version with only positive relevance feedback. Similarly, 7 of the eight subjects identified ZPRISE as the system easier to learn and again the one subject choosing RU-INQUERY had no negative relevance feedback. Interestingly, however, even with the preponderance of subjects identifying ZPRISE as easier to learn and use, only 3 of the 8 subjects selected ZPRISE as the system they liked best. This is illustrated in Figure 3. Three of the four subjects using the positive relevance feedback only version and 2 of the 4 subjects using the version with both negative and positive relevance feedback selected RU-INQUERY as the best system compared to ZPRISE.

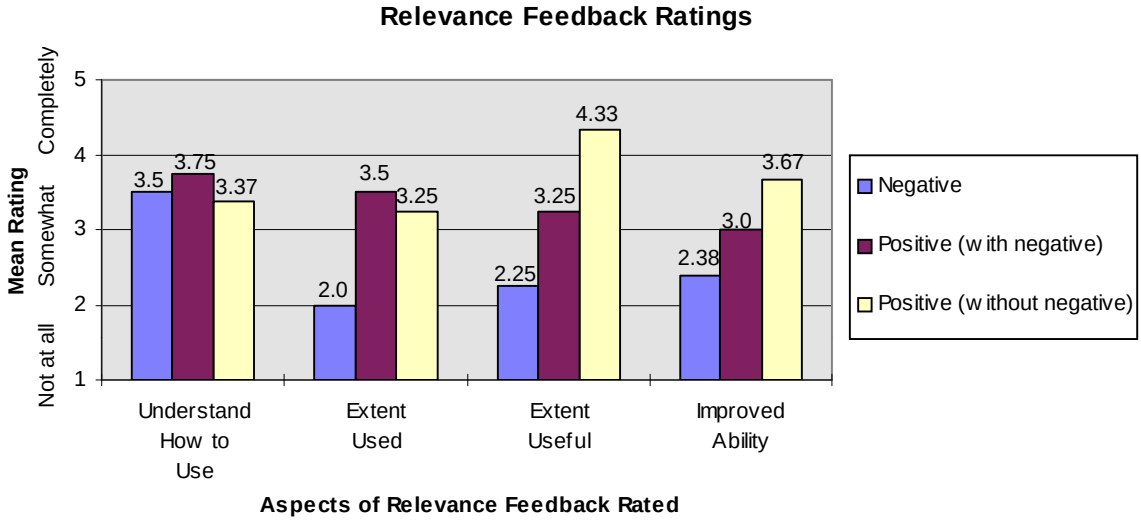
Figure 3: The frequency with which subjects preferred ZPRISE or RU-INQUERY on each of three system aspects.



Subjects provided comments about the two systems, during the exit interview (discussed below in section 3.4). The comments provide a better perspective on their inconsistent ranking of the two systems. It is clear that the subjects felt that RU-INQUERY was powerful and flexible. However, they would forget how to use all the different features or become confused. ZPRISE was seen as simpler and less sophisticated.

Subjects also provided subjective ratings regarding the use of relevance feedback, ranked output and system suggested terms. Subjects responded on a 5-point scale (1 = "not at all," 3 = "somewhat," 5 = "completely"), to questions exploring the extent to which relevance feedback was understood, used, found useful and thought to improve search abilities. The subjects using the version of RU-INQUERY with both positive and negative relevance feedback, provided a response for each type of relevance feedback. The mean ratings for each type of relevance feedback can be seen in Figure 4. The average rating for understanding how to use relevance feedback was above the midpoint of the scale. However, the variance in responses was much greater for subjects only rating positive relevance feedback (positive only SD =1.79, positive SD = .96, negative SD = .58). Subjects were less likely to use negative than positive relevance feedback. When negative relevance feedback was not available, positive relevance feedback was generally rated higher for usefulness and as improving ability to identify different aspects of the topic. Negative relevance feedback was not reported to be very useful or considered to improve searchers' ability. Using the same scale, subjects rated the extent to which ranked output and system suggested terms were useful. Generally, subjects found both features to be moderately to highly useful (ranked output, M = 4.13, SD = .84; system terms, M = 3.71, SD = 1.11).

