

Iterative Exploration, Design and Evaluation of Support for Query Reformulation in Interactive Information Retrieval

N.J. Belkin[†], C. Cool*, D. Kelly, S.-J. Lin, S.Y. Park, J. Perez-Carballo, C. Sikora**
School of Communication, Information & Library Studies, Rutgers University
4 Huntington Street, New Brunswick NJ 08901-1071 USA

*GSLIS, Queens College, CUNY, Flushing NY **Lucent Technologies, Holmdel NJ
nick@belkin.rutgers.edu ccool@qcunix1.qc.edu [dianekel | sy park | carballo]@scils.rutgers.edu
csikora@lucent.com

Abstract

We report on the progressive investigation of techniques for supporting interactive query reformulation in the TREC Interactive Track task. Two major issues were explored over four successive years: various methods of term suggestion; and, interface design to support different system functionalities. Each year's results led to the following year's investigation, with respect to both of these major issues. This paper presents first the general motivation for the entire series of studies; then an overview of each year's investigation, its results, and how they influenced the next year's investigation. We discuss what we believe has been learned through this series of investigations about effective term suggestion, usable and useful interface design, and the relationships between these two in support of the TREC Interactive Track task. We conclude with some comments about the general methodology which we employed over this series of studies, and its relevance to the development and evaluation of interactive information retrieval systems in general.

1. Introduction

1.1 Overview

Query formulation, and especially query *reformulation*, are understood to be among the most difficult tasks that users in interactive information retrieval (IR) systems face. A variety of techniques have been proposed for addressing this general problem, throughout the entire history of IR research (cf. Efthimiadis, 1996). However, very few of them have actually been tested in interactive IR environments. The research described here represents a principled attempt at taking one well-known technique for supporting query reformulation, relevance feedback (RF), and investigating its effectiveness and usability by implementing and evaluating it in the context of a specific interactive IR task.

Interface design for information retrieval (IR) systems, although an important problem for IR in general (cf. Walker, 1971), has only fairly recently become an active area of research in IR (see, e.g. Fox, et al., 1993; Hearst & Karadi, 1997; Swan & Allan, 1998; Williamson & Shneiderman, 1992). Most IR research has focused upon system functionalities, ignoring interface questions, leading to a situation in which interfaces are often seen as things that are put on top of IR systems, rather than being integral parts of them. Here, we present the results of a series of studies of interactive IR systems which attempted to address this problem by integrating interface design with development of the RF and other functionalities related to query reformulation. We construe this

[†] To whom correspondence about this article should be addressed

series of studies as an example of the iterative evaluation/design cycle, as in Egan, et al. (1989), but perhaps extending that model in our emphasis on integration of interface with function.

We describe the progressive development of an IR system and its interface designed to support a particular IR task, and to support specific user problems with that task, and with IR systems in general. This was embodied in a series of studies carried out within the Text Retrieval Conferences (TREC) Interactive Track. The starting point for this series of studies was the idea of testing the usability, use and effectiveness of automatic relevance feedback (RF) in interactive IR, embodied in work reported in Belkin, et al. (1996). Some results from these investigations, and related work (Koenemann, 1996), suggested that although system-controlled RF using only positive relevance judgments was usable and useful, additional functionality might be desirable. Specifically, user-controlled RF (RF as a term suggestion device) and RF using both negative and positive judgments were features that users seemed to desire. These results led to a sequence of four studies, implementing and/or studying these new functionalities in various interface structures, using a basic common underlying IR system (Belkin, et al., 1997, 1998, 1999, 2000). The reports of these individual studies discussed issues specific to each one, and generally did not focus on explicit interface issues. In this paper, we present an integrated discussion of the entire series, with equal emphasis on functionality and interface design. For detailed discussion of each study, see the appropriate TREC publication.

We began this series of studies in the TREC-5 Interactive Track (Belkin, et al., 1997), by investigating the use, usability and effectiveness of a system which implemented RF in the standard way in which it has been suggested for interactive IR; that is, by automatically adding terms to a query, based on the documents which had been judged relevant by the user. This study, like all subsequent ones in the series, was based on the *aspectual* (or *instance*) *recall* task set by the TREC Interactive Track. This task requires subjects to identify the different aspects or instances of a topic, and to save documents which represent those instances. The results of this study (and of a previous study in TREC-4, Belkin, et al., 1996), led to some quite explicit changes in both the RF functionality, and the system and interface in which it was implemented, for our next study, in the TREC-6 interactive track.

Three major changes were made with respect to RF for our TREC-6 investigation (Belkin, et al., 1998). One was the implementation of RF as a term suggestion device, rather than an automatic query expansion device. This followed from both the results of our TREC-5 study, and Koenemann's (1996) results indicating that RF as term suggestion is preferred to, and works at least as well as automatic RF. Another was allowing both positive and negative relevance judgments to be made, leading to the suggestion of both "good" terms to add to the query with positive weights, and "bad" terms to add to the query with negative weights. The third change, related to the second, was to implement what we have elsewhere termed a "revisionist" version of RF (Cool, Belkin & Koenemann, 1996; Belkin, et al. 1997). Since that model is rather different from the "standard" version of RF, and since it was used in our studies in TRECs 6, 7 and 8, we discuss it in some detail in section 1.2, below. The results of our TREC-6 study led us to make changes in the way in which RF was presented to the system user, and in a variety of interface-related issues, whose effects were investigated in our TREC-7 study.

In TREC-7 (Belkin, et al., 1999), the underlying RF implementation was as in our TREC-6 system, and the main goal was still to investigate the utility of negative RF as implemented in our "revisionist" model. However, the conceptual model that was presented to the user became one of "term suggestion" rather than RF, and the interface was redesigned to take account of these

changes, to respond to some specific problems indicated by the subjects in TREC-6, and to better respond to some general HCI design principles. The results of the TREC-7 study were by-and-large positive with respect to usability issues, but effectiveness and perceived usefulness of term suggestion for the task were not what might have been wished. These results led us to investigate a different mode of term suggestion in our TREC-8 study, as well as to make the system design more directly related to the task itself.

In TREC-8 (Belkin, et al. 2000), instead of comparing different version of RF for term suggestion, we compared RF-based term suggestion to another mode which we hypothesized would be better suited to the aspectual recall task. In addition, we continued to change the conceptual model of the system that was presented to the user, and to change various interface characteristics to respond both to the task, and to difficulties experienced by the users in TREC-7. This study concluded the series of investigations reported here. Although there are still a few open questions, the results of our TREC-8 study lead us to believe that we have arrived at what seems to be a reasonable and effective way to support query reformulation for the aspectual recall task in terms of both functionality and interface design.

1.2 A “revisionist” model of relevance feedback

We suggest that there are two ways in which negative relevance judgments can be understood in the context of automatic RF. The ordinary, or “classic” model of RF, based on Rocchio (1971), can be succinctly characterized as follows:

- Based on the concept of reaching an “ideal query”
- The “ideal query” is the best discriminator between relevant and non-relevant documents
- Query terms should be “optimal” discriminators
- Terms which appear only in the original query, or only in positively judged texts are “good”
- Terms which appear in both positively and negatively judged texts are “bad” because they are poor discriminators
- Terms which appear only in negatively judged texts are ignored, since they offer no information about discrimination value

Under this model, the query-term weights of terms which appear in the query and positively judged documents are progressively reduced through RF, until they reach zero weight, when they are typically removed from the query (experiments in non-interactive environments have shown that using negative weights decreases performance). Query expansion is through adding terms, with positive weights, which appear only in positively judged texts.

In contrast to the “classic” model, we suggest a new, “revisionist” model of RF which can be characterized as follows:

- The distinction between relevant and non-relevant texts which have the same terms is:
 - the terms are used in different contexts, or
 - the topics are treated peripherally, or
 - the topics are treated from an inappropriate point of view, or
 - polysemy
- RF should distinguish between appropriate & inappropriate treatments of topics
- Terms which appear in the query, or in positively judged texts, whether or not they also appear in negative texts, are “good”

- Terms which appear only in negatively judged texts are “bad” because they are indicators of inappropriate context, etc.
- Bad terms should be used for query expansion with negative weights, and good terms with positive weights

In this model, 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). Details of how this was implemented are presented in section 4.2, below.

1.3 Structure of this paper

The next section describes the general methodology that we used in all four of the studies in this series, and defines the basic measures that were used to evaluate each system. Then, we give an overview of each of the four studies, describing the study goals, the systems that were used, the experimental subjects, and the results, and concluding with how the results affected what we investigated in the next study. We then discuss the results of the entire series of studies, and conclude with some remarks on the implications of this work for the design and evaluation of interactive IR systems.

