PERGAMON

Information Processing and Management 37 (2001) 521–541

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Passage feedback with IRIS

Kiduk Yang *, Kelly L. Maglaughlin, Gregory B. Newby

School of Information and Library Science, University of North Carolina, Chapel Hill, NC 27599-3360, USA

Accepted 19 September 2000

Abstract

We compare a user-defined passage feedback (pf) system to a document feedback (df) system. Df employed the adaptive linear model for retrieval, while pf used weighted query expansion based on positive and negative feedback. Twenty-four searchers performed the same six tasks in varying search and system-order per TREC-8 guidelines. We hypothesized that pf, which featured interactive query expansion, would outperform df, which relied on automatic query expansion. Initial analysis appeared to reject this hypothesis, as df showed slightly higher overall performance than pf. However, analysis by system-order groups indicates only the first pf use had lower performance. These data suggest that pf was more difficult to learn than df, though the second pf use yielded competitive performance. If performance of pf is indeed affected by learning, an improved pf system with usability enhancements may prove to be an effective mechanism for interactive information retrieval. © 2001 Elsevier Science Ltd. All rights reserved.

Keywords: Passage feedback; Interactive information retrieval; Text Retrieval Conference

1. Introduction

A perspective on information retrieval (IR) was presented by Luhn (1957), who suggested an automatic text retrieval system based on comparison of word representations between the document and the query. In this perspective, the objective of IR is to find information units from an information collection that best match the query put to it by a user. Consequently, the main concerns of IR research have historically been on how best to match documents (i.e., information units) to the query and how to represent them internally to facilitate that matching process.

* Corresponding author. Tel.: +1-919-962-8366; fax: +1-919-962-8071.

E-mail addresses: yangk@ils.unc.edu (K. Yang), maglk@ils.unc.edu (K.L. Maglaughlin), gbnewby@ils.unc.edu (G.B. Newby).

PII: S0306-4573(00)00060-1

When IR approaches are taken out of the laboratory and applied to operational settings involving real users, however, results have often been less than satisfactory. One source for the evidence of shortcomings of experimental IR systems put to real-world tests is the interactive IR experiments in TREC (Beaulieu, Robertson, & Rasmussen, 1996; Robertson, Walker, & Beaulieu, 1997; Sumner, Yang, Akers, & Shaw, 1998). Why has IR research based on Luhn's tradition often failed to yield systems that are effective in operational settings? Are there fundamental flaws in its theories and/or methodologies that invalidate the work of past four decades? Or is IR simply an underdeveloped science, which has yet to find the optimum answers for the questions it asks?

In our opinion, this problem of theory versus practice in traditional IR stems from neither the invalidity of past research nor the immaturity of its science, but is a natural consequence of the questions it asks. The key question IR has traditionally asked, namely, how to find documents that best match an information need, assumes implicitly the existence and availability of expressions of users' information needs, and presupposes an unambiguous and consistent mapping of meaning to words.

This means the scope of the traditional IR perspective is quite restricted and leaves out a large portion of the actual information seeking (IS) process. People, without whom the science of IR is neither possible nor meaningful, have often been left out of the IR equation. More recently, researchers such as Belkin, Brooks, and Oddy (1982); Ingwersen (1996); Saracevic (1996); and Dervin (1997) have begun to ask questions about the IS process itself, often with a focus on human information seekers.

Careful examination of various aspects of IS has revealed the dynamic and complex nature of human information seeking behavior. IR systems based on traditional IR models are often unequipped to deal with such behaviors. The historical IR scenario, where users of IR systems were able to specify a "good" query, has not been the norm observed in interaction between searchers and intermediaries (Saracevic, Mokros, Su, & Spink, 1991). Taylor (1968) characterized information seeking as a process of negotiation and communication that progresses dynamically from visceral to compromised state of information need. Such findings highlight the "pre-search" stage of IS that have been left out of IR models.

Findings from research on information seeking indicated a clear inadequacy in the traditional IR paradigm and suggested a new IR paradigm that can accommodate the entire IS process. As a result, various new models of IR have emerged. Recognizing the difficulty involved in arriving at the precise expression of an information need, Belkin (1980) said that IR systems should focus on exploring the anomalous states of knowledge (ASK) underlying the information need. Ingwersen (1996), viewing IR as a process of cognition, proposed a cognitive model of IR, where he attempted to describe IR in terms of cognitive interactions. Equating IR to an information-seeking activity, Saracevic (1996) argued that the user is the central component and that interaction is the central process of IR.

Though these new models of IR may differ in focus and scope, a common underlying theme can characterize them as "interactive IR models". As the name suggests, the focus of interactive IR research is on the interaction between the IR system and its user rather than on the methodology of information representation and retrieval by the system. The objective of interactive IR research is to answer the question, "how do various types of interaction in IS process affect the retrieval outcome"? As a first step of answering that question, much research in interactive IR has concentrated on identifying the various aspects of the interaction process by modeling and analyzing the user-intermediary interactions (Brooks, Belkin, & Daniels, 1986; Belkin, 1987; Saracevic et al., 1991; Spink & Cool, 1992; Belkin, Cool, Stein, & Thiel, 1994; Saracevic, 1996).

Unfortunately, relatively little interactive IR research till date has ventured beyond this first step to offer a concrete framework in which to measure the effects of IR interactions on the retrieval outcome. Though there currently are numerous suggestions on building an interactive IR system (Meadows, Hewett, & Aversa, 1982; Shuman, 1989; Clarkson, 1992; Beaulieu, 1997; Goker, 1997; Jones, Gatford, Do, & Walker, 1997), none offer a comprehensive solution that is firmly grounded in empirical evidence. Given the complexity and unpredictability of human behavior, it is not surprising that interactive IR models proposed so far remain mostly fragmented and abstract. Consequently, there is a need for IR research that involves real users that interact with real systems, where the effects of various user-system interactions on retrieval results can be studied empirically.

Interactive IR experiments, such as TREC interactive track experiments, present valuable opportunities for such studies. In these studies, various aspects of an interactive retrieval process can be controlled to a certain degree to provide a better understanding of how user-system interactions affect the retrieval outcome. Consequently, our goal in TREC interactive experiments for the past three years has been to enhance our understanding of the impact of user-system interaction on retrieval performance in order to discover the ways to optimize contributions from both the user and the system. The results of our TREC interactive experiments have been confounding and unsatisfactory at times, but they have always been illuminating.

Our first participation in the interactive experiment was in TREC-6, where we developed an interactive retrieval system called information retrieval interactive system (IRIS),[1] which implemented modified versions of the feedback models with a three-valued scale of relevance and reduced feedback query expansion (Sumner et al., 1998). The goal of the IRIS design in TREC-6 was to provide users with ample opportunities to interact with the system throughout the search process. For example, users could supplement the initial query by choosing from a list of statistically significant two-word collocations or add and delete query terms as well as change their weights at each search iteration. Unfortunately, it was difficult to tell how much effect each IRIS feature had on retrieval outcomes due to such factors as strong searcher effects and major differences between the experimental and control systems.

Thus, we attempted to isolate the effect of a given system feature in our TREC-7 interactive experiment by making the experimental and control systems identical, save for the feature we were studying (Yang, Maglaughlin, Meho, & Sumner, 1999). In one interactive experiment, the difference between the experimental and control systems was the display and modification capability of term weights. In another experiment, the difference was the relevance feedback process, where one system utilized user-defined passages to formulate the feedback query, while

[1] A prior version of IRIS was developed by Kiduk Yang, Kristin Chaffin, Sean Semone and Lisa Wilcox at the School of Information and Library Science (SILS) at the University of North Carolina. They worked under the supervision of William Shaw and Robert Losee.

the other system employed document-level relevance judgments to create a feedback query with the adaptive linear model. Despite our attempts to control the experiment to comparison of specific system features, we suspected there were other factors at play that made the results inconclusive and difficult to analyze. Though inconclusive, our TREC-7 results suggested passage feedback (pf) could be an effective alternative to conventional document feedback (df). There were also some indications that searcher intervention might have an advantageous influence on the retrieval performance.

