Relevance Feedback *versus* Local Context Analysis as Term Suggestion Devices: Rutgers' TREC−8 Interactive Track Experience"

N.J. Belkin, C. Cool*, J. Head, J. Jeng, D. Kelly, S. Lin,
L. Lobash, S.Y. Park, P. Savage−Knepshield**, C. Sikora**
School of Communication, Information & Library Studies, Rutgers University
*Graduate School of Library and Information Studies, Queens College, CUNY
**Lucent Technologies, Holmdel NJ

Abstract

Query formulation and reformulation is recognized as one of the most difficult tasks that users in information retrieval systems are asked to perform. This study investigated the use of two different techniques for supporting query reformulation in interactive information retrieval: *relevance feedback* and *Local Context Analysis*, both implemented as term−suggestion devices. The former represents techniques which offer user control and understanding of term suggestion; the latter represents techniques which require relatively little user effort. Using the TREC−8 Interactive Track task and experimental protocol, we found that although there were no significant differences between two systems implementing these techniques in terms of user preference and performance in the task, subjects using the Local Context Analysis system had significantly fewer user−defined query terms than those in the relevance feedback system. We conclude that term suggestion without user guidance/control is the better of the two methods tested, for this task, since it required less effort for the same level of performance. We also found that both number of documents saved and number of instances identified by subjects were significantly correlated with the criterion measures of instance recall and precision, and conclude that this suggests that it is not necessary to rely on external evaluators for measurement of performance of interactive information retrieval in the instance identification task.

1. Introduction

Continuing our program of studying different methods of query expansion in interactive information retrieval (IR), this year our group investigated the effects of varying methods of term suggestion for user−controlled query expansion. The two methods that we compared were user control over suggested terms, implemented as positive relevance feedback (RF), versus magical term suggestion, implemented as a form of Local Context Analysis (LCA). We chose these two since they exemplify two polar methods for supporting interactive query expansion. The effects that we were most interested in were in terms of user preference, usability (as indicated by effort), and effectiveness in task performance.

Previous investigations by us (e.g. Koenemann, 1996; Park, 1999) and others (e.g. Shneiderman, 1998) have indicated that users in IR and similar systems generally prefer to have some measure of control on what the system does for them. This has often been in conjunction with an expressed desire to understand how the system has come to its suggestions/actions. These kinds of results led us to conclude that in interactive IR RF is best implemented as a term−suggestion device, rather than as an automatic query expansion device. In TREC−8, we decided to investigate the issues of control and understanding of system operation in more detail, by comparing a system in which users could control (and therefore presumably understand) where system−suggested terms came from (using RF with positive relevance judgments), with one in which suggested terms appeared as if by magic (*i.e*. LCA). Based on the previous work in this area, we hypothesized that user−controlled term suggestion would be preferred to system−controlled term suggestion.

As do others, we believe that a more usable system is a better system, and further, that a good indicator of usability is the amount of effort (physical, cognitive) that a person has to expend in order to complete a given task. We hypothesized that system−controlled term suggestion would require less effort on the part of the user than one which asked the user to make relevance judgments in order to get suggested terms. Such a difference is indicated by total time taken to perform the task, by the number of documents that a person looks at or reads, by the amount of use of various system features, and by the extent to which system−suggested terms are incorporated into the queries.

The TREC−8 Interactive Track task of instance identification is one which asks users to identify documents which

are about a number of *different* aspects of a general topic. Since RF is based on the idea of constructing an ever better (e.g. more specific) query, and since RF in interactive IR is typically based on a relatively small number of documents, it seems that RF term suggestion based only on positive relevance judgments is not well suited to this task. We can call such term suggestion *directed*. However, the terms identified by LCA for query expansion are based on a system−defined retrieved set of documents, as well as characteristics of the terms in the collection as a whole. Compared to RF, such term suggestion can be characterized as *diffuse*. We hypothesized that for the instance recall task, diffuse term suggestion would be more effective than directed term suggestion, and therefore that users would perform better in the LCA system than the RF system.

The standard measure of performance in the TREC−8 Interactive Track task is instance recall, defined as the proportion of instances of a topic that have been identified by the TREC judges, which have been identified by the searcher (as indicated by the documents the searcher has saved). Since the task that was set the searchers was to identify and save all of the instances of a topic, and since we are interested in developing evaluation measures for interactive IR that do not depend upon external relevance (and related) judgments, we also measured performance according to the number of documents saved, and the number of instances identified.