2. General methodology

All of the studies reported in the following sections were conducted under the rules of the TREC Interactive Task for the relevant years. The specifics of each of the studies are described in some detail in each of those sections, but there are some features common to all of them which we summarize here.

All studies followed a similar pattern in how they were conducted. Subjects came to the laboratory where the experiments were conducted, by appointment. At the beginning of each session, they were interviewed with respect to issues of relevance to that year’s particular research goal (recorded on audiotape), and in addition completed a questionnaire which elicited a variety of demographic data. They then received training (an online tutorial) in the use of one of the IR systems with which they would be searching, after which they conducted a small number of searches using that system. For each search, they were asked to save documents which indicated the different aspects or instances of that search topic. After each search, subjects were asked to evaluate that search in terms of their familiarity with the topic, the difficulty of the search, and their confidence in the completeness of their results. After completing all the searches in a single system, they were asked to evaluate their experience with that system, primarily with respect to the variables of interest in that particular study. In most cases, subjects were asked to search using two different systems; in these cases, after completing the cycle for the first system, they received a tutorial for the second system, and then followed the same pattern. After completing all of the searching and related questionnaires, subjects completed questionnaires and were administered interviews (again recorded on audiotape) which asked them to compare the two systems which they had used, again with respect to the variables of interest to that year’s particular study. All of the

searches were completely logged, the monitors were videotaped during the searches, and the subjects were asked to “think aloud” during the course of each search. The thinking aloud was recorded on the videotape of that search. All of these collection techniques together provided the data according to which we measured the effectiveness and usability of the systems. In each of the studies, the entire experimental process took about three hours for each subject.

For all of our studies, we used some version of InQuery from the Center for Intelligent Information Retrieval, University of Massachusetts, Amherst (cf. Callan, Croft & Harding, 1992), as the basic indexing and retrieval engine, suitably modified, and constructed our own interface on top of that engine. The searches were conducted on SUN SPARCstations of different sorts each year. From year to year both the version of InQuery and the workstations were moderately upgraded.

Measure/Characteristic	Definition
Search	All which took place between the time the subject was handed a sheet of paper with a description of the search topic and instructions for conducting the search, and when the subject invoked the “Exit” button in the system (i.e. when the subject completed the search task).
Cycle	All which took place between one invocation of the “Search” button in the system (i.e. submission of a query), and the next invocation (first cycle starts with the person receiving the topic; last cycle ends with the “Exit” button)
Aspectual recall	The proportion of total number of aspects/instances identified by the TREC judges, which are represented by the documents saved by the searcher
Query	The terms entered into the query window by the subject (whether typed or selected) in one cycle
Query terms	The total number of unique terms that were used for all queries during a search
User query terms	The total number of unique terms that were typed into the query window by the user during a search
System query terms	The total number of unique terms that were selected by the user from the lists of terms suggested by the system, for inclusion in the query window, during a search
System suggested terms	The total number of unique terms that were suggested by the system during a search
Unique titles displayed	The total number of unique titles, in the summary window, which were displayed during the course of a search
Unique texts displayed	The total number of unique documents whose full text the subject selected for viewing during a search

Table 1. Definitions of measures and search characteristics used in the Rutgers TREC studies.

Although all of the four studies did not measure performance according to all the same measures, there were a number of measures that were used in common for TRECs 6, 7 and 8. These measures (and the characteristics of the search process on which they are based) are defined in Table 1. TREC-5 measures were somewhat different, in that query expansion by RF was accomplished automatically, whereas in the other three systems, it was used as a term suggestion device.

3. TREC-5

3.1 Study Goals

In TREC-5, our approach was to further develop the conceptual work in the area of relevance feedback (RF), reported by our group in previous TREC-3 and TREC-4 experiments in the ad hoc task. In TREC-5, this theoretical interest in RF led to an investigation that explicitly conformed to the Interactive Track task. Our work in TREC-5 formed the baseline for the iterative development of the system functionalities and interface modifications reported throughout this paper.

The focus of our investigation in TREC-5 was on the following factors:

- the understanding, use and utility of positive-only RF for the TREC-5 Interactive Track task
- the searchers' understanding of the TREC-5 Interactive Track task
- the range of performance by different searchers on the same topics
- the use of various system functionalities afforded by the RU-INQUERY system

3.2 System

Figure 1 is a screenshot of RU-INQUERY, which offered the following features:

- unstructured (phrases indicated by hyphens between successive terms) query input, in the editable query window at the top of the interface;
- display of the document id numbers and titles, in the middle “summary” window
- ability to mark a document “relevant” by invoking the check-box to the left of the title in the summary window (documents are unmarked by clicking on a marked check-box)
- ability to mark a document “saved” by invoking the check-box to the right of the title in the summary window (documents are “unsaved” by clicking on a marked check-box)
- ability to view the full text of a document in the bottom window, by double-clicking on its title in the summary window
- query words in the document text (including RF terms) are underlined and highlighted
- ability to save a query, to load a saved query in the query window, to clear a query, and to clear all documents marked relevant, in the top command line
- information about how many documents are marked relevant, how many relevance feedback terms have been added to the query, how many documents have been saved, and the rank of the currently viewed document, of the default of 150 retrieved
- ability to scroll directly to the next keyword in the document display, in the bottom command line

[insert figure 1 about here; figure 1 will be mailed separately, and added to this paper]

3.3 Subjects

Twelve volunteer searchers participated in this study, recruited from the community of information professionals in New Jersey, and from the students and faculty of the School of Communication, Information and Library Studies and of the Computer Science Department at Rutgers University. Seven of the subjects were male and 5 were female. A majority of them were between the ages of 41-60. All of the searchers had at least a Bachelor's degree, and nine had an MLS at the time of the study. Four subjects reportedly had a Ph.D. and another two expected to receive a doctorate. Overall, our subject population was a highly educated group

At the time of this study, subjects had a mean average of 6.3 years of searching experience on a variety of information systems. At the same time, their overall experience with systems offering RF, with Ranked-Output systems, and with Full-text Databases was relatively low. Using a 5-point rating scale to measure their searching experiences, in which 1=None and 5=A great Deal, the average level of experience with RF was 2; with Ranked-Output Systems it was 2.5; and with Full-text Databases it was also 2.5.

3.4 Data Sources and Instruments

Each participant conducted six searches on the RU-INQUERY system. Subjects followed the guidelines set forth for the Interactive Track, in that they were instructed to "identify as many aspects as possible for each topic, within a 20 minute time period for each topic." Following TREC Interactive Track protocol, each searcher was given a predefined set of six topics to search.

Prior to conducting their searches, subjects completed a brief questionnaire about their demographic characteristics and about their normal searching behaviors. Some examples of these data are illustrated above. Next, searchers were given a printed tutorial which was designed to interactively guide them through the workings of the RU-INQUERY system, and in particular, to explain the mechanics of the interface; using the features of RU-INQUERY for constructing queries; and interpreting the system output. The functions Ranked Output and Relevance Feedback were conceptually and procedurally explained at length in the tutorial. The tutorial also explained the use of the Save Query and Load Query features and the Save document button. The interactions between Clear Query, Clear Relevance Feedback and Save Query were demonstrated, as was the difference between the Save Document and Relevant Document features. The tutorial attempted to cover all of the RU-INQUERY system features.

The average training time per searcher, as measured by the time spent working through the tutorial, was just over 25 minutes (mean=25.45). The shortest mean time spent in training was 8.77; the maximum mean time was over an hour (67.75).

After completing each of their six searches, subjects were asked, on a 5-point scale, to assess their familiarity with the topic; how difficult it was to do the search on this topic; how satisfied they were with their search results; how confident they were that they had identified all possible aspects for the topic; and, the extent to which they had enough time to do an effective search on that topic.

At the end of the entire searching experience, subjects participated in a semi-structured post-search interview which was designed to elicit beliefs, attitudes and behaviors directly relevant to RF, Ranked Output, and the specific aspectual recall task required of participants in TREC-5.

3.5 Results

The results in this section provide behavioral measures of subjects' uses of a variety of functionalities afforded by RU-INQUERY, their performance on the aspectual recall task, and their subjective assessments of the usability of RF in particular. These data are then discussed in terms of their usefulness in directing future iterations of our general program of study.

