A Case For Interaction: A Study of Interactive Information Retrieval Behavior and Effectiveness

Jürgen Koenemann
Center for Cognitive Science
Rutgers University
Psychology Bldg.
Frelinghuysen Rd.
Piscataway, NJ 08855 USA
+1 908 445 6122
koeneman@ruccs.rutgers.edu

Nicholas J. Belkin
School of Communication,
Information, and Library Studies
Rutgers University
4 Huntington Street
New Brunswick, NJ 08901-1071 USA
+1 908 932 8585
nick@belkin.rutgers.edu

ABSTRACT

This study investigates the use and effectiveness of an advanced information retrieval (IR) system (INQUERY). 64 novice IR system users were studied in their use of a baseline version of INQUERY compared with one of three experimental versions, each offering a different level of interaction with a relevance feedback facility for automatic query reformulation. Results, in an information filtering task, indicate that: these subjects, after minimal training, were able to use the baseline system reasonably effectively; availability and use of relevance feedback increased retrieval effectiveness; and increased opportunity for user interaction with and control of relevance feedback made the interactions more efficient and usable while maintaining or increasing effectiveness.

Keywords:
information retrieval, user interfaces, evaluation, empirical studies, relevance feedback

INTRODUCTION

We are experiencing in our work and home environments a dramatic explosion of information sources that become available to an exponentially growing number of users. This has resulted in a shift in the profiles of users of online information systems: more users with no or minimal training in information retrieval (IR) have gained access to tools that were once the almost exclusive domain of librarians who served as intermediaries between end-users with their particular information needs and the information retrieval tools.

This situation has stimulated increasing interest in computerized tools that support end-users in their information seeking tasks. One important such situation is the information filtering (or *routing*) task, in which streams of information (such as email messages, newswire articles, or net news postings) are automatically filtered by a program based on specifications that are directly or indirectly obtained from the user. How these specifications should be obtained and used, and in particular, whether such programs should be autonomous or interactive are unresolved and controversial issues, which are explicitly addressed in the study reported here. This paper describes an experiment investigating the information seeking behavior of 64 novice searchers who used one of four versions of an advanced IR system to formulate routing queries for two given search topics. Each version offered a different level of interaction with a query formulation support mechanism called *relevance feedback*.

The paper is organized as follows: we first present the rationale for using an interactive, best-match, ranked-output, unstructured input, full-text IR system (INQUERY), and we discuss relevance feedback as a support tool for query reformulation. We then detail the design of the four different interfaces employed in our study, and describe the experiment which we conducted. The major portion of the paper focuses on a comparative description of the information retrieval behavior and effectiveness we observed in the use of these systems. We conclude with some general recommendations for the design of effective interfaces for information retrieval suggested by our results.

SUPPORTING END USERS

Users in all types of IR systems face the central difficulty of effective, interactive (re)formulation of queries which represent their information problems. Professional searchers using commercial IR systems have developed a variety of techniques and heuristics for addressing this difficulty in the context of Boolean query languages for exact-match, set-based retrieval from databases of indexed citations and abstracts of documents. Conversely, the difficulties faced by end-users with no training or experience in the use of these systems have been well documented. From experimental studies it has been known for some time that best-match, ranked-

CHI 96 Vancouver, BC Canada
© 1996 ACM 0-89791-777-4/96/04..$3.50

output retrieval techniques are in general superior in non-interactive settings to exact-match systems, such as commercial Boolean IR systems, in terms of recall and precision performance measures [2].

In response to such results, our study used the INQUERY retrieval engine developed at the University of Massachusetts [3]. The underlying mechanism of INQUERY is a Bayesian probabilistic inference network which provides rules for the computation of probabilistic belief values for each document in the collection. These belief values are based on the terms that are shared by the query and the *full text* of the document, and on the operators used to combine these terms. The system returns a *ranked* list of documents with documents that *best match* a given query being ranked at the top.

Research has shown that end-users with little or no training in query formulation have severe difficulties in making use of available operators and in mapping their intent to the appropriate logical query structure; it seems that systems which allow queries to be put in unstructured form allow easier query formulation than those which require Boolean structure, and are, at least for end-users, more effective [5, 13]. We therefore restricted queries to simple lists of terms. The only (implicit) operator allowed was the concatenation of terms to form multi-term phrases such as "automobile recall".

One particularly interesting and promising tool to support (or even replace) query reformulation in the context of these systems is *relevance feedback*. Relevance feedback modifies an existing query based on available relevance judgements for previously retrieved documents. For example, the system may add key terms from documents that the user has indicated as being relevant to the list of query terms, or the system may assign higher weights to terms in a user query that also appear in documents that have been marked "relevant" by the user. The goal of relevance feedback is to retrieve and rank highly those documents that are similar to the document(s) the user found to be relevant.