Thus, we suggest that although a term−suggestion feature based on RF might be preferred by users to one which is based on LCA, for reasons of control and understanding, the magical method will require less effort, and will lead to better performance in the instance identification task. It will not escape the reader's notice that these hypotheses seem, on the face of it, to be contradictory. That is, it does not follow naturally that a system which required less effort of the user to perform the user's task better, would not also be preferred. This situation provides another rationale for our investigations of these two conditions of term suggestion.

2. System Descriptions

There were two experimental IR systems used in this study. Both systems used Inquery 3.1p1 with its default values for indexing and retrieval (cf. Callan, Croft & Harding, 1992). The sole difference between the two systems lies in the implementation of the term suggestion feature (this leads also to minor differences in the interfaces).

The first system, called INQ−RF, allowed users to make positive relevance judgments on documents. Inquery's RF function was modified so that it displayed a list of terms for positively judged documents, rather than automatically expanding the query. As users made RF judgments about documents, the top *n* terms were presented in a term suggestion window. The number of terms displayed was determined by the formula:

$$n = 5i + 5$$

in which *i* is the number of judged documents, and *n* is no greater than 25. The term ranking algorithm was *rdfidf* (Haines and Croft, 1993), where *rdf* is the number of relevant documents in which the term appears, and *idf* is normalized inverse document frequency as used by Allen (1995) (cf. Belkin, *et al*., 1999).

The second system, INQ−LCA, employed a slight modification of the technique called Local Context Analysis (LCA) (Xu and Croft, 1996) for term suggestion. LCA combines collection−wide identification of concepts, normally nouns and noun phrases, with co−occurrence analysis of those concepts with query terms in the top *n* passages retrieved by a query. The concepts are ranked according to a function of the frequencies of occurrence of the query terms and co−occurring concepts in the retrieved passages, and the inverse passage frequencies for the entire collection of those terms and concepts. The top *m* ranked concepts are then used for query expansion. In our version of LCA, these *m* (*m*=25, to match the RF condition) concepts were displayed in a term suggestion window, after each new query. Based on an experiment using the TREC−7 ad hoc task in which we compared performance of automatic LCA query expansion using different values of *n* and different definitions of passages (with *m* constant at 25), *passage* in our study was defined as the whole document, and *n* was set to 10.[1]

Both systems used the same basic interface, developed at Rutgers, which offers the functions and features described below. Appendix A is a screen shot of the INQ−RF interface. The INQ−LCA interface was identical, except that there were no check boxes to indicate positively judged documents, and no **Clear Good Docs** button.

1 David Harper has pointed out to us that there is an inconsistency in our using the ad hoc task in these experiments, since that task is quite different from the instance recall task, especially in ways that might be relevant to choice of number of passages to be examined.

Suggested terms could be added to the existing query at the user's discretion, which is the same for both systems.

- 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 check boxes for marking documents as good (only in the case of INQ−RF) 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(s).
- Good Terms to Add window −− displayed suggested terms which could be added to the query by clicking on them.
- 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 (only in the case of INQ−RF) −− used to "unmark" previously marked good documents.
- 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.

3. Description of Study

A total of 36 volunteer searchers, recruited from the Rutgers community, participated in this project. Most (89%) of the subjects were full time students, who received compensation in the form of extra course credit for their participation. None had taken part in previous TREC studies. Each subject conducted six searches in accordance with the TREC−8 Interactive Track experimental guidelines. Subjects conducted three searches in both the RU−INQ and RU−LCA systems. We used a Latin square design where six topics were randomized and rotated completely so that each topic may appear only once in each row and only once in each column. The same set of topics was rotated again with a different system order, in order to allow a direct comparison between two different systems. Three different combinations of topic order and system order were used allowing us to run experiments with 36 subjects.

On arrival, the subjects read and signed a consent form explaining their rights and potential risks associated with participation in the experiment. Then they completed a demographic questionnaire that gathered background information and probed their previous searching experience. Next, they received a hands−on tutorial for the first system, describing the various features of that system. After completing the tutorial, subjects were given a general task description and specific instructions for the current search topic. They were allotted 20 minutes to complete each search. As they searched, they labeled instances of topic as they identified them and saved documents. During the sessions participants were asked to continuously "think aloud." A videotape recorded the computer monitor during their searches and also captured their "thinking aloud" utterances. The entire search interaction was logged.

