Elicitation of Term Relevance Feedback: An Investigation of Term Source and Context

Diane Kelly & Xin Fu
University of North Carolina
100 Manning Hall, CB#3360
Chapel Hill, NC 27599-3360 USA
+1 919.962.8065
[dianek | fu] @ email.unc.edu

ABSTRACT
Term relevance feedback has had a long history in information retrieval. However, research on interactive term relevance feedback has yielded mixed results. In this paper, we investigate several aspects related to the elicitation of term relevance feedback: the display of document surrogates, the technique for identifying or selecting terms, and sources of expansion terms. We conduct a between subjects experiment (n=61) of three term relevance feedback interfaces using the 2005 TREC HARD collection, and evaluate each interface with respect to query length and retrieval performance. Results demonstrate that queries created with each experimental interface significantly outperformed corresponding baseline queries, even though there were no differences in performance between interface conditions. Results also demonstrate that pseudo-relevance feedback runs outperformed both baseline and experimental runs as assessed by recall-oriented measures, but that user-generated terms improved precision.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval - relevance feedback, query formulation.

General Terms
Performance, Experimentation, Human Factors

Keywords
Query expansion, relevance feedback interfaces, query length, user feedback, term context, familiarity, elicitation of feedback

1. INTRODUCTION
Despite having a long history in information retrieval, there is no consistent picture regarding the use and effectiveness of explicit term relevance feedback. Studies of term relevance feedback in interactive retrieval settings have yielded conflicting results. Many studies have demonstrated that users are reluctant to provide explicit term relevance feedback [3, 5], while others have demonstrated that users are willing to provide term relevance feedback [2, 14]. Some studies have demonstrated that interactive term relevance feedback is unlikely to lead to improvements in retrieval performance [2, 3, 5, 11, 15, 24, 25]; other studies have demonstrated the opposite [8, 9, 10, 12, 16, 21]. Clearly, there are many factors unique to each testing situation that make it difficult to compare results across studies, including differences in the experimental design and setting, test subjects and tasks, design of term relevance feedback interfaces, and techniques used to suggest terms.

In typical interactive term relevance feedback scenarios, users mark documents that they find relevant, the system suggests potential query expansion terms from these documents to users, and users select which of these terms are added to their queries. Empirical, laboratory-based studies have led to the general finding that users of experimental interactive IR systems desire term relevance feedback features [c.f., 3, 5, 17]. However, much of the evidence from these studies indicates that relevance feedback features are not used, or if they are they are unlikely to result in retrieval improvements. This has been attributed to problems related to the design of relevance feedback interfaces [20], task complexity and the user's lack of additional cognitive resources [5], and the amount of extra time required to use such features. For example, users in a series of studies by Belkin, et al. [5] rarely used relevance feedback features and often expressed confusion over why some terms were suggested by the system. In a study of simulated interactive query expansion, Ruthven [20] demonstrated that users are less likely than systems to select effective terms for query expansion. While Ruthven demonstrated some potential benefit of term relevance feedback if the best terms were used in query expansion, he went on to note that users are unlikely to select these terms because of problems with current relevance feedback interfaces. In a Web-based study, Anick [2] found that users made use of a term suggestion feature to expand and refine their queries. However, this use did not result in improvements in retrieval performance, which indicates that terms users selected were not particularly good. Conversely, in another study of an operational retrieval system, Efthimiadis [8] found that users selected about one-third of terms suggested by the system and that, in general, these terms improved retrieval performance.

One problem with current relevance feedback interfaces is that terms are often presented in isolation, which might make it difficult for users to fully comprehend relationships between terms and their information needs. The display of terms most

SIGIR'06, August 6-11, 2006 Seattle, Washington, USA.

often consists of a single list of terms. However, without appropriate term context, it can be difficult for users to understand how terms are used, why terms are suggested, and how such terms might be used to improve retrieval. Previous research does not provide a clear idea about how term context will affect user behavior and retrieval. Joho, et al. [11] presented users with two types of displays for query expansion, list and menu hierarchy. Joho, et al. found no significant differences in retrieval performance across display types, although subjects selected about 4 more terms on average from the menu hierarchy. Subjects in this study further stated that they believed that the menu hierarchies gave them a better idea of the contents of retrieved documents. Yang, et al. [25] investigated a series of relevance feedback interfaces that allowed users to mark terms, phrases, and documents, and submit passages from documents as relevance feedback. Wu, et al. [24] explored a cluster-based interface for relevance feedback and found that while users preferred this relevance feedback display over a list display, there were no differences in retrieval performance. Finally, during the previous three TREC HARD Tracks [c.f., 1], participants have experimented with a variety of techniques and interface features for providing and eliciting term relevance feedback from users. In general, results have been mixed, although no participant has achieved exceptional performance with any technique or interface.

Will users select more terms or fewer terms if term context is provided? Does term context enable users to make better decisions about term selection? In other words, does term context enable users to be more discriminate when selecting terms? Consequently, will selected terms improve or worsen retrieval performance? One purpose of the current study is to investigate if an interface that provides term context helps users make better query expansion decisions. With respect to this issue we hypothesize the following:

H1: Users will select more terms when they are presented in context than when they are presented in isolation.

H2: Terms users select when using an interface that presents terms in context will result in better retrieval performance than terms selected from an interface that presents terms in isolation.