It is quite clear that automatic relevance feedback significantly improves retrieval performance in *fully automated* retrieval systems without user interaction, and with many relevance judgments[11]. Our concern is with determining how a relevance feedback component impacts the information seeking behavior and effectiveness of *novice* searchers in an *interactive* environment, and therefore with relatively few relevance judgments.

There have been many studies of user interaction with traditional boolean systems (e.g. [8]) and some studies that have focused on novel interaction techniques (e.g. [1, 10]). A few observational studies are concerned with relevance feedback [4, 6] but we are not aware of studies that have looked at relevance feedback in an experimental setting except for our own work in the context of the interactive track of TREC-3 [9].

A central question for the design of interactive systems in general is the amount of knowledge a user is required or expected to have about the functioning of the system and the level of control a user can exert. We share the "task-centered" view that interfaces for the occasional user should hide as much as possible of the inner workings of a system and should instead present users with a view that focuses on the user's task. However, the question arises of *how much* knowledge and control a user should have in order to best interact with components such as relevance feedback that are central to the user task, here the formulation of an information need. At one extreme, the existence of such a tool can be completely hidden from the user: the set of "relevant" documents could be determined by some algorithm that takes as input a user's behavior such as the viewing, saving, or printing of documents. The other extreme would be a system that provides the user with complete control over the feedback mechanism: a user could provide lists of "good" documents to the mechanism, manipulate the query modifications (changed weights and added terms) suggested by the relevance feedback component, and even adjust internal parameters such as belief thresholds. Between these two extremes there is a large space of possible designs; the goal of this study was to explore this space through the design of four interfaces described in the next section.

THE SYSTEMS

We designed and implemented (Tcl/Tk) a **baseline** interface to the INQUERY (V.2.1) retrieval engine, RU-INQUERY, (Figure 1) that allowed users to enter queries and to view the results. Users entered one-by-one single or compound terms (phrases) into the term entry window. An entered term was checked against the database and was either rejected if it did not appear in the collection or added in its stemmed form to the query term list. Words that appeared on a stop word list were ignored. Subjects submitted a query for retrieval by hitting the `Run Query` button. The total number of retrieved documents and the titles of the top five (5) ranked retrieved documents were displayed. Users could scroll through the entire list of titles and look at the full text of any document by double-clicking on its title. A check box next to each title allowed keeping track of documents between iterations. A single-step undo mechanism allowed users to return to the previously run query. The system blocked most inappropriate user activities in order to prevent error episodes. Subjects could view the results of a query, and then reformulate the query by manually adding or deleting terms.

Performance in this baseline system was compared with one of three experimental versions, each offering (in addition to the baseline facilities) a different level of interaction with a relevance feedback facility for automatic query reformulation.

In the **opaque** interface relevance feedback was treated as a black-box, a "magical" tool that hid the functionality of the relevance feedback component from the user. Searchers were simply told that marking documents as relevant would cause the system to retrieve additional documents that were similar to the ones marked as relevant and/or that similar documents would be ranked higher. Thus, users only needed to acquire