We simplified the experiment even further in TREC-8, and focused on comparing the effectiveness of a user-defined pf system with a conventional df system. As in TREC-7, the pf system utilized user-defined passages to formulate a feedback query, while the df system employed document-level relevance judgments to perform relevance feedback with the adaptive linear model. Our TREC-8 df system remained nearly identical to the one used in TREC-7, but the pf system interface was overhauled in TREC-8 to improve its usability.

Our TREC-8 experiment involved 24 searchers who performed the same six search tasks (three per system) in varying search and system-order as specified in the TREC guidelines. Unfortunately, the results of our TREC-8 experiments, described in more detail below, seemed to run contrary to past experimental results and our hypothesis that interactive query expansion by user-defined passage terms would have an advantage.

Initially, we were confounded by the overall experiment results, where the df system showed slightly higher performance statistics than the pf system. When we examined results by the system-order groups (i.e., df1, pf1, df2, pf2, where df stands for document feedback, pf stands for passage feedback, 1 signifies “first system used”, and 2 signifies “second system used”), however, we saw no significant performance difference in systems except for the first used pf system (pf1) that showed noticeably poor results. Analysis of these results as well as transaction logs and questionnaire data suggest to us that the pf system might have been more difficult to learn than the df system. It is possible that the difficulty of learning to use pf resulted in the poor performance of the first used pf system, which in turn degraded overall performance of this system. Though the overall results are once again inconclusive, we are encouraged by the competitive performance of the second use of pf and believe that pf may be an effective mechanism for interactive IR.

In the remainder of this paper, we describe the main components of IRIS, followed by a discussion of our TREC-8 interactive experiment and results. A concluding section considers implications and possibilities for further study.

2. IRIS: system description

IRIS is an interactive retrieval system designed to provide users with many opportunities to interact with the system throughout the search process. For example, users can supplement the initial query with two-word collocations suggested by the system, perform relevance feedback by selecting relevant documents or passages, or add and delete query terms. IRIS, first created in 1996 at the School of Information and Library Science at UNC-CH, has been under continuous development, evolving with each participation in TREC experiments. Below is a description of its key components as used in our TREC-8 experiments.

2.1. Text processing

IRIS processes documents by first removing punctuation, and then excluding 390 high-frequency terms listed in the WAIS default stopwords list as well as "IRIS stopwords,"[2] which were arrived at by examining the inverted index and identifying low frequency terms that appeared to have little value.

After punctuation and stopword removal, IRIS conflates each word by applying one of the four stemmers implemented in the IRIS Nice stemmer module.[3] The module consists of a simple plural remover (Frakes & Baeza-Yates, 1992), a Porter stemmer (Porter, 1980), a modified Krovetz inflectional stemmer (Krovetz, 1993),[4] and a Combo stemmer that uses the shortest whole word (i.e., word that appears in a dictionary) returned by the three stemmers. We used the Krovetz stemmer in TREC-7 for its conservative conflation tendencies, but we opted for the simple plural remover in TREC-8 to speed up indexing time. The simple stemmer was chosen over the Porter stemmer to minimize the overstemming effect, with the hope that understemming effect would be compensated for by the feedback query expansion process.

2.2. Phrase construction

The same phrase construction method employed in our TREC-7 interactive experiments was used to construct phrase indexes for TREC-8. Using an online dictionary and a clause recognition algorithm built into the Nice stemmer, we constructed a two-word noun–noun phrase index by first extracting adjacent word pairs of noun and proper noun combinations within a clause,[5] and then discarding the phrases occurring 20 or fewer times in the collection to reduce indexing time and conserve computing resources. The phrase occurrence frequency threshold of 20 was arrived at by selecting the number that produced the phrase index whose size was most comparable to that of the collocation index used in our TREC-6 experiments (Sumner et al., 1998).

To augment the proper nouns in the online dictionary, all capitalized words not occurring at the beginning of a sentence were considered as proper nouns. Hyphenated words were broken up and stemmed by the simple plural remover before noun–noun phrase construction. Hyphenated words in their raw form (i.e., as they appear in documents sans punctuation) were added to the index as well.

[2] IRIS stopwords are defined as all numeric words, words that start with a special character, words consisting of more than 25 non-special characters, and words with embedded special characters other than a period, apostrophe, hyphen, underline, or forward or backward slash.

[3] Nice stemmer was implemented by Kiduk Yang, Danqi Song, Woo-Seob Jeong and Rong Tang at SILS at UNC. For an interactive demonstration, please visit http://ils.unc.edu/iris/nstem.htm.

[4] The modified Krovetz inflectional stemmer implements a modified version of Krovetz's algorithm and restores the root form of plural ("-s," "-es," "-ies"), past tense ("-ed"), and present participle ("-ing") words, provided this root form is in our online dictionary.

[5] IRIS identifies a clause boundary by the presence of appropriate punctuation marks such as a comma, period, semicolon, question mark, or exclamation mark.

2.3. Ranking function and term weights

IRIS ranks retrieved documents in decreasing order of the inner product of document and query vectors

$$\mathbf{q}^{\mathrm{T}}\mathbf{d}_i = \sum_{k=1}^{t} q_k d_{ik}, \tag{1}$$

where q_k is the weight of term k in the query, d_{ik} the weight of term k in document i, and t is the number of terms in the index. We used SMART *Lnu* weights for document terms (Buckley, Singhal, Mitra, & Salton, 1996,Buckley, Singhal, & Mitra, 1997), and SMART *ltc* weights (Buckley, Salton, Allan, & Singhal, 1995) for query terms. *Lnu* weights attempt to match the probability of retrieval given a document length with the probability of relevance given that length (Singhal, Buckley, & Mitra, 1996). Our implementation of *Lnu* weights was the same as that of Buckley et al. (1996, 1997) except for the value of the *slope* in the formula, which is an adjustable parameter whose optimal value may depend, in part, on the properties of the document collection.

According to our pre-test experiments, an *Lnu* slope of 0.5 performed best with feedback, especially when using both single term and phrase indexes. Based on these findings, we used a slope of 0.5 to optimize performance with feedback.

2.4. Feedback models

2.4.1. The adaptive linear model

We used the same implementation of the adaptive linear model in TREC-8 as in TREC-7 (Wong & Yao, 1990; Wong, Yao, Salton, & Buckley, 1991). The basic approach of the adaptive linear model, which is based on the concept of preference relations from decision theory (Fishburn, 1970), is to find a *solution vector* that, given any two documents in the collection, will rank a more-preferred document before a less-preferred one (Wong, Yao, & Bollmann, 1988).

In the df interface of IRIS, users can evaluate documents as “relevant”, “marginally relevant” or “nonrelevant”. By adapting the concept of user preference relations to extend the relevance scale from binary to three-values, we constructed the following formula for the starting vector. Note that this formula can be adjusted for any multi-valued relevance scale

$$\mathbf{q}_{(0)} = c_0\mathbf{q}_{rk} + \frac{c_1}{N_{\text{new rel}}}\sum_{\text{new rel}}\mathbf{d} + \frac{c_2}{N_{\text{new mrel}}}\sum_{\text{new mrel}}\mathbf{d} - \frac{c_3}{N_{\text{new nonrel}}}\sum_{\text{new nonrel}}\mathbf{d}, \tag{2}$$

where $\mathbf{q}_{\text{rk}}$ is the query vector that produced the current ranking of documents, c_0, c_1, c_2 and c_3 constants, $N_{\text{new rel}}$, $N_{\text{new mrel}}$ and $N_{\text{new nonrel}}$ the number of new relevant, new marginally relevant, and new nonrelevant documents, respectively, in the current iteration and the summations are over the appropriate new documents. A detailed description of the adaptive linear model can be found in Sumner et al. (1998)[6].

[6] Modification and implementation of the adaptive linear model using a multi-level relevance scale was done by Robert Sumner, who has played an indispensable role in the development of IRIS as well as our TREC participation in the past.