Measure	Mean (SD)
Aspectual recall, per search	0.438 (NA)
Number of documents saved, per search	9.40 (7.04)
Time spent per search (minutes)	19.11 (2.72)
Use of RF, per search	3.82 (3.32)
Query terms, per query, for non-RF queries	4.07 (2.79)
Query terms, per query, for all queries	9.93 (8.41)
% of queries using RF, when RF could be used	59.76%
User query terms, per query, for RF queries	5.23 (3.55)
RF terms, per query, for RF queries	10.63 (6.64)
Query terms, per query, for RF queries	15.86 (8.30)
Cycles per search	7.90 (4.37)
Titles displayed per search	107.47 (63.07)
Full text displayed per search	21.86 (10.87)

Table 2. Behavioral measures for the RU-INQUERY system

Table 2 above indicates that on average, participants spent close to the maximum allowable time in searching for documents. However, despite their fairly high degree of time spent in the searching exercise, RF was used, on average, rather infrequently. In the RU-INQUERY system searchers had the opportunity to use RF only after they had originally selected retrieved documents as relevant. This eliminated all first queries. During this experiment, there were 373 cycles for which searchers had the opportunity to use relevance feedback. Of these, subjects used RF in 254, or almost 60%.

It is of some interest to note that there was a substantial difference in the number of query terms supplied by the user in queries which used RF (5.23), as compared to queries which did not use RF (4.07). Although this might be in part attributable to the fact that initial queries are in general shorter than subsequent queries in a search, it also suggests that there may be a relationship between the use of positive, system-controlled RF, and user involvement in manual query expansion.

One of the issues we explored further in our exit interview concerned the extent to which RF was conceptually understood by these searchers who were for the most part unfamiliar with the feature; and also the extent to which they found RF to be useful and usable. After all six searches were completed, the subjects were administered an exit questionnaire and interview, whose foci were: the understanding and use of RF in their searches, and its utility for the interactive track task; and their understanding and experience of the task itself. Here, we are concerned with an analysis of subjects' responses to the following questions asked during the post-search interview:

- To what extent did you understand how to use Relevance Feedback?
- To what extent did you use Relevance Feedback during your searches?

- To what extent did you find Relevance Feedback useful during your searches?
- To what extent did Relevance Feedback improve your ability to identify different aspects of the topics?

For each of these questions, searchers were asked to respond on a 5-point scale, where 1= Not at all and 5=To a great extent. For each response on the scale, subjects were asked to give an open-ended explanation of the reasons for their answer. Our final question during the post-search interview asked subjects if they had any other comments about their experiences with RU-INQUERY. Table 3 summarizes our subjects' responses to the questions about understanding, use, and usefulness of RF in the RU-INQUERY system.

	Number (%) responses on each scale point (N=12)					Mean	SD
	1	2	3	4	5		
Understand RF	0	0	3 (23%)	7 (58%)	2 (17%)	3.83	.54
Use RF	0	0	5 (42%)	3 (25%)	4 (33%)	3.92	.90
RF Useful*	0	3 (27%)	4 (36%)	1 (9%)	3 (27%)	3.41	1.20
RF Good for Task*	2 (18%)	3 (27%)	3 (27%)	1 (9%)	2 (18%)	2.86	1.42

Table 3. Understanding, Use and Usefulness of Relevance Feedback in RU-INQUERY(1=Not at All; 5=A Great Deal) * One case missing.

The general picture we see from Table 3 is that the subjects tend to report that they understand how to use RF, and all of them said that they used it during their searches. At the same time, however, searchers tended to find RF not very useful during their searches, and not helpful in improving their ability to identify different aspects of their search topics. In other words, searchers in our study were able to form a conceptual model of what RF was, and how it should work, but they felt they were not always able to effectively use this functionality, as presented to them in our interface. Our analysis of the data from the post-search interview sheds some light on this problem. Two main themes emerged from these data, having to do with user-control over RF, and general usability of the feature itself.

Over half (7) of the subjects expressed some desire for better control of RF. The most frequently mentioned problems with RF were that it did not permit users to identify documents as “negatively relevant”, and that it did not allow user control over term selection. Several subjects were quite specific about their desire for control over the terms that were to be added or not added by RF. One searcher told us, “There’s no ability to control it. Better if terms could have been seen and either approved or vetoed.” Another subject echoed this feeling by saying, “One problem with relevance feedback was that there was no way to see which words it was using. I would like more control over it.” Another stated more simply, “If I knew how the algorithm worked, I could manipulate it more to find things I wanted to find.” Related to this issue of user control is the feeling of trust that searchers placed in RF. Because much of how RF works was opaque to the searchers, at least one user felt that he “liked it, but felt it was hiding too much.”

Several searchers felt quite strongly that RF would have been more useful if they could specify negatively judged documents. Many of these people were experienced Boolean searchers who also expressed a desire to have a NOT operator in the RU-INQUERY system. “I wanted a negative relevance feedback” explained one subject who found RF to be not very useful. A couple of

subjects mentioned that RF was difficult to use because it didn't eliminate "bad documents" from the retrieved list, which they found to be frustrating. Although these searchers didn't put it this way, an obvious solution to this problem is negative RF.

A majority of our subjects (8) expressed dissatisfaction with the general usability of RF. "I understand the concept of relevance feedback, but I'm not sure it was well-implemented in this system" is how one subject characterized this view. A common complaint was that the searcher and the system had differing definitions of relevance. In such cases, RF returned documents with the right words, but the wrong meanings. "It marked documents as relevant if it had the right type of words, but not necessarily the right treatment of the subject" said one subject. "It would have been useful if it worked; I didn't see the same things as relevant, didn't find the system items relevant" said another. Another complaint we heard was that "relevance feedback brought in unhelpful terms such as names and places."

Although mentioned less frequently, several problems with the RU-INQUERY interface were mentioned that were unrelated to RF. Several searchers desired a mechanism to help them keep track of the documents they had saved, or had marked as relevant. "I wanted to list saved documents and marked relevant documents, in order to re-evaluate. I lost track of saved and relevant documents, and wanted to look at the saved set to see how well the aspects were covered."

3.6 Changes to the RU-INQUERY System

Based upon these results, we made the following significant changes in our system design for TREC-6:

- Included negative as well as positive RF
- Provided user control over positive and negative term selection and de-selection (i.e. treated RF as a method of *term suggestion*).
- Implemented RF according to the revisionist model discussed in section 1.2, in order to attempt to take account of user comments concerning *context*.

These functional changes necessitated significant interface changes (Figure 2). To support positive and negative RF, the single column of "Relevant[]" checkboxes was replaced with two columns labeled "Pos" and "Neg." To support user control, two new windows were added in which terms suggested by the RF algorithm were displayed, for user selection. Two new sections were added to the query formulation window into which terms from the relevant terms suggestion windows could be entered by double clicking on them. Other changes not related to the functionality issues included:

- change in interface colors
- using color to highlight rather than underlining
- providing navigation through document text using best passages
- moving directly from the current document to the previous or next document

4 TREC-6

4.1 Study Goals

The primary goals of this study were:

- to investigate the effectiveness and usability of negative RF in interactive IR.

- to investigate the use and usability of RF as a term suggestion device;
- to investigate aspects of the revisionist model of RF.

We attempted to accomplish these goals by implementing a version of RF which suggested terms for addition to the query, rather than automatically expanding the query with those terms,

4.2 System

We used InQuery 3.1p1 as the basis for our experimental systems. The two versions of InQuery were: 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:

$$5n + 5, \text{ where } n = \text{number of judged documents}$$

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. Figure 2 is a screenshot 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 offered by the systems were:

- Unstructured query input plus phrases in the query formulation window (top center frame);
- Saving, clearing and loading queries;
- Display of rank, date and title of ten retrieved documents at a time (center frame);
- Scrolling the title display ten documents at a time;
- Saving a document to indicate one or more aspects - unsaving by clicking on saved document button (right hand button on the title line);
- 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;
- Display of suggested RF query expansion terms (positive terms displayed in upper leftmost frame; negative terms displayed in lower leftmost frame);
- 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);
- 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);
- Clearing all relevance markings (removes all term suggestions, but not term selections);
- Displaying the full text of a document by double clicking on the title line (displayed in the bottom center frame);
- Scrolling through the full text of the document;
- Highlighting query terms in the full text display;

- Scrolling directly to the next query term in full text display (Show Next Keyword);
- Showing the best (next best, previous best) passage in the full text display, according to default InQuery 3.1p1 method;
- 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.