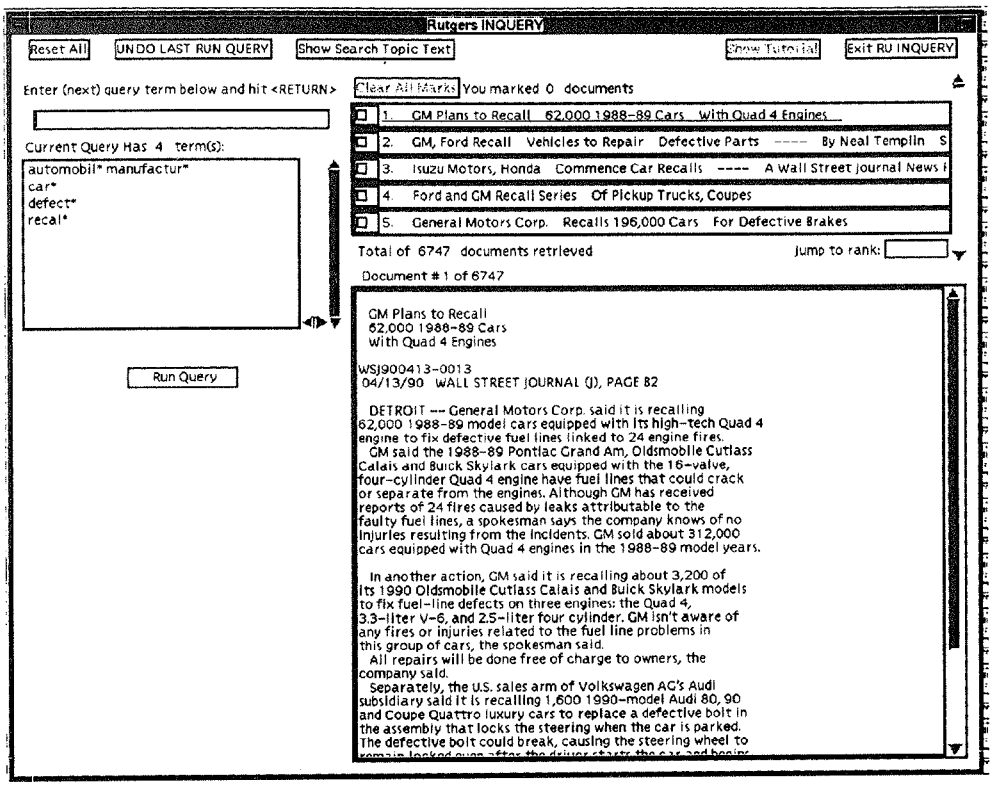

Figure 1: The RU-INQUERY Interface: Base Version Without Relevance Feedback

minimal knowledge about the feedback tool and could focus on the evaluation of documents rather than on the resulting reformulation of queries. Users marked a document "relevant" by clicking on the check box next to its title. Relevance feedback took effect the next time a query was run.

The **transparent** relevance feedback interface had the same functionality as the opaque version with the following addition: after a relevance feedback query had been executed searchers were shown the list of terms that had been added to the user-entered terms by the relevance feedback component. This additional information could be used to develop a more accurate model of the tool. Feedback terms could also be used as a source for future queries.

The **penetrable** relevance feedback interface (Figure 2) took the transparent version one step further: in addition to providing information about the functioning of the relevance feedback mechanism it provided the ability to manipulate the output of the relevance feedback component *prior* to query evaluation. The execution of the query was interrupted and the user was presented with the list of terms suggested by the feedback component. Users had the opportunity to add none, some, or all of the suggested terms prior to the continuation of the query evaluation. For example, the user might select only terms that appear to be central to the search topic.

THE EXPERIMENT

Given the research results and considerations discussed above, our experiment was designed to investigate the following questions:

- Can best-match, ranked-output, full-text retrieval systems combined with an operator-free query language be used in an effective way by end-users with little training?
- Is relevance feedback effective? That is, do users using one of the three systems with relevance feedback perform better on the routing task compared to subjects who use the baseline system without relevance feedback?
- Is user knowledge about the output of the relevance feedback system helpful? If so, users in the transparent and penetrable conditions should perform better than subjects using the opaque relevance feedback system.
- Is user control over the operation of the relevance feedback system helpful? If so, users in the penetrable condition should perform better than subjects using the transparent feedback system and better than users using the opaque system.
- How do different levels of interaction impact the information seeking behavior of users such as the number of queries developed and the way these queries are formed?

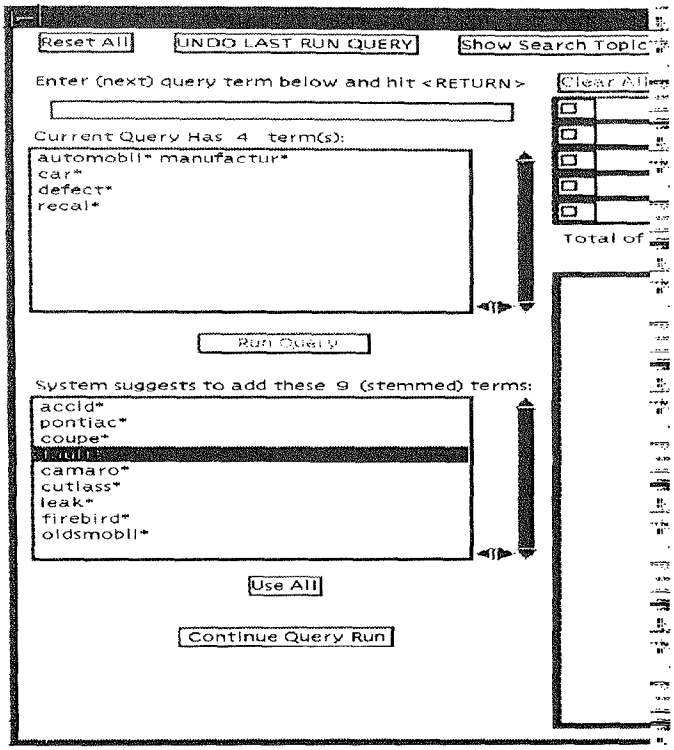

Figure 2: **Penetrable Relevance Feedback Version** (View of Feedback Component)

Subjects

64 Rutgers University Undergraduates (43 females, 21 males) with (self-reported) native-like English competence participated in the study. None of the subjects had any formal training in library and information science. Their IR searching experience was limited to the occasional use of a computerized library card catalog. In addition, a few isolated instances of Psych-Lit searches and WorldWideWeb browsing were reported.