Figure 4: Average subject ratings on different aspects of subjective responses to using relevance feedback. (Note: n = 4).



3.3 Characteristics of the Interactions

We are primarily concerned with differences in measures of interaction between ruinq1 and

ruinq2 which indicate different aspects of usability of the systems. There appears to be no significant difference in the time taken to learn to use the two systems (ruinq1 tutorial mean of 1612.25 seconds, ruinq2 tutorial mean of 1593.5 seconds). Nor is there any significant difference in the time taken per search (ruinq1 mean of 1096.25 seconds, ruinq2 mean of 1192.08 seconds), which seems to indicate that they are equally easy (or difficult) to use. However, differences do arise in other measures of the interaction.

The numbers of iterations, or cycles, per search are quite different (mean for ruinq1 8.92, mean for ruinq2, 5.17), which means that although the total time is more or less the same for both systems, time per cycle is greater for ruinq2. This may also be related to the large difference between the two systems in the numbers of full texts viewed (ruinq1 mean of 25.17, ruinq2 mean of 52.17), and in titles viewed (ruinq1 mean of 378.08, ruinq2 mean of 205.25), which suggests that searchers in the ruinq2 condition spent much more time reading texts than those in the ruinq1 condition, while those in the latter spent more time scrollling through the retrieved document list. There seems to be no obvious relationship between the different features of the two systems, and these differences in behavior, although futher analysis, particularly of the thinking aloud data, may help to explain them. The searchers in the ruinq2 condition made more use of relevance feedback terms (ruinq1 mean of 7.42, ruinq2 mean for positive terms of 11.08, and for negative terms of 4), which effect is heightened since there seems to be no great difference in the number of positively marked documents in the two conditions (ruinq1 mean of 4.25, ruinq2 mean of 5).

Overall, on these quantitative measures of the interaction, although there are some evident differences between behavior in the two systems, they are not easily explained by the presence or absence of support for negative RF, and may be the result of searcher differences rather than system differences.

3.4 Exit Interview Data

During the exit interview, searchers discussed their experiences using relevance feedback. Almost all of the subjects said they understood how to use relevance feedback, at least to some extent. However, as mentioned above, subjects were less likely to use negative than positive relevance feedback; and when negative relevance feedback was not available, positive relevance feedback was generally rated higher for usefulness and for improving ability to identify different aspects of the topic, than negative RF when it was available.

The following are some of the reasons searchers gave for finding *positive relevance* feedback helpful:

1.Positive RF Helps to Identify Relevant Terms and Aspects

For the aspect task, searchers were required to identify as many aspects of the topic as possible. Positive feedback reportedly helped to ensure that all of the relevant terms had been covered. As one searcher told us during the interview, "It helps me like a thesaurus would help me to make sure I'm covering all different terms." (S002)

2. Positive RF Helps to Assist Learning

Positive relevance feedback appears to be helpful in assisting searchers to think not only about what is missing in the search, but also about what he or she is doing in the search process. For example, "There were things as we went through relevant articles, there were things that I that I

just didn't think of at first and its almost like brainstorming. It sort of prompted me to think a little more about exactly what I was doing, and not only did I use them by adding them into my query, it also helped me by knowing what I didn't want to add to my query as well." (S006)

3. Positive RF Helps to Save Time

For this task, searchers were required to finish each search within 20 minutes. Positive relevance feedback allowed them to identify the relevant documents without reading the whole article. For example: "...it allows you to zero in quickly on the ones that would be useful without having to read through the article." (S001)

4. Positive RF Reduces the Retrieved Set Size

One of the reasons that searchers in our study liked positive feedback was that it seemed to keep the set of retrieved documents smaller, thereby making searching more efficient. The following searcher expressed this feeling: "That was useful cause it keeps the pile getting smaller." (S007)

Our exit interview data reveal several reasons why *negative relevance* feedback was difficult to use, or was *not* helpful:

1. Negative RF Sorts Out Relevant or Partially Relevant Documents

The major concern that our searchers had about using negative feedback was that it might sort out, or eliminate, some articles they would want to look at. For example, "The only problem is that its kind of difficult, because some of the articles you want partially, but you say you already have something similar and you don't want anything else to do with that specific topic. You really can't put negative on it because then it might sort out some other articles that you may want." (S003)

2. Negative RF Reduces the Rank Position of Relevant Documents

Another concern about using negative relevance expressed by our searchers was that it might push back relevant documents on the ranked list. As this searcher told us, negative RF was not helpful "because it pushes them back, and maybe those are articles you wanted..." (S003)

3. Negative RF is Difficult to Use Under Pressure

Some of our subjects found negative relevance feedback difficult to use because of the time pressures imposed by the experimental conditions. In other words, using negative RF takes time and a relaxed searching atmosphere. For example, "Actually, for example, if I get used to this search engine, I'll use positive and negative feedback , but now I am a participant, and I got this feeling that I have to do good. Maybe it's just like I'm in a test. Maybe I have too much pressure. Because I was afraid to wonder around, play around. I feel restricted, that I have to complete this task." (S008)

4. The Usefulness of Negative RF is Topic Related

According to searchers, the usefulness of negative RF varies by search topic. Some of the topics are quite straightforward, in which case there is no perceived need to use negative feedback. As this searcher put it, "I just used it to try to get a word...A lot of the searches
I had were straightforward so I didn't need to." (S004)

5. Word Stemming is Problematic in Negative RF

Some searchers thought that word stemming made it difficult to use negative (and positive) relevance feedback effectively. "The negative (RF) and I think positive, too, they only go by word stem, so I got a lot of things about universities, and when I tried to use negative on it, it'll show up on tape, universe, you could use negative on that stem, and then you'd be throwing out things. So maybe negative and positive shouldn't have a stem." (S004)

6. Negative RF is Simply Disliked

Some searchers just did not like negative relevance feedback, for unexplained reasons, so they did not even try to use this function. For example, "I didn't really. I didn't like it the first time. I didn't bother with it." (S007)

7. Negative RF Does Not Offer Term Control

One suggestion from searchers is that they would like to be able to type their own words, and this would make negative RF more effective. For example, "It would be really cool if you could type in, actually type in, what words you didn't want, or what words were cool, other than just putting them in the key word thing." (S004)

3.5 Performance Results

Because of the technical problems we experienced in logging our searches, which we have described above, we are not able to present comparative results between experimental and control systems. Instead, we discuss differences in performance outcomes on the aspect recall task for the two versions of our experimental system, one with positive RF only and the other with both positive and negative RF.

Our original hypothesis was that the system with both positive and negative relevance feedback will lead to better search performance than the system with positive relevance feedback only, and users will prefer negative relevance feedback to positive relevance feedback. This assumption is drawn from our analysis of thinking-aloud protocols, interviews, and questionnaires from TREC4 and TREC5 data.

Tables 1 and 2 summarize the descriptive statistics of precision and aspect recall between our two different systems: ruinq1 indicates the system with positive relevance feedback only, and ruinq2 with positive and negative relevance feedback.

Table 1. Average Precision for Searches on ruinq1 and ruinq2 (N=24)

System	M	SD	Min	Max
ruinq1	.67	.35	.00	1.00
ruinq2	.66	.22	.33	1.00

Table 2.Average Aspect Recall for Searches on ruinq1 and ruinq2 (N=24)

System	M	SD	Min	Max
ruinq1	.46	.37	.00	1.00
ruinq2	.53	.33	.11	1.00

The mean precision of ruinq1 (M = .67, SD = .35) and ruinq2 (M = .66, SD = .22) are almost the same. The result of an independent samples t-test also indicates that there is no significant difference in precision between ruinq1 (INQUERY with positive relevance feedback only) and ruinq2 (INQUERY with both positive and negative relevance feedback).