2.4.2. The passage feedback model

Conventional relevance feedback models treat documents as the information units and assume user relevance judgments to be about entire documents. The document unit in real life, however, is sometimes determined by arbitrary boundaries, based on convenience or convention, rather than content. This can produce a document containing subsections of varying information content. In such instances, the user's determination of relevance is likely to be based on only certain portions of a document. Even in a document of consistent content, the user may be interested in only specific passages.

Findings from passage retrieval research, in which ranking based on passage-query similarity was compared to ranking based on document-query similarity, demonstrated the promise of using passages over documents in retrieval (Kaszkiel & Zobel, 1997; Salton, Allan, & Buckley, 1993; Wilkinson, 1994; Zobel, Moffat, Wilkinson, & Sacks-Davis, 1995). Although passage retrieval and pf are not altogether the same, they do share some common aspects that can be beneficial to retrieval performance. The smaller unit of a passage means that assessment of relevance is not only more precise but also more likely to be semantically oriented. For example, documents with chance occurrences of relevant terms scattered across passages are less likely to be regarded as relevant in a passage-based system. One of main challenges of passage retrieval lies in the determination of passages. How documents should be split into passages in order to maximize its advantageous potential is an important consideration.

However, the determination of passages is not an issue in interactive pf, where users can select many portions of documents they find relevant or not relevant. Furthermore, relevance in an interactive pf system is interactively determined from the contextual evaluation of documents; employing not only the full power of human intelligence but also an intimate knowledge of the information need. Thus, we believe pf employing user-defined passages can be an effective component of an interactive retrieval system that aims to optimize contributions from both the user and the system.

As a first step to building such an interactive system, we implemented a pf model in TREC-7 with the following formula for feedback vector creation:

$$\mathbf{q}_{\text{new}} = \mathbf{q}_{\text{old}} + \sum_{\text{rel}} \mathbf{p} - \sum_{\text{nonrel}} \mathbf{p}, \tag{3}$$

where $\mathbf{q}$ is the query vector and $\mathbf{p}$ is the passage vector determined by the user's selection of relevant and nonrelevant portions of documents. Since the normalization factor of the *Lnu* weight is based on document length, an inverse document frequency weight was used for the passage vector $\mathbf{p}$. The pf approach differs fundamentally from the philosophy of the adaptive linear model in that it simply expands the query vector to make it more "similar" to relevant passages and "dissimilar" to nonrelevant passages rather than trying to rank a document collection in the preference order defined by a training set. Note that the 3-valued relevance judgments of our adaptive linear model are not implemented for pf: passages are simply marked as "relevant" (resulting in terms being added with positive weights to the modified query) or "nonrelevant" (resulting in terms being added with negative weights in the modified query).

We briefly considered a more complex implementation of pf such as more sophisticated query expansion and term weighting or ranking and display of retrieval results by passages, but opted for the simplest implementation for system efficiency and experimental simplicity. Given the 15 min time limit per search task for the TREC-8 interactive track, we believed that system response time was an important factor that could influence user-system interaction. The decision for a simple implementation was also influenced by our past TREC experiences that taught us to test the system by one building block at a time in the interactive retrieval setting, where interaction among multiple factors can rapidly become complex.

Though the underlying implementations of the pf model are the same in our TREC-7 and TREC-8 interactive experiments, we made a significant modification to the user interface in TREC-8. One of the prevalent user comments of our TREC-7 pf system was about the usability of its feedback interface. Users had to first select passages by highlighting with the mouse, copying the highlighted portions, toggling to the pf window, and then pasting the copied selection into appropriate windows. Transaction logs as well as user comments seemed to indicate the difficulty of these steps required for pf that might have kept users from fully enjoying the benefits of the system. Consequently, we simplified the feedback interface by using an embedded Java applet, which consolidated the document display and feedback windows as well as simplifying the overall pf operation.

3. Interactive experiment

3.1. Research question

One of the key questions in interactive IR is how to optimize contributions from the user and the system to improve retrieval performance. The user, being the source of the information need and armed with the powers of human intelligence, can contribute a great deal to the retrieval outcome. The system, with knowledge of the indexing language by which the collection is represented as well as the advantage of computational speed to apply appropriate mathematical principles on the fly, also contributes to the information retrieval process.

As mentioned in the introduction, most IR research in the past has focused on either the system or the human component of the retrieval process. As a result, the question of how the system and its users should interact to optimize mutual contributions to the retrieval process remains largely unanswered. Salton and Buckley (1990) demonstrated that relevance feedback is an effective automatic query expansion mechanism, but their experiment implicitly assumed not only the static nature of the information need but also user's full and unchanging awareness of what is relevant and what is not relevant throughout the search process.

According to the views of interactive IR research that regard information seeking as a dynamic discovery process, such assumptions may not hold true very often. Such factors as searcher's understanding and expression of the information need, the criteria for relevance and the utility of a relevant document among numerous other things may affect how the searcher interacts with the system. It is easy to imagine a scenario where the document found relevant by the searcher in one feedback iteration is judged as non-relevant in another. How well would automatic query expansion based on such relevance judgments work?

Consider another situation where the searcher needs to change the direction of his search in some way as in the case in the instance retrieval task of TREC interactive experiments. Having found one instance of the information need, the searcher now needs to find a different instance from the one already found. Would the searcher be able to influence the automatic query expansion process enough to redirect the search process to suit his or her needs?

An alternative to automatic query expansion via relevance feedback is interactive query expansion, where the searcher manually selects the feedback terms from a system-suggested term list based on his interactive relevance judgments (Beaulieu & Gatford, 1998; Belkin et al., 1998). The potential advantage of such systems lies in the ability of users to actively intervene in the query reformulation process to better reflect their needs rather than leaving the task of feedback query expansion to the system entirely. However, such systems still rely on the system process of identifying potential query expansion terms, which can suffer from the same shortcomings as a fully automated system. On the other hand, if a searcher were to select portions of documents with which to expand the query vector as he or she evaluates search results, it may prove advantageous in situations where system contributions to the query expansion process are inadequate or inappropriate. The effectiveness of such a system would depend heavily on the searcher's ability to recognize "potent" portions of documents since the system's contribution to feedback query expansion would be minimized to ranking the document collection to be displayed for evaluation.

Many questions can be asked at this point. Can searchers recognize "potent" passages for feedback query expansion? If so, how? What are the significant roles that the system and the searcher can play in the feedback query expansion process and how do they affect retrieval outcomes? What user-system interaction combinations work best in what situations? Motivated by these questions, we started our inquiry in the TREC-8 interactive experiments by comparing automatic feedback query expansion performed by the system with "manual" feedback query expansion controlled by the searcher. By starting at the extremes and working our way along the spectrum of user-system interactions, we hoped eventually to arrive at an optimal user-system interaction model of feedback query expansion. What follows is the first step in our journey to that goal.

3.2. Methodology

In our TREC-7 interactive experiments, we examined the effects of the user interface on retrieval performance by comparing a system with a complex interface to one with a simpler interface. We also compared the effectiveness of a pf system, where user-defined passages were utilized to expand the feedback query, with a conventional df system that employed relevance judgments based on documents to perform relevance feedback based on the adaptive linear model.

In our TREC-7 experiments, pf system results were better than df results. However, the results of the simple interface system versus the complex pf interface system were unexpected in that the simple interface performed slightly worse than the complex interface. After further examination, which revealed more searcher intervention steps in the transaction logs of the complex system, we concluded that complex system allowed users more opportunities to intervene, thereby positively affecting retrieval performance.