Materials

The system and interfaces used are described in the previous section. The INQUERY system and the RU-INQUERY interfaces were installed on a SPARC 2 workstation and users interacted with it via a networked SUN 3/50 with monochrome monitor, standard keyboard, and mouse.

Recently, in the context of the ARPA-sponsored TIPSTER project and the ARPA/NIST TREC studies [7], a test collection of reasonably realistic information problems, together with a large database (ca. 3GB) of the full texts of a wide variety of documents, and relevance judgments on up to 2000 documents per information problem has become available. For our experiment, we used a subset of the TREC test collection, consisting of 74,520 articles from the Wall Street Journal between 1990 and 1992.

Two search topics (162 - *Automobile Recalls* and 165 - *Tobacco Advertising and the Young*) were selected from the set of TREC search topics. Each topic consisted of a title, a short description, and narrative that spelled out what would constitute a relevant article (figure 3).

Topic: Tobacco company advertising and the young
Description: A document will provide information on what is a widely held opinion that the tobacco industry aims its advertising at the young.
Narrative: A relevant document must report on tobacco company advertising and its relation to young people. A relevant document can address either side of the question: (1) Do tobacco companies consciously target the young, or (2) As the tobacco industry argues, is this an erroneous public perception. The "young" may be identified as youth, children, adolescents, teenagers, high school students, and college students.

Figure 3: Search Topic Definition For Topic 165

Each of the unique 2000 retrieved documents was rated by the first author as being relevant or not relevant to the topic on hand. These ratings were compared to TREC relevance judgments made by the originators of the topic descriptions that were available for a subset of about 560 of the 2000 retrieved documents. The inter-rater agreement between the experimenter and the TREC evaluators was almost perfect (98%) for topic 162 and very good for topic 165 (94%); cases of disagreement were resolved by careful reexamination of the documents. These judgements served as the basis for the performance evaluation.

The task for this study was a query construction task, i.e. subjects had to develop a final routing query for a given information problem, or *topic*, that could be run repeatedly against changing document collections. Specifically, we asked subjects to devise a single final query that retrieved at least 30 documents from the current collection, of which as many documents as possible in the top 30 were relevant to the provided topic. A focus on the top 30 articles mirrors realistic retrieval situations in which users are only interested in a small number of documents. We focused on retrieval for the current collection because a previous study [9] had shown that even experienced searchers did not reason about the effect of changing collection content and collection characteristics on query effectiveness. The task of developing a final query is different from a standard adhoc retrieval task in which users focus on finding relevant documents throughout the interaction and across multiple queries and where the end result is a set of documents, not a single final query. Our routing task is typical of a situation where a user wants to develop a profile for the filtering of information and uses an existing current collection to develop and test this profile.

A short (30-40 minutes long) interactive tutorial was integrated into the interface. It guided subjects through a sequence of exercises using the baseline system without relevance feedback in the form of a sample search. Three additional, very short online tutorials were developed that taught the use of the respective relevance feedback interface.

Experimental Design and Procedure

Subjects performed two (2) searches: all subjects used the baseline system without relevance feedback for their first search, followed by a second search on a different topic using either one of the three relevance feedback systems or continuing the use of the baseline system (control group). The order of topics was counterbalanced between searches, leading to eight (8) different conditions based on the topic order employed (2) and the type of system used during the second search (4). Subjects were assigned to one of the eight conditions in a block-randomized fashion.

After giving their informed consent and filling out the online questionnaire subjects worked through the online tutorial at their own pace. After subjects ended the tutorial or after the allotted time of 45 minutes had expired they were given twenty minutes to formulate a routing query for their first topic. At the twenty minute mark subjects were told that time was up and that they should wrap up the current action. After a short break subjects worked for up to ten minutes either through one of the three relevance feedback tutorials (experimental groups) or returned to a review of the original baseline tutorial (control group). Next, subjects were again given twenty minutes to formulate a routing query for the second topic, using the system and interface they just had learned about. A short interview concluded the experiment. Users were asked to comment on their overall experience and to state which interface they liked better.

All interactions with the system were automatically recorded by the system, creating a timed log of user and system actions. Subjects were instructed to think aloud during their two searches and the utterances were video-taped along with the current screen image. Independent of actual performance, subjects were told that they did well.

RESULTS AND DISCUSSION

Information Retrieval Evaluation

Although IR has a long history and extensive experience of evaluation (see, e.g. [12]), evaluation of interactive IR systems is still in its infancy. The two standard measures of retrieval effectiveness are **precision**, the number of relevant retrieved documents over the total number of retrieved documents, and **recall**, the ratio of relevant retrieved documents to the total number of (known) relevant documents. For the purposes of this paper we use *precision at 30 retrieved documents*, the cut-off level determined by the task that we set for our searchers in the experiment.