In this experiment, we were more interested in the measure of "aspect recall" than "precision", because the focus of the searchers' task was on the identification of as many aspects of the specific topic as possible. The mean aspect recall of ruinq2 ($\underline{M}$ = .53, $\underline{SD}$ = .33) is higher than the mean aspect recall of ruinq1 ($\underline{M}$ = .46, $\underline{SD}$ = .37). Contrary to our initial expectation, there is no significant difference in aspect recall between ruinq1 and ruinq2. This insignificant result is partly a result of there being too few subjects for analysis (we had only four subjects for each system). However, the comparison is in the expected direction: the system with both negative and positive relevance feedback leads to better performance than the system with only positive relevance feedback. A replication of this study with a larger sample size, or different sampling method, might reveal significant differences in performance between these two different systems. This remains an open area of investigation.

We were also interested in the possible relationships between demographic characteristics of the searchers and their performance, and also in the relationships between their subjective evaluation of their searches and their actual performance. Contrary to our expectation, none of the demographic or experience variables obtained from the pre-search questionnaire is significantly related to performance measures (aspect recall and precision). Also, there is no significant relationship between searchers' subjective evaluations and their actual performance. Again this result can be partly explained by small sample size.

3.5 System Comparisons

As a final step in our analysis we compare performance measures of all of the participating systems in the TREC6 Interactive Track. We compare the average recall and precision of all the participant systems (except for Rutgers and UMASS's INQ4int, which did not provide results of the common control system). We find that although there is a positive correlation on recall between experimental minus control systems and experimental systems (E-C vs. E, $r = .84$, $p < .001$), the recall of experimental systems is also correlated to that of control systems (E vs. C, $r = .57$, $p < .05$). Such results imply that searcher effects are greater than system effects in general at any one site. In other words, searchers at Berkeley had better recall performance than other sites in terms of both experimental and control systems. A comparison of characteristics of searchers at different sites may provide explanations for searcher effects. Secondly, the same thing happened to precision. While there is a positive correlation in precision between experiment minus control systems and experiment systems (E-C vs. E, $r = .73$, $p < .01$), the precision of experimental systems is also positively correlated to that of control systems (E vs. C, $r = .75$, $p < .01$). Searcher effects were thus dominant in precision. It seems, therefore, that experiment minus control measures appear to be a better indicator for system performance than the measures of the experimental systems alone. This confirms the design of the interactive track for comparison of different experimental systems. Thus, we are unable to fairly compare the performance of our systems with those of other participants.

4. Conclusions

It is difficult to draw any firm conclusions with respect to our initial hypotheses about the benefit of negative RF in the aspectual recall task. Clearly, this is in part a result of the small number of subjects, and perhaps also a result of the lack of a control system correcting for searcher variability. Given the somewhat contradictory nature of the evaluations of the systems

by the subjects in the scale measures as opposed to the free descriptive comments about the system features, and also the fact that ruinq2 performed at least as well as ruinq1, it may be that the most that we can say now is that ruinq2 offered our subjects a useful functionality, implemented in a rather unhelpful way.

Looked at from a slightly more optimistic point of view, it does appear that our results indicate that negative RF, implemented in this way, and subject to the control of the searcher, at the very least does not harm interactive IR performance, and may enhance it. This interpretation is of some interest, since it contradicts previous results using negative RF, especially those in which negative weights have been used. Thus, we tend to consider this study as a promising beginning for more extensive and controlled research on how best to implement and support negative RF in interactive IR.

5. References

Belkin, N.J., Cool, C. & Koenemann, J. (1996). On the potential utility of negative relevance feedback in interactive information retrieval. In: SIGIR '96. Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, ACM: 341 (abstract of a poster presentation).

Koenemann, J. (1996) Relevance feedback: Usage, usability, utility. Ph.D. Dissertation, Department of Psychology, Rutgers University.

Koenemann, J. & Belkin, N.J. (1996). A case for interaction: A study of interactive information retrieval behavior and effectiveness. In: CHI '96. Proceedings of the Conference on Human Factors in Computing Systems. New York, ACM: 205-212.

APPENDIX A: SCREEN DUMP OF RUINQ2 INTERFACE