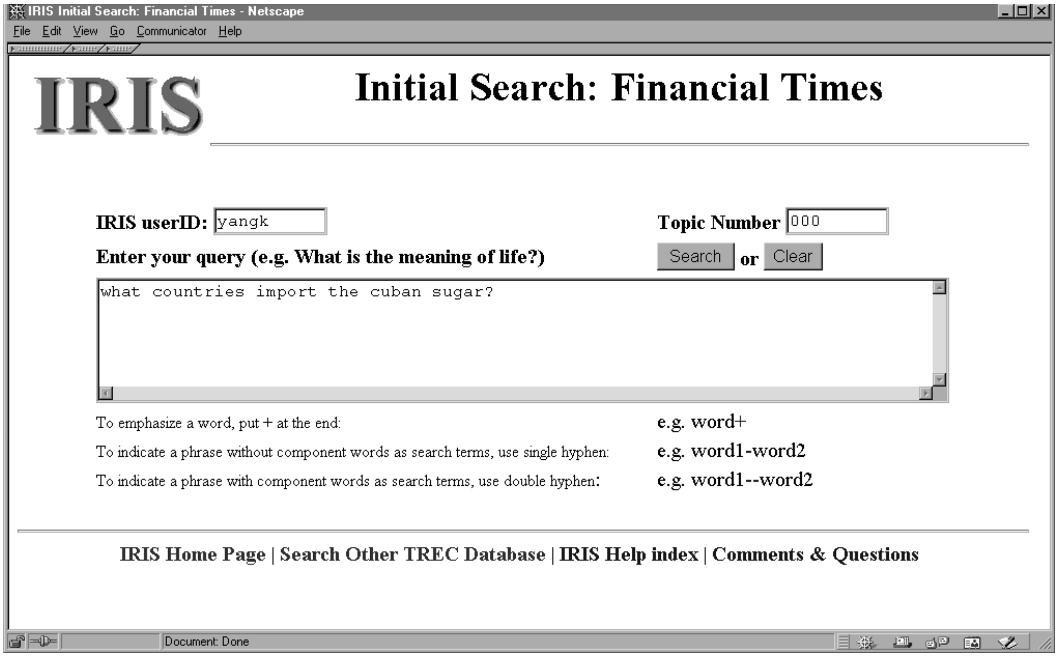
Furthermore, we noticed that pf features in TREC-7 were somewhat underutilized. Though there were more retrieval iterations per query in the pf system than in the df system, there were more reformulations of the initial query than expansion of the feedback vector with the pf interface. This supports our view that the value of a richer interactive framework may have partially masked the value of pf. One of the prevalent user comments of our TREC-7 pf system was about usability problems in its feedback interface. Indeed, our TREC-7 pf interface required several keystrokes or mouse clicks and toggling between two windows.

Based on these observations, we hypothesized the following in our TREC-8 experiment:

- pf in an interactive system will perform better than df;
- improving the usability of the pf interface will invite more utilization of it, thereby resulting in more positive user intervention;
- user intervention can positively affect retrieval performance.

To test our hypotheses, we constructed a pf system with a streamlined feedback interface for our TREC-8 interactive experiment and compared its performance with that of a df system. If our hypotheses were correct, our TREC-8 pf system should show better results and more user intervention steps than the df system.

The two systems in TREC-8 were identical in all aspects except for how relevance feedback was implemented. Both systems had exactly the same features and interfaces for initial query formulation (Fig. 1), initial query modification (Fig. 2), and feedback query modification (Fig. 3). The only difference between systems occurred during the relevance feedback process. The df system employed a conventional feedback mechanism for soliciting a 3-valued relevance judgment on the relevance of a whole document (Fig. 4(a)), then performing a new query based on the



Initial Search: Financial Times

IRIS userID: yangk

Topic Number 000

Enter your query (e.g. What is the meaning of life?)

what countries import the cuban sugar?

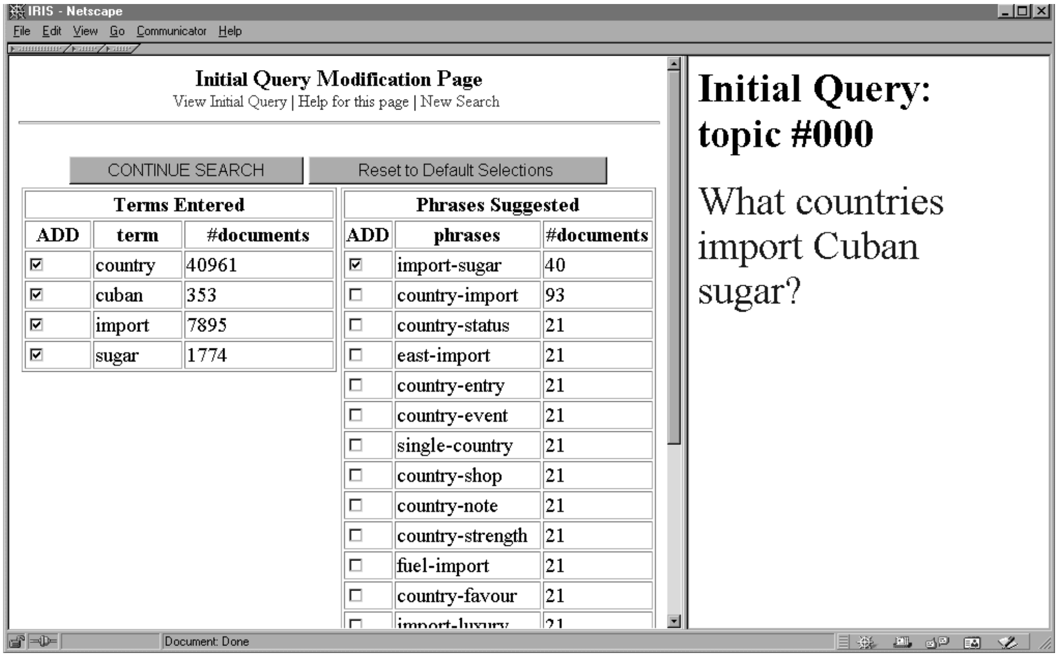
To emphasize a word, put + at the end: e.g. word+

To indicate a phrase without component words as search terms, use single hyphen: e.g. word1-word2

To indicate a phrase with component words as search terms, use double hyphen: e.g. word1--word2

IRIS Home Page | Search Other TREC Database | IRIS Help index | Comments & Questions

Fig. 1. Initial query formulation interface.


IRIS - Netscape
File Edit View Go Communicator Help
Document: Done

Initial Query Modification Page

View Initial Query | Help for this page | New Search

CONTINUE SEARCH
Reset to Default Selections

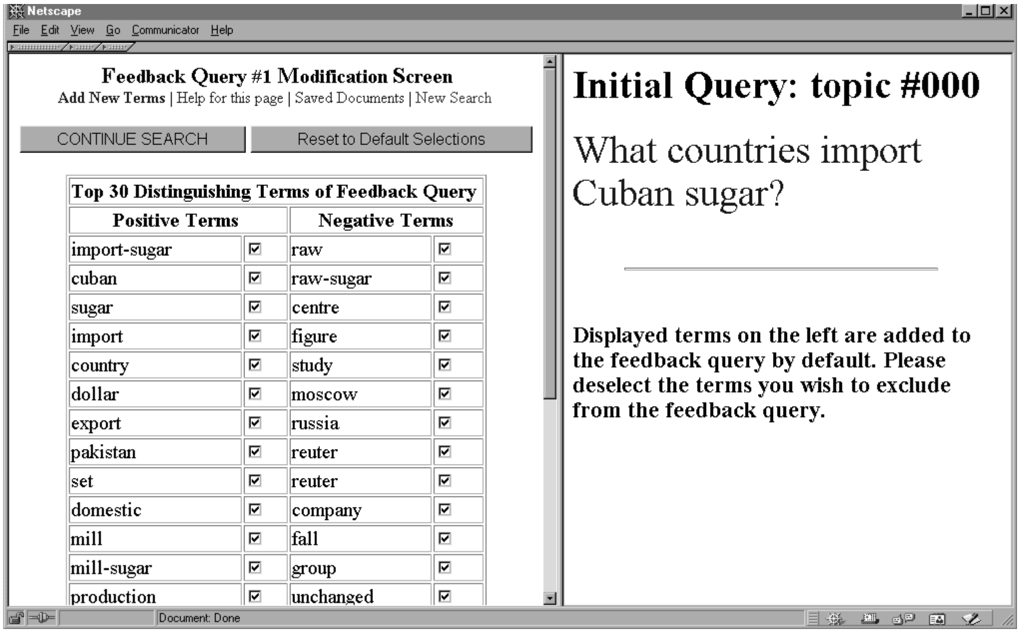
Terms Entered		
ADD	term	#documents
☑	country	40961
☑	cuban	353
☑	import	7895
☑	sugar	1774

Phrases Suggested		
ADD	phrases	#documents
☑	import-sugar	40
□	country-import	93
□	country-status	21
□	east-import	21
□	country-entry	21
□	country-event	21
□	single-country	21
□	country-shop	21
□	country-note	21
□	country-strength	21
□	fuel-import	21
□	country-favour	21
□	import-luxury	21

Initial Query: topic #000

What countries import Cuban sugar?