We used the non-parametric Kruskal-Wallis and Mann-Whitney U tests with corrections for tied ranks to analyze precision results since the normality and homogeneity assumptions for an ANOVA were violated. We report medians (M), the interquartile range (IQR), means ($\overline{X}$), and standard deviations (s) for descriptive purposes.

Training

Searchers trained on the baseline system on average about 26 minutes; the fastest subject finished the tutorial in 9.5 minutes whereas one subject had to be stopped after 45 minutes. There were no differences in training time between conditions. There was no correlation between training times and performance on the first search topic ($r = .08, p > .5$).

Note that we designed the training to ensure (for experimental purposes) that all subjects were comfortable with the baseline system and had used all its features. Thus, the training times we observed do not reflect the minimum time that would be required to make use of the system in a walk-up situation. Indeed, a few subjects "jumped the gun" and issued quite successful queries prior to studying the tutorial.

Baseline: Search Task 1

There were no differences in mean search-times for the first search by topic or experimental condition. All subjects used up (most of) the allotted time (17-21 minutes).

Three quarters of all subjects retrieved at least 8 relevant documents and half the subjects located over 12 relevant documents in the top 30 with the baseline system ($M = .40$, $IQR = .27 - .67$, $\overline{X} = .45$, $s = .24$). The lowest precision was 0, i.e. no relevant documents were ranked in the top 30 (but appeared in lower ranks); the best final queries had a precision of .87. Topic 162 on "Automobile Recall" had 2.5 times as many relevant documents in the collection compared to topic 165 on "Tobacco Advertising And the Young" (83 versus 33). As a result, finding relevant documents and ranking them high was significantly more difficult for the latter: the median precision for topic 162 was $M = .67$ ($IQR = .28 - .78$, $\overline{X} = .56$, $s = .28$) as compared to a median precision of $M = .33$ ($IQR = .27 - .42$, $\overline{X} = .34$, $s = .09$) for topic 165 (Kruskal-Wallis $H' = 10.17$, $df = 1, p \leq .01$). The average and median performance on the first search did not differ for the four experimental groups that searched for topic 165 in their baseline search. Among the groups that were given topic 162 first, only the group that would later use the opaque relevance feedback version performed on a level that was borderline significantly lower than the other groups.

Relevance Feedback Training

After completion of the first search subjects were trained on the feedback mechanism or reviewed the baseline tutorial (control). Subjects completed the second part of the tutorial on average within 8 minutes. Subjects in the control group finished their review on average in 6.5 minutes; the opaque feedback group who had only to learn about the marking of documents for relevance feedback but not about interaction with the output of the relevance feedback mechanism finished equally fast ($\overline{X} = 6.5$ minutes). Subjects in the transparent and penetrable condition used the allotted 10 minutes to finish the tutorial.

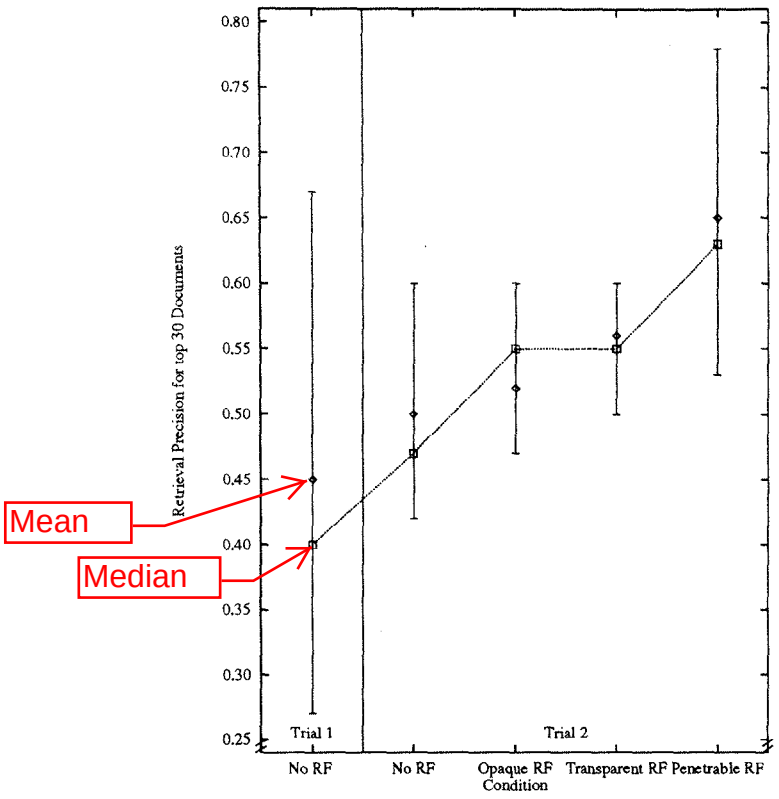

Figure 4: **Retrieval Precision at 30 Documents for Final Queries.** Given are median precision (□), mean (◇), and the interquartile range for search task 1 (No RF, n=64) and for each of the four conditions used for search task 2 (n=16).