After conducting each search, subjects answered several questions about their familiarity with the search topic, experiences with the searching task, their satisfaction with the search result, and satisfaction with the amount of time allotted for the search. After completing three searches for the first system, subjects answered several questions about the system in general. After a short break, the subjects were given a tutorial for the second system, searched another three topics, completed a post−search questionnaire for each topic, and a post−system questionnaire. After completing all six searches, the subjects completed an exit interview. The entire session took between 3 and 3 and one−half hours.

As mentioned above, an overwhelming majority (89%) of the subjects were students, and 83% were female. The average age of these searchers was 24 years old. Seventy−five percent of the subjects held, or expected to receive, a bachelor's degree at the time of the experiment. Nineteen percent had, or expected to receive, an MLS. On average, these searchers had been doing online searching for just over three years (M= 3.48).

We asked a series of questions about the background experiences of our volunteer searchers, using a 5 point scale, wherein 1= no experience and 5= a great deal. Overall, the searchers were quite familiar with the use of GUIs (*M*= 4.19), and with the WWW search engines (*M*= 4.06). A majority reported having had some experience with library OPACs (*M*= 3.22), and with searching on CD ROM systems (*M*= 3.0). In light of this, it is not surprising that on average our searchers reported conducting searches greater than twice a month.

Of note is that experience searching on commercial online systems in general was reported to be fairly low for our subjects (*M*= 2.19) and experience searching on systems other than the web was markedly low (*M*= 1.19). On a final note, the searchers in our study tended to say that they enjoyed doing information searches (*M*= 3.56) as measured by the 5 point scale wherein 1= strongly disagree and 5= strongly agree.

4. Results: Descriptive Statistics

The interaction between the searcher and the system was similar for the two systems as can be seen by the means presented in Table 1. The mean number of documents retrieved during a search topic was almost identical for LCA (*M* = 653.59) and RF (*M* = 655.49). The number of iterations (queries) in a search was roughly the same for the two systems (LCA *M*=5.93, RF *M*=5.76). Consistent with the data for documents retrieved, mean number of unique titles displayed for a search topic was also similar for LCA and RF (123.53 and 128.81, respectively). The searchers viewed the full text of almost 20% of the unique document titles displayed in each system (LCA *M* = 21.88 and RF *M* = 25.41). The similarity between the systems, in terms of interaction, was also demonstrated by the number of instances identified and documents saved. The mean number of instances identified by a searcher for a particular topic was 9.36 for LCA and 9.75 for RF. The mean cumulative number of documents saved was 8.66 for LCA and 8.69 for RF. Given that multiple instances could be found in one document, more instances were identified than documents saved in both systems. System errors generally did not occur in either system (LCA *M* = .01 and RF *M* = .01).

	LCA *M* (*SD*)	**RF** *M* (*SD*)
Number of documents retrieved:	653.59 (383.88)	655.49 (533.83)
Number of unique titles displayed:	123.53 (60.77)	128.81 (69.86)
Number of unique full−texts viewed:	21.88 (10.82)	25.41 (18.94)
Number of instances identified:	9.36 (4.79)	9.75 (5.69)
Number of documents saved:	8.66 (4.33)	8.69 (4.79)
Number of system errors:	.01 (.30)	.01 (.30)
Number of documents identified as “good”:	NA	2.85 (3.40)
Number of times suggested term list was cleared:	NA	.21 (1.01)
Number of times the good mark was removed:	NA	.74 (1.88)
Number of times the query window was cleared:	.95 (1.54)	.87 (1.62)
Number of times the save mark was removed:	.17 (.50)	.18 (.41)
Number of times paging−style scrolling was used:	22.26 (17.48)	25.44 (19.49)
Number of times dragging−style scrolling was used:	1.36 (2.32)	1.80 (3.58)
Number of suggested terms:	126.61 (83.82)	113.03 (148.93)
Number of *unique* suggested terms:	62.73 (33.59)	22.81 (24.62)
Number of unique terms used in the query:	10.0 (5.60)	8.91 (5.71)
Number of suggested terms selected to use in the queries:	4.41 (5.23)	1.87 (2.65)

Table 1: Means and standard deviations associated with system interaction and feature use.
Note: For LCA and RF, *n* = 108. Each mean is based on one search topic session.