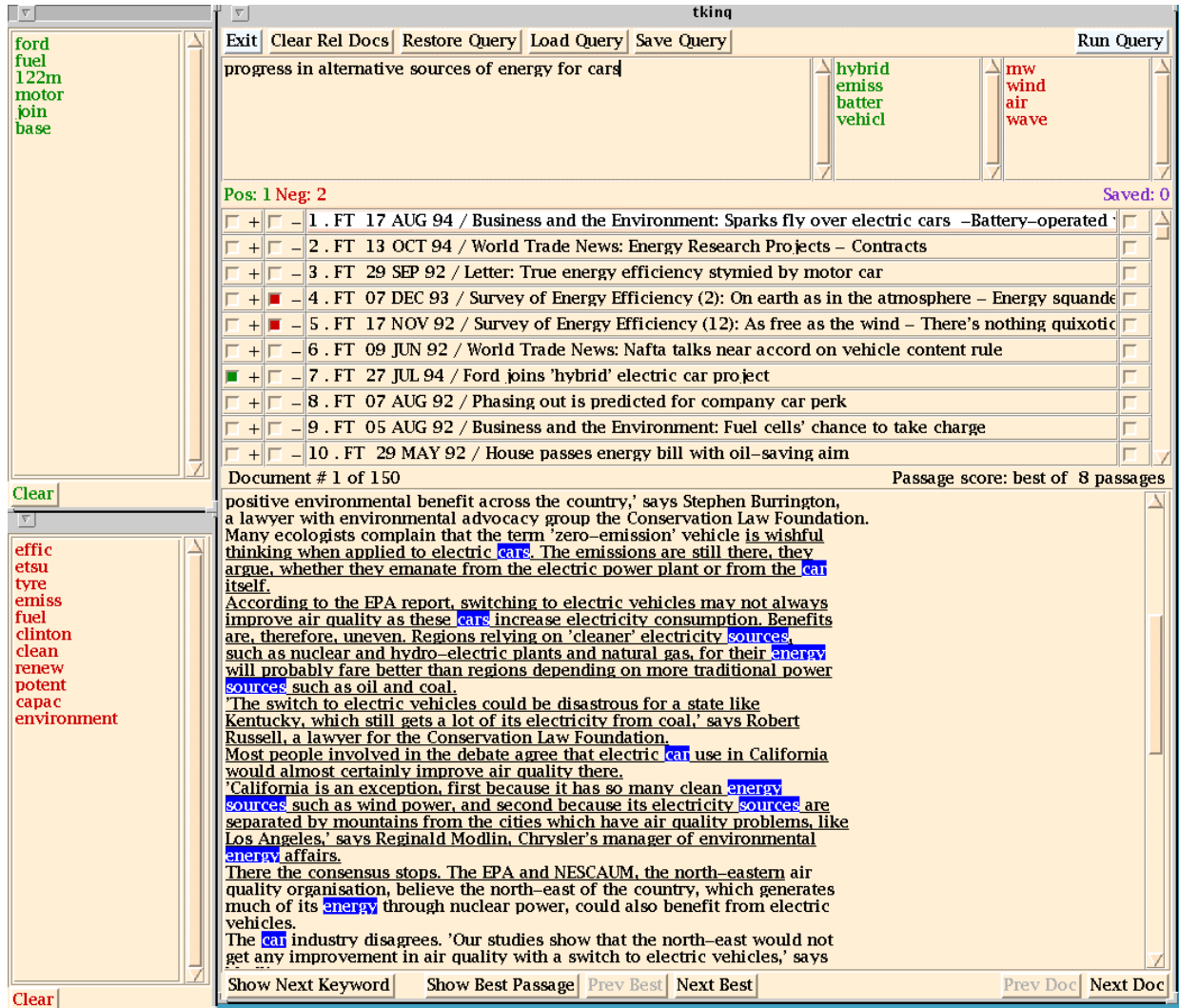


Figure 2. TREC 6 Interface Supporting Positive & Negative Relevance Feedback (RUINQ2)

4.3 Subjects

Eight volunteer subjects included 6 females and 2 males. The subjects ranged in age from under 21 to between 51 and 60. Five of the eight subjects had, or were pursuing, a graduate degree in library science, and the other three subjects indicated no education in library science and had, or pursuing, Bachelor degrees in other fields. The median number of years reported for overall online search

experience was 3.5. The range was 1.5 years to 10 years. None of the subjects had prior experience on the systems being investigated.

4.4 Data Sources/Instruments

Subjects conducted three searches on different topics on each version of the system. Each subject searched with two systems: one of the two versions of the system described above, and a control system. Questionnaires after each search, each system and at the end of the experiment were used to assess quantitative ratings of the task and systems. Specifically, subjects responded regarding the systems' ease of learning, ease of use and understanding of how to use. Qualitative data were collected through the use of a think aloud protocol during the search process, and from comments made on the questionnaires and during the exit interview.

4.5 Results

The quantitative ratings of the two systems provided subjects' reactions to the ease of learning to use the systems, the ease of use and the ability of the subjects to understand how to use the systems. These ratings are shown in Table 4. Subjects responded on a five-point scale where 1 represented "not at all," 3 indicated "somewhat" and 5 suggested "extremely".

	RUINQ1 Mean (SD)	RUINQ2 Mean (SD)
Easy to learn to use	3.30 (1.27)	3.67 (0.27)
Easy to use	3.33 (1.19)	4.00 (0.72)
Understand how to use	3.25 (1.28)	3.67 (0.98)

Table 4. Usability ratings of the RUINQ1 (positive RF only) and RUINQ2 (positive and negative RF) systems.

As in TREC-5, subjects claimed to have understood RF even though there were clearly misconceptions about its use as implemented in the RUINQ systems. In particular, the five-window arrangement for query formulation and term suggestion (Figure 2) confused the subjects. They did not recognize the association between the original query window and the windows displaying the selected suggested terms. They often wanted to type their own words into the windows displaying the selected terms. Subjects would mark documents positive or negative to try to find specific terms to add. They did not realize they could type the terms into the original query formulation window. The confusion experienced by the users demonstrates a need for better functional grouping of elements within the interface.

Stemming was another problem related to RF. The terms suggested by RF were displayed as word stems, which subjects did not always understand. They also found it difficult to match the stems to the words of the documents which they read. In general, the use of stems added cognitive effort to the task.

Another problem that became clear from the transaction logs of subjects who did not fully utilize the RF feature was that they seemed to think that changes would occur automatically. They would mark a document as positively or negatively relevant, but not choose any of the suggested terms before re-running the query. Or, they would choose some of the suggested terms, but not re-run the new query. Both cases resulted in no new search, and no change in the retrieved document list.

The interface did not provide sufficient understanding of the impact of marking items as relevant or not relevant. With the lack of clear understanding, subjects made assumptions about what was occurring based on their internal model of how IR systems work or should work.

Subjects were sometimes confused when they saved a document, marked it negatively relevant to avoid getting more of the same, and then found that the new document list did not have the saved document. Although this was the appropriate use of the negative RF feature, it led to discomfort and uncertainty in the mind of the subject not to see the previously saved document.

Displaying the number of documents retrieved by a query (the system default was to return the top 150 documents) led some subject to believe that *only* 150 documents matched their query, and others to believe that their query modifications were having no effect.

	RUINQ1 Mean (SD)	RUINQ2 Mean (SD)
Query terms, per search	12.83	10.75
User query terms, per search	9.75	6.25
System query terms, per search	3.08	4.5
Ratio of user terms to all query terms	0.79	0.58
Positive system suggested terms	25.08	23.58
Negative system suggested terms	0	13
Positive system terms selected	3.08	3.33
Negative system terms selected	0	1.17
Number of documents marked positive	3.67	4.33
Number of documents marked negative	0	1.58
Titles displayed	378.08	205.25
Documents displayed	25.17	52.17
Cycles per search	8.92	5.17
Aspectual recall	0.46	0.53
Documents saved	6.25	8.33

Table 5. Use and effectiveness of RUINQ1 (positive RF only) and RUINQ2 (positive and negative RF)

An analysis of the log data (Table 5) provided insight into the use and effectiveness of RUINQ1 and RUINQ2. The numbers of iterations, or cycles, per search are quite different (mean for RUINQ1 8.92, mean for RUINQ2, 5.17). 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 scrolling through the retrieved document list. The searchers in the RUINQ2 condition made more use of relevance feedback terms (RUINQ1 mean 7.42, RUINQ2 mean for positive terms 11.08, and for negative terms 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). Although aspectual recall was higher for RUINQ2 (0.53) than RUINQ1 (0.46), this difference was not statistically significant, possibly because of the small number of subjects. One can note, however, that the

number of documents saved in RUINQ2 is substantially higher than in RUINQ1 (8.33 vs 6.25, respectively), and the actual task set the users was to save documents identifying as many instances as they could.