Second Search Task - Effectiveness

At the top 30 cutoff level we found (see Figure 4) that overall retrieval performance differed significantly between conditions (Kruskal-Wallis test, $H' = 8.27, df = 3, p < .05$). Subjects who used relevance feedback had 17% to 34% better performance than subjects who continued in the control condition. Subjects in the penetrable feedback condition performed as a group 15% better than subjects in the opaque and transparent feedback conditions. Pairwise comparisons were, however, not significant except for the difference between the penetrable and baseline conditions (Mann-Whitney $U' = 72, p < .05$) due to large within-group variations and the relatively small sample size. The same general pattern held when data were analyzed separately for both topics. Again, topic 165 proved to be more difficult compared to topic 162, but the difference was less pronounced compared to the baseline search ($M_{165} = .50$ vs. $M_{162} = .63$, Mann-Whitney $U' = 212, p < .05$).

This overall performance is quite impressive, considering that our users had no search experience. The median performance exceeded the performance of INQUERY in fully automated mode (.23 for topic 162 and .40 for topic 165 on this collection), INQUERY being one of the best performing systems in the TREC context.

We can compare the performance on the two topics by considering R-precision, the proportion of relevant retrieved documents at the number of relevant documents in the collection, i.e. precision at the point where 100% recall would be possible. Topic 162 had 83 relevant documents compared to only 33 relevant documents for topic 165. The median R-precision for all subjects searching on topic 162 was .39 as compared to .48 for all subjects working on topic 165. Thus, subjects who searched on topic 165 were able to retrieve a larger proportion of those relevant documents known to be in the collection

Although the task specified 30 documents as the cutoff level, it is instructive to note the performance at other cutoffs as well. At 100 retrieved documents the relative performances mirror the results from the top 30 cutoff: the control group did worst ($M = .22$), the penetrable group did almost 50% better ($M = .32$), and the groups in the opaque and transparent condition fell in between ($M = .27$ and $M = .26$, respectively). If one only considers the top 5 and top 10 ranked documents, an even more dramatic (and significant) difference in favor of relevance feedback materializes: typically 5 out of 5 and 9 out of the top 10 documents were relevant for subjects in the feedback conditions whereas subjects in the control group managed to manually design final queries that had 3 out of the top 5 and 7 out of the top 10 being relevant.

A potential limitation of our study is the use of relevance feedback as a simple query expansion tool without a re-weighting of user terms. It remains an open issue whether or not the more massive query expansion through automatic, opaque relevance feedback may do better under particular term re-weighting schemes and how search tasks and collection characteristics impact performance of the various interfaces.

Second Search Task – Behavior

Search times for the second search did not differ significantly for either topics or conditions: 80% of the subjects used up the allotted time of 20 minutes, 19% searched for 16-19 min., and one subject stopped after 11 minutes, having found a very good query.

Interactive query formulation is an iterative process of query design and entry, query execution, and query evaluation. The number of iterations for each condition is depicted in figure 5. There were significant differences in the number of iterations (Kruskal Wallis $H' = 8.52, p < .05$). The control group ($M = 7, \overline{X} = 8.2$) and opaque feedback group ($M = 7, \overline{X} = 7.7$) went through roughly the same number of iterations, whereas subjects in the transparent condition performed an extra iteration ($M = 8, \overline{X} = 8.8$). The penetrable feedback group needed less iterations ($M = 6, \overline{X} = 5.8$) to develop equally good or better queries. Interaction was clearly beneficial: initial queries had much lower precision (Topic 162: $M_1 = .06$ vs. $M = .63$, Topic 165: $M_1 = .33$ vs. M=.50) and on average precision increased with the first few iterations.