Fig. 2. Initial query modification interface.


Netscape
File Edit View Go Communicator Help
Document: Done

Feedback Query #1 Modification Screen

Add New Terms | Help for this page | Saved Documents | New Search

CONTINUE SEARCH
Reset to Default Selections

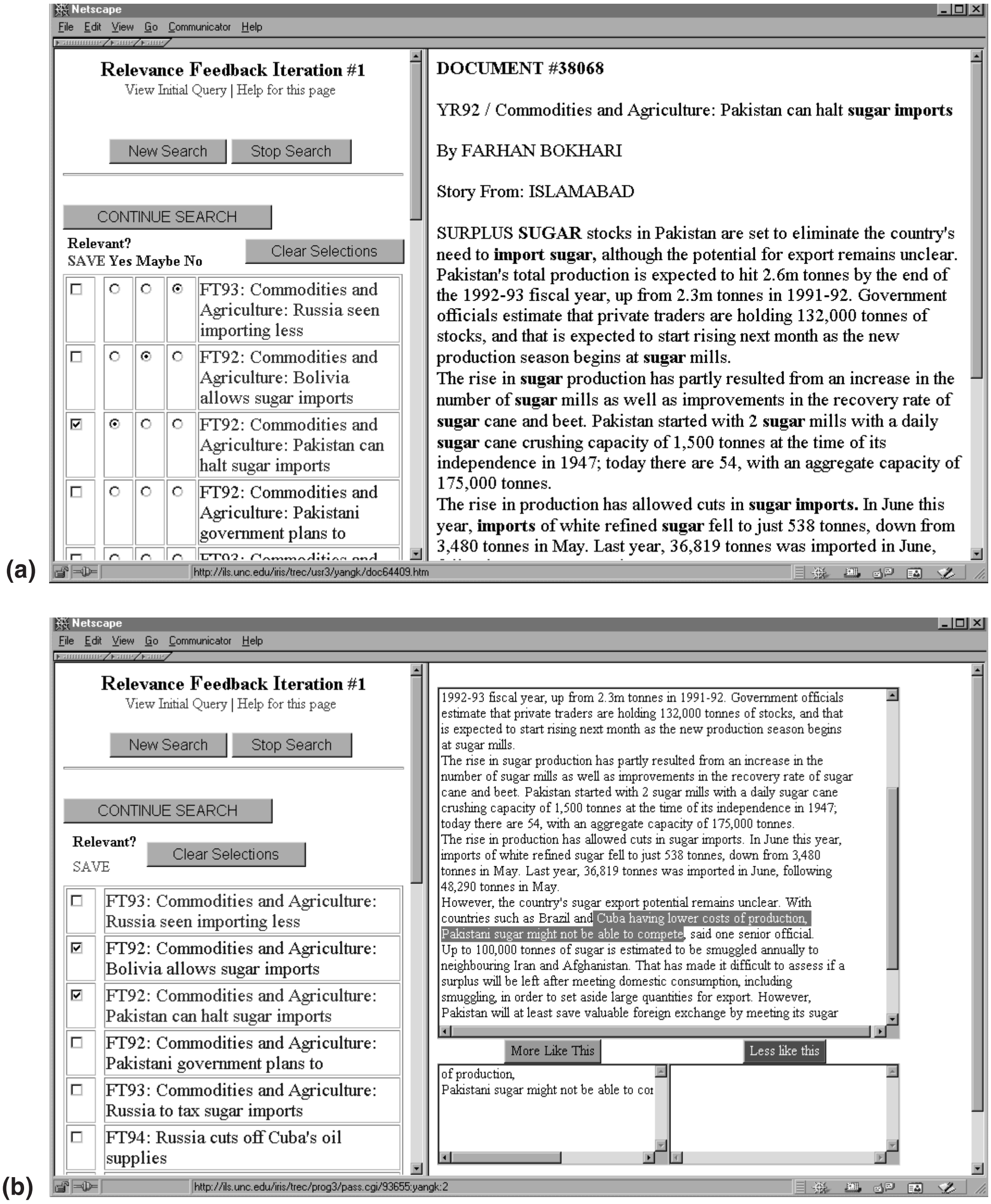
Top 30 Distinguishing Terms of Feedback Query			
Positive Terms		Negative Terms	
import-sugar	☑	raw	☑
cuban	☑	raw-sugar	☑
sugar	☑	centre	☑
import	☑	figure	☑
country	☑	study	☑
dollar	☑	moscow	☑
export	☑	russia	☑
pakistan	☑	reuter	☑
set	☑	reuter	☑
domestic	☑	company	☑
mill	☑	fall	☑
mill-sugar	☑	group	☑
production	☑	unchanged	☑

Initial Query: topic #000

What countries import Cuban sugar?

Displayed terms on the left are added to the feedback query by default. Please deselect the terms you wish to exclude from the feedback query.

Fig. 3. Feedback query modification interface.


(a)
Netscape
File Edit View Go Communicator Help
New Search
Stop Search
CONTINUE SEARCH
Clear Selections
http://ils.unc.edu/iris/trec/usr3/yangk/doc64409.htm
(b)
Netscape
File Edit View Go Communicator Help
New Search
Stop Search
CONTINUE SEARCH
Clear Selections
More Like This
Less like this
http://ils.unc.edu/iris/trec/prog3/pass.cgi/93655:yangk:2

Relevance Feedback Iteration #1
View Initial Query | Help for this page

Relevant?

SAVE	Yes	Maybe	No	
□	○	○	●	FT93: Commodities and Agriculture: Russia seen importing less
□	○	●	○	FT92: Commodities and Agriculture: Bolivia allows sugar imports
☑	●	○	○	FT92: Commodities and Agriculture: Pakistan can halt sugar imports
□	○	○	○	FT92: Commodities and Agriculture: Pakistani government plans to

DOCUMENT #38068

YR92 / Commodities and Agriculture: Pakistan can halt **sugar imports**

By FARHAN BOKHARI

Story From: ISLAMABAD

SURPLUS **SUGAR** stocks in Pakistan are set to eliminate the country's need to **import sugar,** although the potential for export remains unclear. Pakistan's total production is expected to hit 2.6m tonnes by the end of the 1992-93 fiscal year, up from 2.3m tonnes in 1991-92. Government officials estimate that private traders are holding 132,000 tonnes of stocks, and that is expected to start rising next month as the new production season begins at **sugar** mills.
The rise in **sugar** production has partly resulted from an increase in the number of **sugar** mills as well as improvements in the recovery rate of **sugar** cane and beet. Pakistan started with 2 **sugar** mills with a daily **sugar** cane crushing capacity of 1,500 tonnes at the time of its independence in 1947; today there are 54, with an aggregate capacity of 175,000 tonnes.
The rise in production has allowed cuts in **sugar imports.** In June this year, **imports** of white refined **sugar** fell to just 538 tonnes, down from 3,480 tonnes in May. Last year, 36,819 tonnes was imported in June,

Relevance Feedback Iteration #1
View Initial Query | Help for this page

Relevant?
SAVE

□	FT93: Commodities and Agriculture: Russia seen importing less
☑	FT92: Commodities and Agriculture: Bolivia allows sugar imports
☑	FT92: Commodities and Agriculture: Pakistan can halt sugar imports
□	FT92: Commodities and Agriculture: Pakistani government plans to
□	FT93: Commodities and Agriculture: Russia to tax sugar imports
□	FT94: Russia cuts off Cuba's oil supplies