Overall, on these quantitative measures of the interaction, although there are some differences between behavior in the two systems, they are not easily explained by the presence or absence of system support for negative RF. However, one should note that the percentage of terms selected from those suggested is substantially higher in RUINQ2 than in RUINQ1, which suggests that the subjects in RUINQ2 perhaps had to exert less effort (i.e. think up fewer terms on their own) than did those in RUINQ1.

4.6 Changes to System

Based upon these results, we made the following significant changes in our system design for TREC-7:

- Selected terms were entered into the single, editable query window, eliminating the two separate windows. This is to reinforce the searchers' ability to manually enter positive or negative (with a minus sign) terms, and to give a single query view.
- Stemming was changed from Porter to K Stem, which presents the terms in a canonical, rather than stemmed form. This addressed the problems associated with the confusion and ambiguity of stemmed words in the term suggestion window.
- An error message was introduced to warn subjects that no changes had been made since the query was last run. This addressed the problem of users marking documents positive or negative and then not selecting any items before re-running the search.
- Double clicking was changed to single clicking. This change is consistent with a general effort to accommodate people with difficulties double clicking.
- Display of the number of documents retrieved by the search was removed from the interface. It appears that seeing the number of documents retrieved led subjects to a set-based exact-match model of IR, rather than the more appropriate best-match model.
- The interface terminology was simplified. The new terms were selected to be more intuitive and easier to understand. In particular, "Pos[itive]" and "Neg[ative]" were replaced by "Good" and "Bad" to better map to RF as term suggestion. Functions not used were eliminated and complex functions were separated into multiple buttons.
- Titles were added to all of the windows within the application screen. This partially addressed the confusion about the relationships of the windows to one another. Balloon help, invoked when the cursor based over the title bar, also addressed this issue.
- The buttons were moved to provide better functional mapping and better visibility.
- A saved documents list was implemented. This separate window lists the documents saved and requires naming the instances identified. This addresses the confusion subjects experienced when a document marked to be saved and marked as negatively relevant disappeared when the search was re-run. Requiring the subjects to enter the aspects or notes about why they saved the document provide a reminder of what aspects have already been identified.

TREC-7

5.1 Study Goals

As in TREC-6, a main goal of this study was to investigate the effectiveness and usability of negative RF. A second goal was to investigate both positive and negative RF as user-controlled term suggestion. We also wanted to investigate the effects of the changes that were motivated by our TREC-6 results. These goals were accomplished by comparing two systems using the same interface, one offering positive and negative RF (INQ-R, Figure 3), and the other offering only the positive RF feature (INQ-G).

5.2 Systems

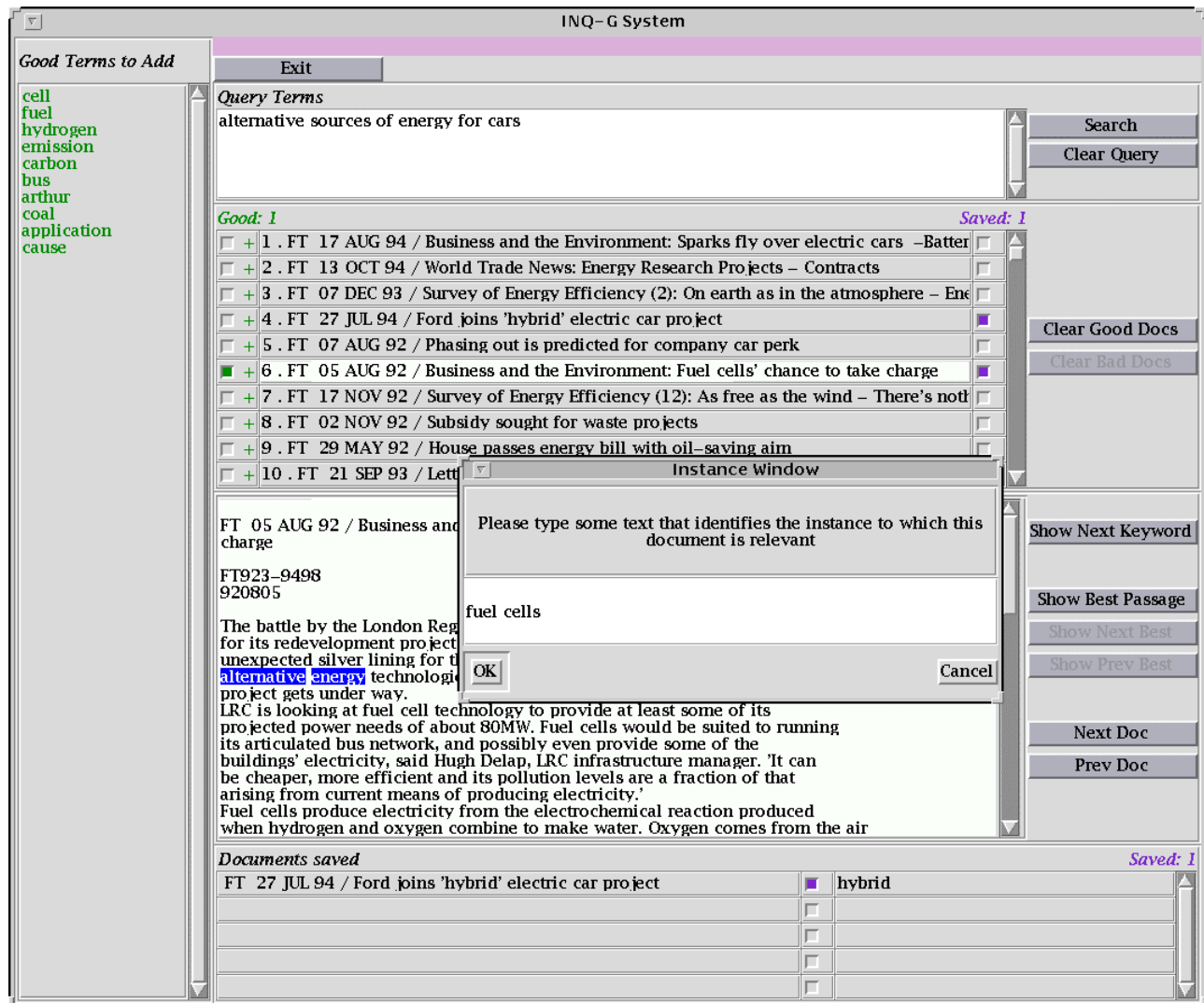


Figure 3. Screen shot of INQ-R system (positive and negative RF)

We used exactly the same implementation of InQuery and the RF method as for TREC-6. Figure 3 is a screen shot of the positive and negative RF system (INQ-R). The positive only RF system (INQ-G) was identical, except that there were no Bad Terms to Avoid window, no Clear Bad Docs button, and no Bad RF radio buttons. The system 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.

5.3 Subjects

There were 16 subjects, including 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.

5.4 Data Sources/Instruments

Subjects conducted searches on eight different topics, four in one of the INQ interfaces, and then four more in the other. The data collection instruments were the same as in TREC-6

5.5 Results

An analysis of the three-item post-system questionnaire (Table 6) yielded no statistically significant difference between subjects' mean ratings of the usability of INQ-G compared to INQ-R. However, the directional tendency is one in which the system offering positive only RF was perceived to be slightly easier to learn to use, easier to use and easier to understand.

	INQ-G Mean (SD)	INQ-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 6. Usability of INQ-G (positive RF only) and INQ-R (positive and negative RF)

An analysis of the log data (Table 7) provided insight into the use and effectiveness of INQ-G and INQ-R. The way in which the subjects used the two systems was generally similar. The number of cycles per search was not visibly different. The number of user generated terms in those queries was nearly identical. However, the total number of terms used in the queries was higher for the system offering both positive and negative term suggestions, and the percentage of system generated terms was also higher. As in the TREC-6 results, this suggests that the use of negative term suggestion was substantial, and also that the subjects in INQ-R may have had to expend less effort for longer queries. The results may not reflect user entered query terms that were prompted by the term suggestion feature, but were typed in rather than chosen from the tool. The number of documents in which the subjects viewed the full-text was very consistent between the two systems. However, the system only providing positive term suggestions had fifty percent more titles viewed compared to the system with positive and negative terms suggestions, again, corresponding to the results of TREC-6. The effectiveness of the two systems also proved to be roughly equivalent. The means for aspectual recall on the task were very similar, as were the means for number of documents saved. Overall, this data do not demonstrate differences in general use and effectiveness of the two systems. However, some differences in the use of the features were evident.