Feature use is another aspect of system interaction. Feature use was fairly comparable in the two systems with the exception of the additional actions required to obtain suggested terms in the RF system. In the RF system, the average number of documents that were identified as relevant and used to generate suggested terms was 2.85 per search topic. Searchers generally did not use the features to clear the suggested terms list or uncheck a document as relevant (*M* = .21 and *M* = .74, respectively). On average for both systems, searchers cleared the query window no more than one time per search topic (LCA *M* = .95 and RF *M* = .87). Searchers generally did not change their mind and 'unsave' a document in either system (LCA *M* = .17 and RF *M* = .18). Across the two systems, similar frequency of use within a search session was found for paging−style scrolling (LCA *M* = 22.26 and RF *M* = 25.44) and dragging−style scrolling (LCA *M* = 1.36 and RF *M* = 1.80). Overall, these descriptive statistics demonstrate similar interactions for the two system.

When looking specifically at query formulation and term suggestion, interaction differences between the two systems emerge. Although the total number of suggested terms was similar for LCA and RF (*M* = 126.61 and *M* = 113.03, respectively), the number of unique suggested terms provided by LCA and RF differed substantially (*M* = 62.73 and *M* = 22.81, respectively). Additionally, the total number of unique terms used in the query by the searcher was similar for LCA and RF (*M* = 10.0 and *M* = 8.91, respectively). However, the average number of suggested terms that the user selected to use in their query was very different (LCA *M* = 4.41 and RF *M* = 1.87). The strongest contrasts in the systems beyond those associated with the use of relevance feedback, are the differences in the number of unique terms suggested by the systems and the number of those selected by the searcher.

5. Results: Preference, Effort, Performance

In this section, we present the results of our study with respect to the three hypotheses which motivated it. Under each hypothesis, we show the results on the relevant measures, and indicate whether the hypothesis is supported or rejected according to those results.

5.1 Hypothesis 1: User−control (RF) will be preferred to system−control (LCA)

System preference was measured by subjective response to the following question: "Which of the systems did you like best overall?"" System preference was distributed roughly evenly across the RF (39%), LCA (31%) and No Difference (31%) categories. This measure was not significantly related to system order nor to performance, but there was a relationship between preference and *perceived* system effectiveness.

Perceived system effectiveness was measured by response to questions which asked subjects about the extent to which they used each term suggestion feature to modify their searches, the extent to which they found the terms that were suggested by each of the systems useful and the extent to which the term suggestion feature improved their ability to identify different aspects of the topics in each of the systems. Each of these questions was measured on a 5−point Likert scale, where 1=not at all; 3=somewhat; and 5=a great deal. A factor analysis was performed in order to create an index variable of perceived system effectiveness for each system. One factor emerged for each of the systems that included all of the measures. The reliability coefficients for each of the system factors, Perceived System Effectiveness of LCA and Perceived System Effectiveness of RF, were high (LCA *Alpha* = .849, RF *Alpha* = .777). The mean Perceived System Effectiveness for LCA was 3.02 (SD=1.11). For RF, the mean Perceived System Effectiveness was 2.94 (SD=1.03). T−tests indicate that there is a significant difference in the Perceived System Effectiveness between subjects who prefer LCA [$t(214)=2.9$, $p<.01$] and those who prefer RF [$t(214)=-2.85$, $p<01$]. Those who preferred LCA scored significantly higher on Perceived System Effectiveness for LCA. Those who preferred RF scored significantly higher on Perceived System Effectiveness for RF.

In sum, as might be expected, subjects preferred the system that they perceived to be more effective. Their perceptions, however, were not related to objective performance measures.

Based on the relatively even distribution of system preference, hypothesis 1 is rejected.

5.2 Hypothesis 2: LCA will require less effort than RF

There was little difference in subjective responses to questions intended to measure effort on the two systems. When asked which system they found easier to *learn* to use, seventy−five percent of subjects indicated that there was 'no difference' between the two systems. The remainder of the subjects were closely split between

preferences for the two systems (LCA = 14% and RF = 11%). When asked which system was simply easier to use, fifty percent of subjects expressed no preference for one system over the other. The other fifty percent were again closely divided in their preferences between the two systems (LCA = 22% and RF = 28%). When the question focused on the ease of using the systems' term suggestion feature, only twenty−five percent indicated no clear preference. Of those searchers who had a preference, LCA's term suggestion feature was indicated as preferred more often (LCA = 42% and RF = 33%). There was no system order effect on these results.