1992-93 fiscal year, up from 2.3m tonnes in 1991-92. Government officials estimate that private traders are holding 132,000 tonnes of stocks, and that is expected to start rising next month as the new production season begins at sugar mills.
The rise in sugar production has partly resulted from an increase in the number of sugar mills as well as improvements in the recovery rate of sugar cane and beet. Pakistan started with 2 sugar mills with a daily sugar cane crushing capacity of 1,500 tonnes at the time of its independence in 1947; today there are 54, with an aggregate capacity of 175,000 tonnes.
The rise in production has allowed cuts in sugar imports. In June this year, imports of white refined sugar fell to just 538 tonnes, down from 3,480 tonnes in May. Last year, 36,819 tonnes was imported in June, following 48,290 tonnes in May.
However, the country's sugar export potential remains unclear. With countries such as Brazil and Cuba having lower costs of production, Pakistani sugar might not be able to compete, said one senior official.
Up to 100,000 tonnes of sugar is estimated to be smuggled annually to neighbouring Iran and Afghanistan. That has made it difficult to assess if a surplus will be left after meeting domestic consumption, including smuggling, in order to set aside large quantities for export. However, Pakistan will at least save valuable foreign exchange by meeting its sugar

of production,
Pakistani sugar might not be able to cor

Fig. 4. (a) Document feedback interface; (b) Passage feedback interface.

adaptive linear model. The pf system allowed users to select relevant and nonrelevant portions of a document with which to modify the feedback query vector (Fig. 4(b)). Additionally, an unintended system difference was introduced in the feedback interface due to a design oversight:

query terms contained in documents were not displayed in bold characters in the pf system as they were in the df system.

In both systems, user intervention could occur at several points throughout the search process: the initial query modification phase where the user may supplement their initial query with “suggested phrases” selected by the system; the feedback query modification phase where the user may add new terms or delete existing terms from the feedback query; and the relevance feedback phase where the user evaluates retrieved documents for relevance.

The underlying system constructs for both systems were same except for the feedback mechanism. The df system implemented the adaptive linear model to construct a feedback query with 250 terms with highest positive weights and 50 terms with lowest negative weights, whereas the pf system constructed the feedback query by simply adding terms in relevant passages and subtracting terms in nonrelevant passages selected by users. (Details of how the 3-valued judgments are processed are in Section 2.4.1.)

3.3. Searchers

Our TREC-8 interactive experiment involved 24 searchers, each of whom performed, in one session, three search tasks using one system and then another three search tasks using the other system. To minimize the potential effects of searcher and search topic differences, the order of both search topics and systems used were varied across searchers as specified in the TREC-8 guidelines. Table 1 shows information about each searcher’s background and search experience gathered by pre-study questionnaires. All searchers were either working on or had received a graduate degree, most (17) of which were in the field of Information and Library Science. The searchers had between 1 and 14 yr searching experience, (mean = 4.5). Nine of the 24 searchers were male.

Table 1
Response frequency of searchers on pre-study questionnaire

	No experience		Some experience		Great deal of experience
1. Using a point-and-click interface			3	5	18
2. Searching elec. library catalogs		2	3	9	10
3. Searching on CD ROM systems	1	3	12	8	
4. Searching commercial systems	5	11	2	5	1
5. Using WWW search services			5	8	11
6. Searching other systems	4		2	1	2
	Never	Once or twice a year	Once or twice a month	Once or twice a week	Once or twice a day
7. Searching frequency			2	9	13
	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
8. Enjoys information searches		1	6	10	7

3.4. Results[7]

We were confounded by the initial analysis of our TREC-8 experiment results, which seemed contrary to both our hypothesis and findings from previous experiments. The performance of systems measured by the mean instance precision[8] (MIP) and mean instance recall[9] (MIR) was slightly better for the df system than the pf system. The difference between systems was statistically significant at $\alpha = 0.10$ but not significant in $\alpha = 0.05$ (Table 2).

When we examined the results with system-order consideration, we noticed an interesting trend. The performance statistics of system-order groups in Table 3 show noticeably poor results by pf1 and relatively comparable results by all other system-order groups. As can be seen in Table 4, there is no significant performance difference between system-order groups unless the comparisons involve the first used pf system (pf1). When the system comparison is restricted to the first used systems only, the df system (df1) clearly outperforms the pf1 with *p*-value of 0.05. However, this is not true when they were the second systems searched (df2 vs pf2). Also, the second pf system (pf2) showed much better performance statistics than the pf1 with significant MIP difference ($\alpha = 0.10$). The MIR difference between pf1 and pf2 is not statistically significant, possibly due to large standard deviation values.

Two-way ANOVA with factors *system* and *order*[10] shows the MIR difference between passage and df system to be significant at 0.10 level but not at 0.05 level, with no system $*$ order interaction or order effect for MIR (Table 5). There is some system $*$ order interaction for MIP, though MIP difference between systems is not significant (Table 5). These ANOVA tables, when considered in the context of system-order performance comparisons of Tables 3 and 4, seem to indicate that both systems performed comparably overall despite the markedly poor performance by the first used pf system.

In addition to retrieval performance measures (MIR and MIP), searchers were asked to provide evaluation data on each search, each system and the overall searching process. Although none of the mean scores from the system questionnaire were different statistically, there does seem to be a difference in system rating depending on system-order (Table 6). The ratings remained fairly consistent between systems when the searcher used the document system first but the document system was given a higher score when the searcher used the passage system first.

The searchers also showed a difference in system preference, depending on system-order, when explicitly comparing the two systems in the exit questionnaire (Table 7). The searchers who used the document system first (df1) were equally divided on the system they felt was easier to learn, but 83% of the searchers who used the passage system first (pf1) indicated that the document system was easier to learn (the difference was significant at $\alpha = 0.10$ but not at $\alpha = 0.05$). When asked about the system that was easier to use and the system that they liked the best, again the searchers who used the document system first were divided fairly equally (50%, 58%, respectively),

[7] Performance statistics presented in this paper differ from those in our TREC-8 paper (Yang & Maglaughlin, 2000) due to errors in our submitted results to NIST. The performance statistics presented in this paper were computed from the corrected results.

[8] Instance precision is the fraction of documents selected by the searcher that contain one or more instances.

[9] Instance recall is the fraction of total instances contained in documents selected by the searcher.

[10] System is defined by the feedback type (i.e., pf system, df system) and *order* signifies the order of systems used.

Table 2
Overall performance statistics of document feedback and passage feedback systems

	Document feedback (df)	Passage feedback (pf)	$H_a : \mu_{df} \neq \mu_{pf}$
Mean instance precision (MIP)	0.676 (0.224[a])	0.629 (0.261[a])	0.10[b]
Mean instance recall (MIR)	0.293 (0.239[a])	0.226 (0.198[a])	0.09[b]

[a] Standard deviation.
[b] *p*-values from matched-pair *t*-test.

Table 3
Performance statistics of system-order groups

	1st used system (f1)	2nd used system (f2)	1st used df (df1)	1st used pf (pf1)	2nd used df (df2)	2nd used pf (pf2)
MIP (STD[a])	0.635 (0.253)	0.671 (0.233)	0.695 (0.215)	**0.575 (0.277)**	0.657 (0.234)	0.684 (0.235)
MIR (STD[a])	0.244 (0.206)	0.275 (0.235)	0.292 (0.227)	**0.196 (0.174)**	0.293 (0.254)	0.256 (0.217)

[a] Standard deviation.

Table 4
Statistical significance (*p*-values) of system differences

	$H_a : \mu_{f1} \neq \mu_{f2}$	$H_a : \mu_{df1} \neq \mu_{pf1}$	$H_a : \mu_{df2} \neq \mu_{pf2}$	$H_a : \mu_{df1} \neq \mu_{df2}$	$H_a : \mu_{pf1} \neq \mu_{pf2}$	$H_a : \mu_{df1} \neq \mu_{pf2}$	$H_a : \mu_{pf1} \neq \mu_{df2}$
MIP	0.38[a]	**0.05**[a]	0.63[a]	0.47[a]	**0.08**[a]	0.78[b]	**0.05**[b]
MIR	0.41[a]	**0.05**[a]	0.51[a]	0.99[a]	0.21[a]	0.53[b]	**0.08**[b]

[a] *p*-values from two sample *t*-test.
[b] *p*-values from matched-pair *t*-test.