	INQ-G Mean (SD)	INQ-R Mean (SD)
Query terms	9.72	11.39
User query terms	7.70	7.72
Positive system generated terms in query	2.02	1.74
Negative system generated terms in query	NA	1.98
Total system generated terms in query	2.02	3.72
% System generated terms per query	82%	72%
Positive suggested terms per session	2.31 (3.34)	1.76
Negative suggested terms per session	NA	1.97
Number of documents marked good	3.08	2.55
Number of documents marked bad	NA	2.25
Unique titles displayed	304.02 (339.69)	203.33 (250.65)
Unique full-text viewed	27.24 (16.30)	24.80 (12.15)
Cycles per search	7.19 (5.30)	6.41 (3.77)
Aspectual recall	.39 (0.24)	.35 (0.23)
Documents saved	6.46 (3.78)	5.87 (3.86)

Table 7. Use and effectiveness of INQ-G (positive RF only) and INQ-R (positive and negative RF).

Overall, the use of the RF features was not as frequent as was expected. Subjects marked an average of three documents as good per topic in an effort to obtain positive suggested terms. They only included two of the terms on average during a search. When choosing good and bad documents to obtain positive and negative suggested terms respectively, on average subjects only marked two and one half documents good and two and a quarter bad. On average, they included fewer than two of each of the positive and negative suggested terms in their queries. The effort required to obtain suggested terms may have inhibited the use of the feature.

The qualitative interview data provide some insight into how searchers understood and used the RF features. Many subjects reported that they used, and liked, both the negative and positive term suggestion features. A common feeling was that the feature was easy to use and easy to learn to use. However, in their open-ended comments subjects often referred to the terms that the suggestion feature offered as synonyms. This suggests that searchers who used the term suggestion feature may not have had an appropriate model of how it functioned.

No subjects expressed any difficulty executing the term suggestion feature. They seemed to understand the mechanics of selecting and adding good and bad terms to their query as well as deleting these from their query. However, many subjects expressed frustration because they were not able to highlight and delete search terms from their query. Instead, they were only able to backspace over a query term in order to get rid of it or modify it. This particular shortcoming was consciously implemented during the design phase in order to allow subjects to manually change one or several characters of a search term. Subjects were especially appreciative that they could add their own negative terms to the query using the minus (-) sign, suggesting that subjects had a desire to use the negative term suggestion feature.

One feature of the INQ interfaces that is different from many other search interfaces is its one-screen design. This characteristic of the interface was a result of the previous iteration study where subjects consistently complained about the five-window screen. Although the implementation of the features within the interface changed, the capabilities of the system did not. Subjects use a single screen to enter queries, view titles, view full text and save documents. The status of all of these functions is available at any given time to the user without requiring him/her to view another page or go back to a page. Nothing is hidden or covered up by multiple windows. Subjects do not have to keep track of multiple windows and screens that might be used to carry out various search activities nor do they have to keep up with the information that might be displayed on these various and changing screens. Subjects made repeated comments about all of these advantages of the one-screen design. This integration of activities reduced the cognitive requirements of the user by making extensive and efficient use of the screen real estate.

This integration of activities also contributed to subjects' understanding and trust of the system. Subjects liked the stability and consistency of the interface. They were pleased that things did not "pop" out at them while they were carrying out their search activities. In terms of subject control, subjects appreciated that all options were displayed and available to them at all times and that they had the ability to determine if, and when, these options would be used.

Two features of the interface that were particularly liked by most subjects were the "Show best passage" and the "show keyword" buttons. Subjects liked these features because they reduced the amount of time needed to navigate to specific terms and passages in the document. This, in turn, reduced the amount of time that subjects had to spend evaluating documents. This seems to have

been especially important since the aspectual recall task required a good deal of reading and interpretation of the documents.

The mechanism for saving documents required subjects to create a label for each document saved. Subjects reported liking this feature because the labels acted as a memory aid during subsequent searching. However, the mechanics of creating the labels often frustrated users. The labels were created in an instance window that was triggered when subjects saved a document. While subjects could move this window around on the screen, they could not scroll through the document while the window was present. Many subjects expressed a desire to do this, especially when one document contained many instances.

The system design for TREC-7 was improved from that of TREC-6 by reducing a five-window arrangement to a single window, by requiring subjects to label the saved documents and providing those documents with their labels for reference. These two previously identified problems seemed to be remedied by these solutions. However, additional opportunities for improvement became evident in the TREC-7 study. The suggested terms generated by marking documents as negatively or positively relevant were not used as much as anticipated, suggesting that the effort required to generate suggested terms should be reduced. The users commented that the term suggestion feature was not useful with respect to the task. Relevance feedback may be too directed for the task, since this feature tries to narrow the search to a single aspect. Providing automatic term generation, selected from a larger range of documents, might increase the likelihood of including terms that address different aspects. This *diffuse* approach may be more appropriate for the aspectual recall task than the more *directed* RF term-suggestion.

5.6 Issues to be Addressed

The results of the TREC-7 study raised a number of system functionality and interface issues, which we addressed in our TREC-8 study. These include:

- Presenting a more accurate conceptual model of the terms that are offered by RF.
- Presenting a conceptual model of best-match retrieval in which relationships between ranking and numbers of documents retrieved are made clearer.
- Providing more direct query editing facilities, and direct cut-and-paste between document and query.
- Cleaning up and extending the instance-identification facility to a more general information organization and structuring feature.
- Reducing the amount of effort required to make suggested terms available to the user.
- Introducing a term-suggestion feature hypothesized to be more directly useful for the aspectual recall task.

6. TREC-8

6.1 Study Goals

The primary goal of the TREC-8 study was to investigate the effectiveness and usability of two different term suggestion methods for interactive IR. This focus was a result of the observations in TREC-7 that the term suggestion feature was not used as often as we had expected, and that the subjects did not find it all that useful with respect to the task. We were also concerned to further investigate the issue of user control of term suggestion. The two methods that we compared were user control over suggested terms, implemented as positive (RF), versus “magical” (or system-

controlled) 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 especially effectiveness in task performance.

As far as the term suggestion functionality was concerned, in TREC-8 we had two hypotheses. Based on previous work on user preference for control (e.g. Koenemann, 1996; Shneiderman, 1998), we hypothesized that user-controlled term suggestion (i.e. RF-based term suggestion) would be preferred to system-controlled term suggestion. But in the RF system, term suggestion was based on a relatively small number of documents that had to be selected as relevant by the user (i.e. terms are generated from the document level). Thus, in this system term suggestion was *directed* by some small set of documents. In the LCA system, term suggestion was based on a system-defined set of documents (e.g. the top ten documents retrieved by a query), as well as by characteristics of the terms in the collection as a whole. Thus, in this system, term suggestion was based on a more *diffuse* set of sources. We felt that the “diffuse” system better matches the aspectual recall task than does the “directed” system, since that task asked the subjects to find different aspects of a topic, rather than to narrow in on some specific aspects. Therefore, we hypothesized that subjects would be more effective in the LCA system than the RF system..

6.2 Systems

Two experimental IR systems were used in this study. Both systems used Inquiry 3.1p1 with its default values for indexing and retrieval. 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. In order to simplify the experimental design, and because we did not know quite how to implement LCA in a negative mode, we employed only the positive judgment feature from our TREC-7 study in TREC-8. The functionality and implementation of RF as positive term suggestion was identical to that of TREC-7.

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

¹David 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.