The effort associated with interacting with the two systems was similar based on the use of features, number of iterations (queries), and the viewing of items (see Table 1 for the data on these measures). Neither page−style scrolling nor dragging−style scrolling yielded significant differences between the two systems [t(214) = −1.26, ns and t(214) = −1.06, ns, respectively]. The number of iterations (queries) in a search was roughly the same for the two systems (LCA M=5.93, RF M=5.76). The difference between the two systems was also insignificant for total number of documents viewed, total number of unique documents viewed, total number of titles displayed and total number of unique titles displayed [t(214) = −1.71, ns; t(214) = −1.68, ns; t(214) = −1.14, ns; t(214) = −.59, ns; respectively].

The total number of query terms used in a single query was roughly equivalent regardless of the system the user was using (LCA M = 10.0, RF M = 8.91). However, the way in which the terms were acquired for use in the query did vary across systems. The number of suggested query terms selected by the user was significantly higher when using the LCA system compared to the same users searching on the RF system, (LCA M = 4.41; RF M = 1.87; t(214) = 4.50, $p < .001$.) The number of user−defined terms entered into the query by the user, that is, those query terms *not* selected from the suggested terms list, was significantly higher for RF than LCA, (LCA M = 5.59; RF M = 7.04; t(214) = 2.04, $p < .05$). This suggests that in the RF system users spent more effort generating terms themselves, while in the LCA system users spent less effort thinking of terms and selected more terms from those provided.

Based on the measure of effort defined as the user's having to think of good query terms, hypothesis 2 is supported.

5.3 *Hypothesis 3: LCA (diffuse term suggestion) will be more effective than RF (directed term suggestion)*

Performance was measured by instance recall, number of instances identified and number of documents saved. The mean instance recall for the two systems was close (LCA M = .24, RF M = .26), as was the number of instances identified (LCA M = 9.19, RF M = 9.56). For number of documents saved, subjects' 'performance was almost identical (LCA M = 8.48, RF M = 8.49). These differences were all insignificant, which suggests that the effectiveness of the two systems is similar [t(214) = −.69, ns; t(214) = .51, ns; t(214) = −.06, ns; respectively]. There was no system order effect on these results. The means and standard deviations for each of the performance measures are displayed in Table 2.

	TOTAL M (SD)	**LCA** M (SD)	**RF** M (SD)
Instance Recall	.25 (.17)	.24 (.16)	.26 (.17)
Number of Instances Identified	9.38 (5.18)	9.19 (4.81)	9.56 (5.55)
Number of Documents Saved	8.49 (4.52)	8.48 (4.33)	8.49 (4.72)

Table 2. Means and Standard Deviations of Performance Measures

There was little difference in the subjective response to a question intended to measure effectiveness of the two systems. When asked which of the systems'' terms they found more effective, forty−two percent of subjects indicated that RF suggested more helpful terms, thirty−three percent of subjects indicated that LCA suggested more helpful terms and twenty−five percent of subjects indicated that there was no difference in the helpfulness of the terms suggested by the two systems.

Correlations between the performance measures were computed in order to determine the relationship between measures that depend on external relevance judgments (instance recall) and measures that depend upon user

relevance judgments (number of instances identified and number of documents saved). All of the performance measures were significantly correlated, which suggests that number of instances identified and number of documents saved might be considered as alternative evaluation measures for interactive IR. These results are displayed in Table 3.

	Instance Recall	Number of Instances Identified	Number of Documents Saved
Instance Recall	1.00	–	–
Number of Instances Identified	.412*	1.00	–
Number of Documents Saved	.315*	.896*	1.00
*Correlation is significant at the .01 level			

Table 3. Correlation Matrix of Performance Measures

Since no significant differences in performance were found between LCA and RF, hypothesis 3 is rejected.

6. Discussion and Conclusions

Our first reaction to our results is HOORAY!! For the first time in the history of the TREC Interactive Track, it's been possible to realize a statistically significant difference between two treatments. Our second reaction is, however, somewhat more muted. Two out of three of our hypotheses were rejected, and the third was supported based on only one measure out of several. What should we make of these results?

To discuss the positive first, we found a significant difference in one measure of effort between the LCA and RF systems, while finding no significant differences in preference for the two systems, nor in the objective measures of performance in the task. We conclude from these results that the LCA system is better (i.e. more *usable*) for supporting the instance identification IR task than is the RF system. This finding also lends support to the idea that an IR system which suggests terms for query modification without user control is better than one which requires user control, that is that, as Croft (1995) suggested, users want magic. Less flippantly, we note that the mean number of unique terms that were suggested by LCA was about three times the number of unique terms suggested by RF. The ratio of suggested terms added to the queries in the two systems was also roughly 3:1. This suggests that having more terms to choose from leads to including more of those terms in the query.