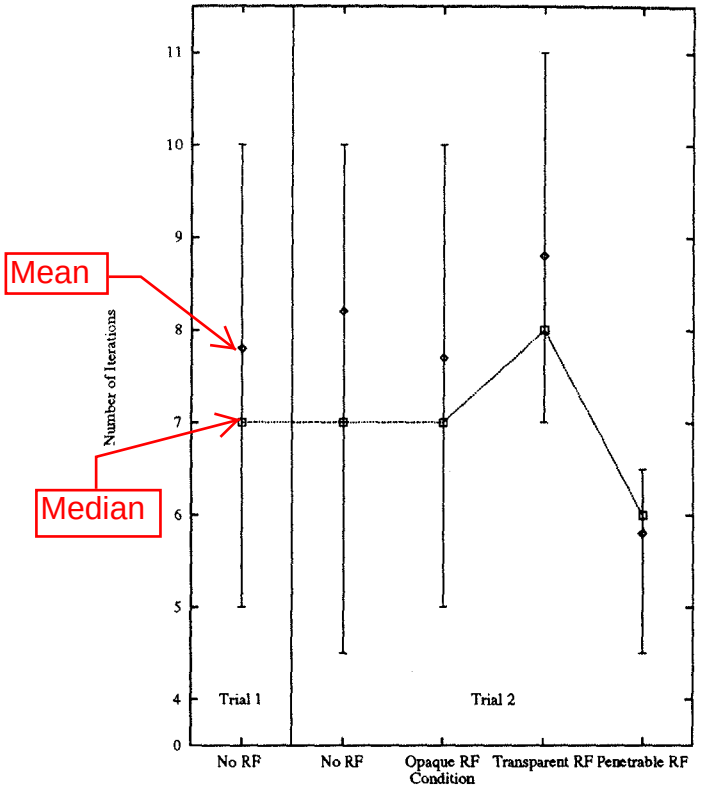

Figure 5: **Number of Iterations** (Number of Unique Queries Developed, Run, and Evaluated); both topics combined. Given are median number of iterations (□), the mean (◇), and the interquartile range for search task 1 (No RF, n=64) and for each of the four search task 2 conditions (n=16).

The four different interfaces shaped how subjects constructed their final queries over the course of the interaction. Table 1 provides a summary of the analysis of final queries and their constituent terms. The rightmost column lists the total number of query terms for each of the four conditions and for the two search topics separately and combined. Users in the baseline condition without relevance feedback entered on average 6.4 terms ($s = 4.2$), 1.8 of which were compound terms of 2 or 3 words each. The use of relevance feedback led to a dramatic increase in query length. In the opaque condition, the final query comprised on average 35.5 terms ($s = 17.92$). This was not a result of users entering more terms ($\overline{X} = 7.3, s = 9.36$), but a result of the automatic query expansion through relevance feedback which added on average 28.2 terms ($s = 15.38$). This expansion was caused by searchers having marked on average 12.6 documents as being relevant when running the final query. The average query length for queries by subjects in the transparent condition is even larger: final queries had on average 41.2 terms, 30.3 of which were automatically added by the feedback component based on an average of 13.6 documents being marked as relevant. User queries had on average about 11 terms; only 3.8 of these terms were entered by the user, almost twice that many

Mean Number & Sources of Query Terms					
Relevance Feedback Condition	User Controlled: User Typed	User Controlled: Copy from RF	User Controlled: Σ	Added by RF SYS	Σ
Topic 162:					
None	6.9	n/a	6.9	n/a	6.9
Opaque	10.9	n/a	10.9	35.9	46.8
Transparent	3.3	9.1	12.4	42.8	55.1
Penetrable	6.3	24.4	30.6	n/a	30.6
Topic 165:					
None	6.0	n/a	6.0	n/a	6.0
Opaque	3.8	n/a	3.8	20.5	24.3
Transparent	4.3	5.3	9.5	17.8	27.3
Penetrable	3.3	9.5	12.8	n/a	12.8
162&165:					
None	6.4	n/a	6.4	n/a	6.4
Opaque	7.3	n/a	7.3	28.2	35.5
Transparent	3.8	7.2	10.9	30.3	41.2
Penetrable	4.8	16.9	21.7	n/a	21.7

Table 1: Number and Sources of Query Terms in Final Queries for Second Search, Both Topics. Given are the mean number of query terms for each of three possible sources: user entry, copied by user from relevance feedback output, and automatic query expansion through relevance feedback.

($\overline{X} = 7.2$) were terms the user had copied from the display of automatically added terms or through relevance feedback expansion during prior iterations.

Finally, the average length of final queries by subjects in the penetrable feedback condition was 21.7 ($s = 29.4$). This number is lower compared to the other feedback conditions because on average final queries contained only 16.9 terms ($s = 27.3$) suggested by the feedback component. Subjects in the penetrable condition marked a comparable number of documents as being relevant but were quite selective in using suggested feedback terms: throughout the search they used only 349 of a total of 1259 suggested terms. Furthermore, five (5) instances in which users copied all suggested terms in one step accounted for 233 of these terms. An initial content analysis of copied terms suggests that users primarily copied those terms which had a clear and central semantic relation to the search topic. At the same time, subjects entered fewer terms ($\overline{X} = 4.8, s = 3.94$) manually. Indeed, subjects commented that they preferred the "lazy" approach of term selection over term generation.