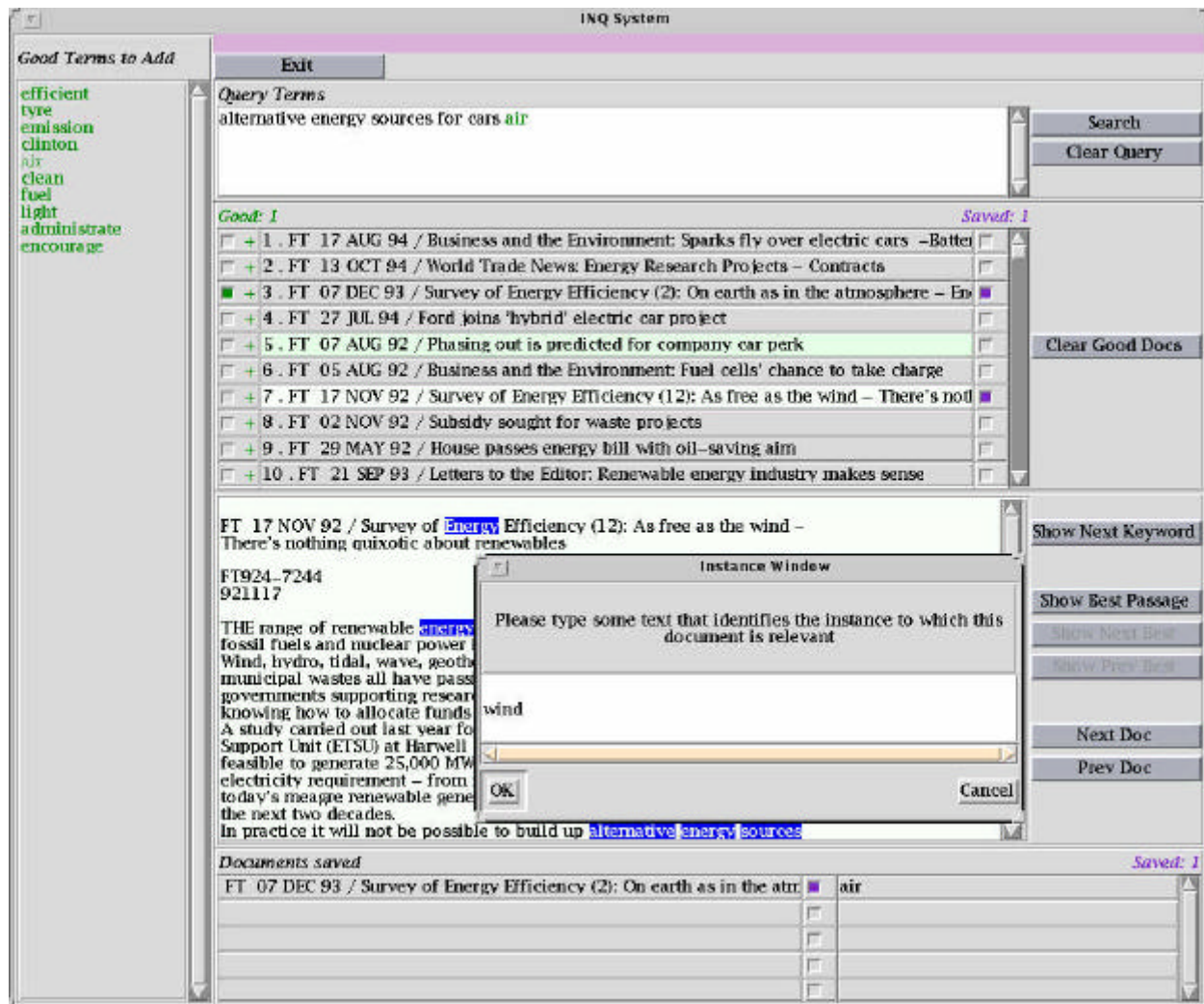


Figure 4. Screen shot of INQ-R interface.

Both systems used the same basic interface, which offers the functions and features described below. Figure 4 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. 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.

6.3 Subjects

As compared to all previous iterations, the backgrounds of subjects for this experiment were quite different. A total of 36 volunteer searchers, recruited from the Rutgers community, participated in this experiment. Most (89%) of the subjects were full time undergraduate students, who received compensation in the form of extra course credit for their participation. The subject group included 30 females and 6 males, whose ages ranged from 18 to 55. Twenty-seven of the subjects were pursuing a Bachelor’s degree, 7 were pursuing a graduate degree in library science and 2 were pursuing a PharmD.

6.4 Data Sources/Instruments

Each subject conducted six searches on six different topics. Three searches were conducted on the RF system and three more were conducted on the LCA system. Data collection instruments were identical to those in TREC-7, except that there were no questions about negative RF.

In order to assess the effectiveness of the two systems, subjects’ interactions with the systems were logged. Subjects’ performance on each of the systems was measured by aspectual recall, number of documents saved and the number of cycles per search.

6.5 Results

The quantitative ratings of the two systems provided subjects’ reactions to the ease of learning and ease of use of each of the systems and their understanding of how to use the systems. These ratings are shown in Table 8. Subjects responded using a five-point scale, where 1 represented “not at all,” 3 “somewhat” and 5 “extremely.” As indicated by Table 8, there was little difference in subjective responses to questions intended to measure the usability of the two systems.

	LCA Mean (SD)	RF Mean (SD)
Easy to learn to use	4.34 (.68)	4.39 (.60)
Easy to use	4.20 (.83)	4.28 (.74)
Understand how to use	4.03 (.94)	4.31 (.67)

Table 8 Usability ratings of the RF and LCA systems.

The use and effectiveness of RF and LCA was provided by the log data. Table 9 shows the data for these criteria. In the RF system, the average number of documents that were identified as relevant

and used to generate suggested query terms was 2.85 per search topic. While the total number of query terms per search was roughly equivalent in each system, there was a significant difference in the sources for these terms. Specifically, the number of suggested query terms selected by the user was significantly higher when using the LCA system ($M=4.41$) compared to the same users searching on the RF system ($M=1.87$), $t(214)=4.50$, $p<.001$. When converted to a percentage, terms suggested by the system comprised 44% of the total query terms for the LCA system and only 21% of the total query terms for the RF system. The number of terms entered into the query by the user (i.e. those not selected from the suggested terms list), was significantly higher for RF ($M=7.04$) than LCA ($M=5.59$), $t(214)=2.04$, $p<.05$. This suggests that in the RF system, users spent more effort generating query terms themselves, while in the LCA system users spent less effort generating query terms and selected more terms from those provided. Thus, our speculation from TREC-7 concerning why RF was not used, that it may have required too much effort to operate as implemented, appears to be supported with the results from TREC-8. In particular, it seems fair to say, if we assume that a system which requires less effort to accomplish the same level of performance is a better system, that the LCA system was significantly better than the RF system. That is, it is easier for users to accomplish the aspectual recall task with magical, diffuse term suggestion than with controlled, directed term suggestion.

	LCA Mean (SD)	RF Mean (SD)
Query terms, per search	10.0 (5.60)	8.91 (1.87)
User generated terms in query	5.59 (5.70)	7.04 (4.63)
System generated terms in query	4.41 (5.23)	1.87 (2.65)
% System generated terms per query	44%	21%
Uses of RF	NA	2.85 (3.40)
Unique titles displayed	123.53 (60.77)	128.81 (69.86)
Unique full-text viewed	21.88 (10.82)	25.41 (18.94)
Cycles per search	5.93 (3.25)	5.76 (3.95)
Aspectual recall	.24 (.16)	.26 (.17)
Documents saved	8.48 (4.33)	8.49 (4.72)

Table 9. Effort and performance measures for the LCA and RF systems

The average number of cycles per search was similar when subjects used the LCA system (5.93) and when subjects used the RF system (5.76). The total aspectual recall for the two systems was close (LCA $M=.24$, RF $M=.26$), as was the number of documents saved (LCA $M=8.48$, RF $M=8.49$). All of these differences were insignificant, suggesting that the effectiveness of the two systems is similar. Thus, while the actual use of the term suggestion feature of each of the systems differed significantly, there was no difference in the effectiveness of the two systems.

The qualitative data provide more specific insight into how searchers understood and used the features of the two retrieval systems. In accord with the quantitative data on subjects' perceived usability, subjects made many comments about how easy each system was to use, how enjoyable searching was on each of the systems and how understandable each of the systems were.

Although comments regarding the term suggestion feature of each system were generally positive, subjects tended to favor the execution of the feature in LCA. These comments lend support to our

hypothesis that a magical term suggestion feature would be preferred over one that requires additional effort to operate.

There were several comments regarding the quality of the terms suggested by LCA. One subject characterized many of the suggested terms as foreign, while another subject felt that the terms were too specific. These comments provide some insight into why there was no difference in the effectiveness measures of the two systems, even though there was a significant difference in the number of system suggested terms subjects added to their queries. It may be the case that the algorithm used for diffuse term suggestion does not optimally support subjects in the task in which they were engaged in TREC-9. While the implementation of the feature at the interface level may have been effective, its implementation at the system level may not have been.

Many subjects expressed frustration with the way in which the minus sign was implemented in both systems. Subjects were surprised to find that many terms in which they placed a minus sign in front of appeared in documents near the top of the retrieval list. These subjects made specific comparisons to the minus sign in Boolean retrieval models, indicating interference between that retrieval model and the model for our two systems. Interestingly, these comments regarding the minus sign are quite disparate from those made by subjects in TREC-7. Subjects in TREC-7 expressed approval of the feature and made no comments that indicated that the implementation of the feature caused confusion.

As in previous studies, subjects were quite impressed with the document navigation facilities for the systems. Subjects made specific comments about the how the highlighted query words in the document sped up the evaluation process. Subjects also made comments regarding the ease by which they could navigate through the documents. These comments lend continual support to the interface recommendations that were made at the end of TREC-5 and implemented in TREC-6 and TREC-7.