How can we explain the result that user−controlled term suggestion was not preferred to "black box" term suggestion? We have three possible answers. One is based on general principles of interface design (e.g. Shneiderman, 1998), which suggest that as task complexity increases, the desire for, and effectiveness of user control, decreases. In the case of this study, we note that users had three tasks in common in the two systems: developing effective queries; deciding on whether a document should be saved; and labeling the instance associated with the document. However, in the RF system, the users had also to make decisions about which documents to mark good, and to consider the relationships between these documents and the terms that were suggested. Thus, there arose an explicit task associated with term−suggestion, which in and of itself was complex, and added a layer of complexity that didn't exist in the LCA system. Thus, the measure of control that was gained, was not worth the extra complexity it required.

Another possible explanation for non−preference of user−controlled term suggestion is that user control itself was not enough to overcome the effect of the other factors which might affect preference, in particular that of effort. Finally, it could also be the case that the difference between the two systems which we hypothesized to be quite significant, was not perceived as such by the subjects. This could be due to the novel task and situation in which the subjects found themselves, and the great similarity between the two systems on other dimensions.

In terms of performance, LCA turned out not to be better than RF, contrary to our hypothesis. For this result, we also have a potential explanation, based primarily on comments made by the subjects with respect to the nature of the terms suggested by LCA. The general opinion seems to have been that many of these terms were difficult to

understand: they were in languages other than English; they were proper nouns of unusual sorts; they were sometimes only numbers. It appears that some characteristics of the term−ranking algorithm used by this version of LCA (and perhaps the values we chose for the LCA parameters) favored quite rare co−occurrences (i.e. low document frequencies). A reasonably consistent comment in the exit interview was that if LCA presented "better terms", then it might have been preferred over RF, whose suggested terms were more understandable. This suggests experimenting with the term−selection algorithm used by "magic term suggestion", to see if "better" terms will lead to better task performance.

Overall, we conclude on the basis of this study, that magical term suggestion is likely to be a better mode of support for query modification than user−controlled term suggestion, in that control of term suggestion is less important to users of IR systems than is ease of use of a term suggestion feature.

We wish to make one final observation with respect to evaluation of interactive IR. In our study, the measures of performance applied to the Interactive Track task, instance recall and precision, were significantly correlated with the number of instances identified by the subjects, and with the number of documents saved by the subjects. This result suggests to us that it would be reasonable to evaluate performance in the instance identification task without having to depend upon external judgments, relying rather solely upon how "well" the subjects performed on the task that they were in fact set: identifying as many instances of a topic as possible in a given time period. But the other groups in the Interactive Track did not find significant correlations between these measures, which suggests that factors associated with our subjects, or our interface, might have led to our result. We nevertheless believe that this result could have important implications for the methodology of experimentation in interactive IR.

7. Acknowledgments

We would like to offer our heartfelt thanks to James Allan and Victor Lavrenko, both of the Center for Intelligent Information Retrieval at the University of Massachusetts, Amherst, for all of the help that they gave us in installing LCA at Rutgers.

8. References

Allan, J. (1995) Relevance feedback with too much data. In *SIGIR '95.* Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 337−343.

Belkin, N.J., Perez Carballo, J., Cool, C., Kelly, D., Lin, S., Park, S.Y., Rieh, S.Y., Savage−Knepshield, P. & Sikora, C. (1999) Rutgers' TREC−7 interactive track experience. In *TREC−7.* Proceedings of the Seventh Text REtrieval Conference. Washington, D.C.: NIST, 275−283.

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 System Applications. Berlin: Springer Verlag, 78−83.

Croft, W.B. (1995) What do people want from information retrieval (the top ten research issues for companies that sell and use IR systems). *D−Lib Magazine,* November 1995. http://www.dlib.org/dlib/november95/11croft.html

Koenemann, J. (1996) *Relevance feedback: usage, usability, utility*. Ph.D. Dissertation, Department of Psychology, Rutgers University, New Brunswick, NJ.

Park, S.Y. (1999) *Supporting interaction with distributed and heterogeneous information resources*. Ph.D. Dissertation, School of Communication, Information and Library Studies, Rutgers University, New Brunswick, NJ.

Shneiderman, B. (1998) *Designing the user interface,* 3d edition. Reading, MA: Addison−Wesley.

Xu, J. & Croft, W.B. (1996) Query expansion using local and global document analysis. In *SIGIR '96.* Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 4−11.

Appendix: Screen Shot of System INQ−RF Interface