Given that relevance feedback depends on the marking of relevant documents and that topic 165 was much harder than topic 162, it is not surprising that there were significant differences in query length between the two topics ($F(1, 56) = 16.3, p < .001$). However, table 1 shows that the pattern described above also holds for each topic in isolation.

CONCLUSIONS

Subjects used our system and interface quite effectively and very few usability problems surfaced. Users had little problem formulating their queries and the observed retrieval effectiveness of final queries supports the view that interactive best-match, ranked-output, unstructured input, full-text retrieval systems are suitable tools for end users with limited search experience. Users clearly benefited from the opportunity to revise queries in an iterative process.

Overall, relevance feedback appears to be a beneficial mechanism. All users declared their preference for the relevance feedback mechanism over the baseline system in a post-search interview. Relevance Feedback did improve median performance by 10% (or more) at 30 retrieved documents. Even larger gains were observed if one only considers the top 10 or top 20 documents retrieved, a situation quite common in many application domains. However, substantial subject and topic variation precluded statistical significant findings at the chosen level of $p \leq .05$.

Subjects that interacted with the penetrable version of relevance feedback did best in our study overall and significantly better than users of the baseline system without relevance feedback.

Perceived performance, trust in the system, and subjective usability are important issues, in particular for such sensitive domains as the filtering of one's personal email or news. Subjects really "liked" the penetrable version that allowed them to manipulate the list of suggested terms. Indeed, subjects in the opaque condition and in another study currently in progress routinely expressed their desire to "see and control" what the feedback component did to their queries. People commented that using the feedback component as a suggestion device made them "lazy": the task of generating terms was replaced by the easier task of term selection. Furthermore, users in the penetrable condition needed fewer iterations to achieve results comparable to, or better than the other, less interactive, feedback conditions.

Although a number of issues need to be further addressed, these results suggest that interfaces for IR systems for end users should be designed to support interactive collaboration in query formulation and reformulation between the users and the "intelligent" programs which support them.

ACKNOWLEDGMENTS

This work is supported by NIST Cooperative Agreement 70NANB5H0050. Thanks to the Rutgers University Center for Cognitive Science (RuCCS) for providing equipment loans and infrastructure support and to Jamie Callan, Bruce Croft, and Steve Harding of the Center for Intelligent Information Retrieval at the University of Massachusetts at Amherst for their unstinting support of our use of INQUERY.

REFERENCES

1 Ahlberg, C., and Shneiderman, B. Visual information seeking: Tight coupling of dynamic query filters with starfield displays. In *Proc. CHI'94*. ACM Press, New York, 1994, pp. 313–317.

2 Belkin, N. J., and Croft, W. B. Retrieval techniques. In *ARIST*, M. E. Williams, Ed. Elsevier, 1987, ch. 4, pp. 109–145.

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

4 Efthimiadis, E. *Interactive Query expansion and Relevance Feedback for document Retrieval Systems*. PhD thesis, City University, London, UK, 1992.

5 Frei, H., and Qiu, Y. Effectiveness of weighted searching in an operational IR environment. In *Information Retrieval 93: von der Modellierung zur Anwendung; Proc. der 1. Tagung Information Retrieval '93* (Konstanz, 1993), Universitätsverlag Konstanz, pp. 41–54.

6 Hancock-Beaulieu, M., and Walker, S. An evaluation of automatic query expansion in an online library catalogue. *Journal of Documentation 48*, 4 (1992), 406–421.

7 Harman, D. Overview of the second text retrieval conference. *IPM 31*, 3 (1995), 271–289.

8 Hewett, T., and Scott, S. The use of thinking-out-loud and protocol analysis in development of a process model of interactive database searching. In *Proceedings of INTERACT'87* (Amsterdam, 1987), Elsevier, pp. 51–56.

9 Koenemann, J., Quatrain, R., Cool, C., and Belkin, N. J. New tools and old habits: The interactive searching behavior of expert online searchers using inquery. In *TREC-3. Proceedings of the Third Text REtrieval Conference* (Washington, D.C., 1995), D. Harman, Ed., Government Printing Office, pp. 144–177.

10 Landauer, T., Egan, D., Remde, J., Lesk, M., Lochbaum, C., and Ketchum, D. Enhancing the usability of text through computer delivery. In *Hypertext: A psychological perspective*, C. McKnight, A. Dillon, and J. Richardson, Eds. Ellis Horwood, New York:, 1993, pp. 72–136.

11 Salton, G., and Buckley, C. Improving retrieval performance by relevance feedback. *JASIS 41*, 4 (1990), 288–297.

12 Sparck Jones, K. *Information Retrieval Experiment*. Butterworths, London, 1981.

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