6.6 Discussion

The results from TREC-8 have allowed us to identify several possible problems with the system interface and its functionality. These problems concern how subjects engage the term suggestion feature at the interface level and how the term suggestion feature functions at the system level.

The results for TREC-8 revealed that when using LCA, subjects used the term suggestion feature to add terms to their queries significantly more often than they did when using RF, even though subjects' ratings of the usability of the two features were similar. In the LCA system, the term suggestion feature automatically generated words each time a subject performed a search. In the RF system, subjects had to mark a document good, in order to see any suggested terms. These results suggest that, although control over term suggestion is valued and useful, control over which terms are suggested is not. One reason for this is that as task complexity increases, users have fewer and fewer cognitive resources available for other activities (Norman, 1990). 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 document 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 did not exist in the LCA system. Hence, the measure of control that was gained was not worth the extra complexity it required.

The source of the suggested terms differed for each of the systems, so it is difficult to make a direct comparison with regard to the way in which the feature was implemented at the interface. Although subjects used significantly more of the terms suggested by LCA than RF in their queries, LCA did not outperform RF. The difference in use and performance of the two term suggestion features might be a result of the quality of terms suggested by each of the systems. According to the qualitative data, the general opinion seems to have been that many of the terms suggested in LCA 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 reasonable 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. It appears that diffuse term suggestion did not support the searching task that was required of users in TREC-8. This suggests that experimenting with the term-suggestion algorithm used by LCA in order to increase the quality of terms might lead to better task performance.

7. Discussion

The series of studies described in this paper began with an investigation of the use of a relatively standard version of automatic RF for support of query reformulation in the TREC Interactive Track task of aspectual recall, in a rather standard IR interface. Based on our analyses of the results of each study, modifications in both the functionality and interface of each system were made, and new issues were investigated in each successive study. The final study in the series investigated differences in preference, usability and effectiveness of two versions of term suggestion for user-controlled query reformulation, in a system with a significantly redesigned interface which responded to support for the new types of functionality, to user responses to the various interfaces and to the implementation of the different functionalities, and to HCI design principles. Although there are some obvious modifications and a few further issues to be investigated with respect to the system which implemented LCA in TREC-8 (for instance, somewhat better and more obvious explanation of the ranking and negation features; better implementation of LCA, or of some other similar term suggestion technique; testing negative term suggestion in the LCA mode), it seems to us that the end result is a total system design which is well suited to supporting interactive IR in at least the aspectual recall task. And it also seems to us that our results with respect to mode of term suggestion may well be applicable to other interactive IR tasks.

Although it is not possible to strictly compare the results of the different studies, especially because of differences in our subjects, it may nevertheless be possible to draw some general inferences from them. One is that suggestion of negative terms to be added to a query is something that was both valued and used by our subjects. The major problems with this feature seem to arise not through the feature itself, but rather through its implementation in the interface. Another is that term suggestion, as opposed to automatic query expansion, did indeed seem to be preferred by our subjects, and that it was not difficult for them to understand and use (even in a mode which was perhaps not well-suited to the task in which they were engaged). A third is that having a clear and adequate conceptual model of the system functionalities, whether implemented in a tutorial, in a help system, or directly through interface features, is absolutely necessary in order for the functions to be taken up. Just how this should be done still remains an open issue, but we believe that it would be best if this were to be accomplished to the greatest extent possible by making the system operation explicit in the interface and interaction design. We note, with respect to this issue, that

tutorial time appeared to decrease over the course of the last three studies, as our interface design paid more attention to presentation of the conceptual model.

8. Conclusions

We believe that more can be concluded from this series of studies in addition to the specific results with respect to the problem of support for query reformulation in the aspectual recall task. In particular, it seems that having conducted a series of related studies, using the same or similar methods and measures in each one, allowed us to develop a meaningful sequence of principled changes and issues to be investigated. Being able to do this within the TREC context, especially within the context of considering a single IR task throughout the entire series, meant that we were able to focus in depth on the effect of particular changes on a variety of measures. Although the strategy we adopted did not allow us to compare directly the results of one study to subsequent ones, it did allow us to build upon the results of earlier studies in a principled way. And this seems to us especially important, in that this particular characteristic has been sorely lacking in studies of interactive IR. It is also the case that we were able, through this series of studies, to investigate not only issues associated with system implementation and effectiveness, but also the deeper issues of what kinds of functions ought to be offered to IR system users, and why. An important reason that we were able to do this, we think, was the combination of many different types of measures and data that we used, in particular the use of both quantitative and qualitative methods and measures, and the explicit study of information seeking behavior. Finally, we believe that our general methodology may be most useful as an example of how to take account of the study and understanding of information seeking behavior (or, more generally, interaction with information) in the design of interactive IR systems.

9. Acknowledgements

We wish to thank first all of the subjects who participated in our TREC studies, donating over three hours of their time to helping us in our research. We also wish to thank all of the other people who participated with us as researchers in our TREC studies; their names are all indicated in the author lists in the citations to our TREC publications. Some of the research reported here was supported by the DARPA TIPSTER Phase 3 Program, under contract number , and by funding for Graduate Student support from the Rutgers Distributed Laboratory for Digital Libraries.

10. References

- Belkin, N.J., Cool, C., Koenemann, J., Ng, K.B., Park, S.Y. (1996) Using relevance feedback and ranking in interactive searching. In *The Fourth Text REtrieval Conference (TREC-4)*, D.K. Harman, ed. Washington, D.C.: GPO, 181-132.
- Belkin, N.J., Cabezas, A., Cool, C., Kim, K., Ng, K.B., Park, S.Y., Pressman, R., Rieh, S.Y., Savage, P., Xie, H. (1997) Rutgers interactive track at TREC-5. In *The Fifth Text REtrieval Conference (TREC-5)*, E. Voorhees and D. K. Harman, eds. Washington, D.C.: GPO, 257-266.
- Belkin, N.J., Perez Carballo, J., Cool, C., Lin, S., Park, S.Y., Rieh, S.Y., Savage, P., Sikora, C., Xie, H. (1998) Rutgers' TREC-6 interactive track experience. In *The Sixth Text REtrieval Conference (TREC-6)*, E. Voorhees and D.K. Harman, eds. Washington, D.C.: GPO, 597-610.
- 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.

Belkin, N.J., Cool, C., Head, J. Jeng, J., Kelly, D., Lin, S.J., Lobash, L., Park, S.Y., Savage-Knepshield, P. & Sikora, C. (2000) Relevance feedback versus Local Context Analysis as term suggestion devices: Rutgers' TREC-8 Interactive Track experience. In *TREC-8. Proceedings of the Eighth Text Retrieval Conference*. Washington, D.C.: GPO, in press.
http://trec.nist.gov/pubs/trec8/t8_proceedings.html

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.

Cool, C., Belkin, N.J. & 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).

Efthimiades, E.N. (1996) Query expansion. *Annual Review of Information Science and Technology*, v. 31: 121ff.

Egan, D.E., Remde, J.R., Gomez, L.M., Landauer, T.K., Eberhardt, J. & Lochbaum, C.C. (1989). Formative design-evaluation of SuperBook. *ACM Transactions on Information Systems*, 7(1), 30-57.

Fox, E.A., Hix, D., Nowell, L.T., Brueni, D.J., Wake, W.C., Heath, L.S., Rao, D. (1993) Users, user interfaces, and objects: Envision, a digital library. *Journal of the American Society for Information Science*, 44(8), 480-491.

Hearst, M.A., Karadi, C. (1997) Cat-a-Cone: An interactive interface for specifying searches and viewing retrieval results using a large category hierarchy. In *SIGIR 97, Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: ACM, 246-257.

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

Norman, K.L. (1990). *The Psychology of Menu Selection: Designing Cognitive Control at the Human/Computer Interface*. NJ: Ablex.

Rocchio, J.J., Jr. (1971). Relevance feedback in information retrieval. In: G. Salton, ed. *The SMART retrieval system: Experiments in automatic document processing*. Englewood Cliffs, NJ, Prentice-Hall: 313-323.

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

Swan, R.C., Allan, J. Aspect windows, 3-D visualizations, and indirect comparisons of information retrieval systems. In *SIGIR 98, Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: ACM, 173-181.

Walker, D.E., ed. (1971) *Interactive bibliographic search: The user/computer interface*. Montvale, NJ: AFIPS Press.

Williamson, C., Shneiderman, B. (1992) The Dynamic HomeFinder: Evaluating dynamic queries in a real-estate information exploration system. In *SIGIR 92, Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: ACM, 338-346.

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.