Table 5
Two-way ANOVA table

Source	Degrees of freedom	Sum of squares	Mean square	*F* value	*p*-value
For Mean Instance Precision					
System	1	0.0798	0.0798	1.37	0.24
Order	1	0.0459	0.0459	0.79	0.38
System * Order	1	0.1957	0.1957	3.36	0.07
For Mean Instance Recall					
System	1	0.1592	0.1592	3.29	0.07
Order	1	0.0331	0.0331	0.68	0.41
System * Order	1	0.0314	0.0314	0.65	0.42

while the searchers who used the passage system first tended to prefer the document system (75%, 75%, respectively).

Examination of transaction logs and questionnaire data suggest that the searchers might have found the pf system difficult to learn. Using the pf system in the first phase of the experiment

Table 6
Mean scores and standard deviation of system questionnaire (1 = not at all, 5 = extremely)

	Document feedback (df)	Passage feedback (pf)	1st used df (df1)	2nd used pf (pf2)	1st used pf (pf1)	2nd used df (df2)
How easy the searcher felt the system was *to learn*	3.57 (0.79)	3.43 (0.66)	3.08 (0.51)	3.09 (0.70)	3.14 (0.67)	4.09 (0.70)
How easy the searcher felt the system was *to use*	3.78 (0.80)	3.26 (0.92)	3.42 (0.79)	3.18 (0.87)	3.33 (0.98)	4.18 (0.60)
How well the searcher *understood how to use* the system	3.70 (0.76)	3.35 (1.03)	3.42 (0.68)	3.45 (0.70)	3.58 (1.08)	4.00 (0.77)

Table 7
System preference[a] in the exit questionnaire

	Total df	1st used df (df1)	1st used pf (pf1)	*t*-statistic
Easier to learn	16 (67%)	6 (50%)	10 (83%)	1.773, $p < 0.1$
Easier to use	15 (63%)	6 (50%)	9 (75%)	1.254, ns[b]
Liked the best	16 (67%)	7 (58%)	9 (75%)	0.842, ns[b]

[a] Each cell contains number and percentage of searchers choosing the document feedback system over the passage feedback system.

[b] Statistically not significant ($\alpha = 0.1$).

might have imposed additional cognitive burdens that searchers were ill equipped to deal with initially. We believe that the searcher is better able to utilize the pf system once familiar with general system use as well as the interactive search tasks.

Table 8 shows that searchers typically spent more time evaluating documents with the pf system than the df system. Searchers evaluated documents more quickly when using the df system second (df2), but actually took more time when they used the pf system second (pf2). This may be due to our system design oversight that did not boldface the query terms contained in documents retrieved by the pf system. It is also possible that the task of identifying relevant portions of documents is just a more specific subset of making relevance judgments on whole documents (i.e., the searcher is doing the same thing, but at a higher level of granularity). Thus, searchers who used

Table 8
Relevance feedback statistics from transaction logs

	df	pf	df1	pf1	df2	pf2
Mean time of documents evaluation (s)	206	244	236	240	175	247
Mean number of positively weighted terms	224	19	235	14	212	23
Mean number of negatively weighted terms	22	4	22	1	22	7

the pf system first automatically gain familiarity with the df system, which helps to speed up things for the second df system, but when searchers encounter the pf system after first using the df system, more time is required to learn the higher granularity of passage evaluation.

Comparison of the feedback query length between systems show a decrease in the second df system and an increase in the second pf system. Since the df system has a fixed feedback query length by default, the only possible explanation of the reduced query length is by searcher intervention of term exclusion in the feedback query modification interface (Fig. 3). It is conceivable that searchers would intervene more by excluding inappropriate terms as they became more familiar with the system and the search task in the second df system. The terms displayed in the feedback query modification interface of the pf systems are what the searchers selected as relevant passages in the first place, so taking them out would be unlikely. On the contrary, the familiarity with the system use, search task, and pf is likely to engender more utilization of the pf feature, therefore, resulting in the increase of the feedback query length in the second pf system. This is evidenced by the increase in pf2 term counts in Table 8.

If we ignore the outlier (D at 10), the scatter plots of overall search iterations per topic (Fig. 5) tells us that searchers did more searches per topic in the pf system than the df system. In other words, they did more feedback iterations in the df system overall. The scatter plot of first systems (Fig. 6) resembles the overall plot, but the second system plots (Fig. 7) show similar number of search iterations between the document and pf systems. There were also fewer searches per topic (i.e., more feedback) in the second systems in general than the first systems. The patterns in these plots tell us that searchers tended to do new searches rather than utilizing the feedback feature in the first pf system, but utilized the feedback feature more in the second pf system as they became more familiar with the system. The scatter plots reinforce our belief about the learnability of the pf system, namely, that the pf system might have been a more difficult system to learn than the df system.

```
      |
      |  D
   40 +
      |
      |  P
   35 +
      |
      |
   30 +
F     |
r     |
e  25 +
q     |
u     |
e  20 +
n     |
c     |
y  15 +
      |       P
      |       D    D
   10 +
      |            P
      |
    5 +                      P
      |                 P
      |                      D
    1 +                 D         P    P    PD        D
      |
      ---+----+----+----+----+----+----+----+----+----+--
         1    2    3    4    5    6    7    8    9   10

                 Number of Searches per Topic
```

Fig. 5. P = Passage Feedback, D = Document Feedback.

```
          |
          |
       25 +
          |
          |
     F    |
     r    |
     e 20 +
     q    |  D
     u    |  P
     e    |
     n    |       PD   D
     c  5 +            P    P
     y    |                 P
          |
          |
          |
        1 +                 D    D         PD   P         D
          |
          ---+----+----+----+----+----+----+----+----+----+--
             1    2    3    4    5    6    7    8    9   10

                      Number of Searches per Topic
```

Fig. 6. First systems searched: P = Passage Feedback, D = Document Feedback.

```
          |
          |
       25 +
          |  PD
          |
     F    |
     r    |
     e 20 +
     q    |
     u    |
     e    |
     n    |       P
     c  5 +       D    D
     y    |            P
          |
          |
          |                      D
        1 +                           P
          |
          ---+----+----+----+----+----+----+----+----+----+--
             1    2    3    4    5    6    7    8    9   10

                      Number of Searches per Topic
```

Fig. 7. Second systems searched: P = Passage Feedback, D = Document Feedback.

Evidence from performance statistics as well as transaction logs and searcher questionnaire data suggest that the pf system was more difficult to learn than the df system in our TREC-8 experiment. Consequently, we believe the poor result of the pf system in TREC-8 was largely due to our failure to make the pf system more usable. Though we improved the passage selection interface from TREC-7, we did little to ease the cognitive burden of having to identify relevant passages. Thus, we believe that one of the main challenges for pf lies in helping searchers identify relevant passages quickly and easily.

4. Suggestions for further study

As the scope of retrieval is broadened to include the searcher and the dynamic cognitive process of information seeking in an interactive IR environment, the universe of possible IR “moves” grows considerably. Even with the TREC interactive experiment guidelines designed to minimize the effect of lurking variables, our TREC-8 interactive experiment did not provide clear results.

As mentioned in preceding sections, our implementation of pf in TREC-8 was not only simplistic by design but also disadvantaged by oversight (i.e., query terms not boldfaced in documents). It would be interesting to see what would happen if the system was "improved" to increase user-system contributions. In order to assist the searcher in identifying useful passages more quickly and easily, the design oversight of not boldfacing query terms in documents should first be corrected, after which some visual cues indicating passage importance based on passage retrieval can be investigated. The effect of more sophisticated query expansion and term weighting methods should also be examined to find the system construct that maximizes its contribution to the feedback query expansion process.

In future work, we will consider modifying the experimental design to better suit our inquiry. One of the difficulties with assessing whether the searchers were effective in selecting useful passages stemmed from an experimental design that did not control what documents searchers evaluated. For a given topic, the set of documents presented to searchers for evaluation could vary widely depending on the initial query, which differed across searchers. Since documents evaluated for a given topic between systems mostly differed, direct comparison of feedback query use between systems is not possible. To focus the examination on the feedback query term selection process, one may consider controlling the experiment so that the same sets of documents are presented to searchers for given topics.

Finally, qualitative research methods such as videotaping think aloud protocols, and content analysis of user-defined passages should be incorporated to supplement the quantitative methods. In our opinion, the domain of interactive IR stretches beyond the realm of quantitative methods and should be arrived at by incorporating multiple approaches.

References

Beaulieu, M. (1997). Experiments on interfaces to support query expansion. *Journal of Documentation*, *53*, 8–19.

Beaulieu, M. M., & Gatford, M. J. (1998). Interactive Okapi at TREC-6. In E. M. Voorhees, & D. K. Harman, *The sixth text retrieval conference* (*TREC-6*) (NIST Special Publication 500-240, pp. 143–167). Washington, DC: US Government Printing Office.

Beaulieu, M., Robertson, S., & Rasmussen, E. (1996). Evaluating interactive systems in TREC. *Journal of the American Society for Information Science*, *47*, 85–94.

Belkin, N. J. (1980). Anomalous states of Knowledge as a basis for information retrieval. *Canadian Journal of Information Science*, *5*, 133–143.

Belkin, N. J. (1987). Discourse analysis of human information interaction for specification of human-computer information interaction. *Canadian Journal of Information Science*, *12*, 31–42.

Belkin, N. J., Brooks, H. M., & Oddy, R. N. (1982). ASK for information retrieval. Part I: background and theory. *Journal of Documentation*, *38*, 61–71.

Belkin, N. J., Cool, C., Stein, A., & Thiel, U. (1994). Cases, scripts, and information-seeking strategies: on the design of information retrieval systems. *Information Processing and Management*, *29*, 325–344.

Belkin, N. J., Perez Carballo, J., Cool, C., Lin, S., Park, S. Y., Rieh, S. Y., Savage, P., Sikora, C., Xie, H., & Allan, J. (1998). Rutgers' TREC-6 interactive track experience. In E. M. Voorhees, & D. K. Harman, *The sixth text retrieval conference* (*TREC-6*) (NIST Special Publication 500-240, pp. 597–610). Washington, DC: US Government Printing Office.

Brooks, H. M., Belkin, N. J., & Daniels, P. J. (1986). Research on information interaction and intelligent information provision mechanisms. *Journal of Information Science: Principles & Practices*, *12*, 37–44.

Buckley, C., Salton, G., Allan, J., & Singhal, A. (1995). Automatic query expansion using SMART: TREC 3. In D. K. Harman, *Overview of the third text retrieval conference* (*TREC-3*) (NIST Special Publication 500-225, pp. 69–80). Washington, DC: US Government Printing Office.

Buckley, C., Singhal, A., & Mitra, M. (1997). Using query zoning and correlation within SMART: TREC 5. In E. M. Voorhees, & D. K. Harman, *The fifth text retrieval conference* (*TREC-5*) (NIST Special Publication 500-238, pp. 105–118). Washington, DC: US Government Printing Office.

Buckley, C., Singhal, A., Mitra, M., & Salton, G. (1996). New retrieval approaches using SMART: TREC 4. In D. K. Harman, *The fourth text retrieval conference* (*TREC-4*) (NIST Special Publication 500-236, pp. 25–48). Washington, DC: US Government Printing Office.

Clarkson, M. A. (1992). The information theater. *Byte*, *17*, 145–152.

Dervin, B. (1997). Given a context by any other name: methodological tools for taming the unruly beast. In P. Vakkari, R. Savolainen, & B. Dervin, *Information seeking in context* (pp. 13–38). London: Taylor.

Fishburn, P. C. (1970). *Utility theory for decision making*. New York: Wiley.

Frakes W. B., BaezaYates, R. (Eds.), *Information retrieval: data structures and algorithms*. Englewood Cliffs, NJ: Prentice-Hall.

Goker, A. (1997). Context learning in Okapi. *Journal of Documentation*, *53*, 80–83.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory. *Journal of Documentation*, *52*, 3–50.

Jones, S., Gatford, M., Do, T., & Walker, S. (1997). Transaction logging. *Journal of Documentation*, *53*, 35–50.

Kaszkiel, M., & Zobel, J. (1997). Passage retrieval revisited. *Proceedings of the ACM SIGIR conference on research and development in information retrieval*, 178–185.

Krovetz, R. (1993). Viewing morphology as an inference process. *Proceedings of the 16th annual international ACM SIGIR conference on research and development in information retrieval*, 191–203.

Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, *1*, 309–317.

Meadows, C. T., Hewett, T. T., & Aversa, E. S. (1982). A computer intermediary for interactive database searching. 1. Design. *Journal of the American Society for Information Science*, *33*, 325–332.

Porter, M. (1980). An algorithm for suffix stripping. *Program*, *14*, 130–137.

Robertson, S. E., Walker, S., & Beaulieu, M. (1997). Laboratory experiments with Okapi: participation in the TREC program. *Journal of Documentation*, *53*, 20–34.

Salton, G., Allan, J., & Buckley, C. (1993). Approaches to passage retrieval in full text information systems. *Proceedings of the ACM SIGIR conference on research and development in information retrieval*, 49–58.

Salton, G., & Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, *41*, 288–297.

Saracevic, T. (1996). Modeling interaction in information retrieval (IR): a review and proposal. In S. Hardin, *Proceedings of the 59th ASIS annual meeting* (pp. 3–9). Medford, NJ: Information Today, Inc.

Saracevic, T., Mokros, H., Su, L. T., Spink, A. (1991). Interaction between users and intermediaries in online searching. In M. E. Williams, *Proceedings of the 12th national online meeting* (pp. 329–341). Medford, NJ: Learned Information, Inc.

Shuman, B. A. (1989). Expert systems and the future of interactive searching. In *Proceedings, national online meeting*, (pp. 405–411).

Singhal, A., Buckley, C., & Mitra, M. (1996). Pivoted document length normalization. In *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval*, (pp. 21–29).

Spink, A., & Cool, C. (1992). Recognition of stages in the user's information-seeking process during online searching by novice searchers. *Online Review*, *16*, 297–301.

Sumner, R. G., Jr., Yang, K., Akers, R., & Shaw, W. M., Jr. (1998). Interactive retrieval using IRIS: TREC-6 experiments. In E. M. Voorhees, & D. K. Harman, *The sixth text retrieval conference* (*TREC-6*) (NIST Special Publication 500-240, pp. 711–734). Washington, DC: US Government Printing Office.

Taylor, R. S. (1968). Question-negotiation and information seeking in libraries. *College and Research Libraries*, *29*, 178–194.

Wilkinson, R. (1994). Effective retrieval of structured documents. In *Proceedings of the ACM SIGIR conference on research and development in information retrieval*, (pp. 311–317).

Wong, S. K. M., & Yao, Y. Y. (1990). Query formulation in linear retrieval models. *Journal of the American Society for Information Science*, *41*, 334–341.

Wong, S. K. M., Yao, Y. Y., & Bollmann, P. (1988). Linear structure in information retrieval. In *Proceedings of the ACM SIGIR conference on research and development in information retrieval*, (pp. 219–232).

Wong, S. K. M., Yao, Y. Y., Salton, G., & Buckley, C. (1991). Evaluation of an adaptive linear model. *Journal of the American Society for Information Science*, *42*, 723–730.

Yang, K., Maglaughlin, K., Meho, L., & Sumner, R.G., Jr. (1999). IRIS at TREC-7. In E. M. Voorhees, & D. K. Harman, *The seventh text retrieval conference* (*TREC-7*) (NIST Special Publication 500-242, pp. 555–566). Washington, DC: US Government Printing Office.

Yang, K. & Maglaughlin, K. (2000). IRIS at TREC-8. In E. M. Voorhees, & D. K. Harman, *The eighth text retrieval conference* (*TREC-8*). Washington, DC: US Government Printing Office (in press).

Zobel, J., Moffat, A., Wilkinson, R., & Sacks-Davis, R. (1995). Efficient retrieval of partial documents. *Information Processing and Management*, *31*(3), 361–377.