Much work has been conducted examining the effectiveness of various techniques for selecting terms to display for term relevance feedback since the potential benefit of interactive term relevance feedback is still related to the quality of terms suggested by the system [c.f., 9, 10, 16]. No matter how many terms a user chooses, if terms are all of poor quality to begin with, then they are unlikely to increase retrieval performance. In retrieval situations where users pose ambiguous queries, this problem is more acute since there is a large chance that documents retrieved in response to such queries will be irrelevant.

Instead of relying on the system to identify potentially useful terms, some researchers have explored how users can be exploited as sources of terms for query expansion [7, 8, 12, 15]. For instance, Larson [15] designed an interface with a large text box aligned beside the full text of retrieval documents. This interface allowed users to easily enter potential query expansion terms while they evaluated documents. Kelly, et al. [12] designed an interface to elicit terms for query expansion from users by probing users more fully about their information needs. This study demonstrated that users were able to articulate additional information about their information needs beyond what they articulated in their initial queries and that this information improved retrieval performance significantly. Belkin, et al. [6] found the lengths of users' queries varied with the size of the query box. These studies suggest that term relevance feedback interfaces should be designed to elicit suggestions from users via interface features that are more fluid than the standard query box.

There is further evidence from studies of reference interactions that users can contribute good expansion terms via relevance feedback [21]. For instance, in a study of term sources for query expansion during user-intermediary retrieval, Spink [21] found that the majority of terms (38%) came from user question statements, and that these terms on average retrieved about 82% of the relevant documents. The most effective sources of search terms for query expansion were terms from users' written questions statements. Based on these results, Spink suggests that IR interfaces should encourage users to use their own knowledge as a source of terms for query expansion.

In this experiment, we are further interested in investigating users' abilities to suggest terms to add to their queries given appropriate stimulation. Along with the empirical evidence cited above, there is also theoretical support for this interest. Belkin [4] suggests that only through interaction with texts can users come to understand and learn about their information needs. Specifically, Belkin states that "interaction with texts implies at least the possibility of an unpredictable, and therefore unspecifiable, change in the condition which led to the interaction in the first place (e.g., the information need)" (p.59). This work suggests that as users interact with retrieved documents their information needs are likely to change. With respect to term relevance feedback, this indicates that terms may no longer be useful since the query on which they are based may no longer be appropriate. This also suggests that as users interact with documents, they may identify, recognize, or realize potentially useful query expansion terms; Pennanen and Vakkari [18] have found empirical support for this.

Based on this previous research, we propose that interactions with text surrogates (what we consider as context in this study) can stimulate users' thinking about their information needs and that this stimulation can help users identify additional terms to add to their queries. We anticipate that text surrogates will provide users with ideas about terms for query expansion in both a direct fashion (i.e., terms contained within surrogates can be identified by users) and an indirect fashion (i.e., terms contained within surrogates can stimulate users to think of additional terms not contained within surrogates). We hypothesize the following:

H3: Users will identify more terms using an interface that presents sentences and elicits free-form text input than an interface that presents sentences with check boxes.

H4: Terms suggested by users via the interface that presents sentences and elicits free-form text input will result in better retrieval performance than those selected with the interface that presents sentences with check boxes.

2. METHOD

We created three term relevance feedback interfaces and designed a between-subjects experiment to investigate our hypotheses. Each subject was assigned to a single interface condition and asked to build queries for ten search topics. Subjects spent no longer than one hour completing the experiment. Although subjects completed the study in group-settings in a computer

laboratory, each subject worked independently and interface condition did not vary within session. In total, there were seven experimental sessions, each containing a varying number of subjects. Details of the experiment are presented below.

2.1 Interfaces

We created three term relevance feedback interfaces each demonstrating a different method of displaying document surrogates and eliciting query expansion terms. Screen shots of the interfaces are displayed in Figures 1-3 (screen shots of interfaces two and three are truncated). The first interface (Interface_1) displayed a list of twenty terms; users were asked to mark check-boxes next to terms they wanted to add to their queries. The second interface (Interface_2) displayed a list of the same twenty terms, plus sentences in which these terms appeared; users were asked to mark check-boxes next to terms they wanted to add to their queries. Terms were emphasized in bold within their corresponding sentence. The final interface (Interface_3) displayed sentences used in Interface_2 and a text box. Users were asked to enter terms they wanted to add to their queries. Users were further instructed that terms could be from sentences or their own terms.

The comparison between Interface_1 and Interface_2 allowed us to explore H1 and H2, while the comparison between Interface_1 and Interface_3 allowed us to test H3 and H4. These interfaces were piloted previously with five subjects [13].

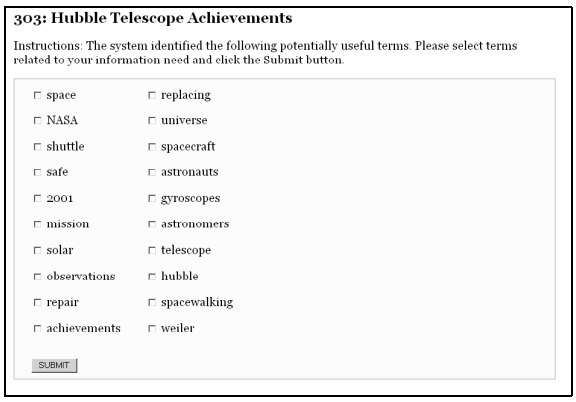

Figure 1. Interface_1: Terms and Check-boxes

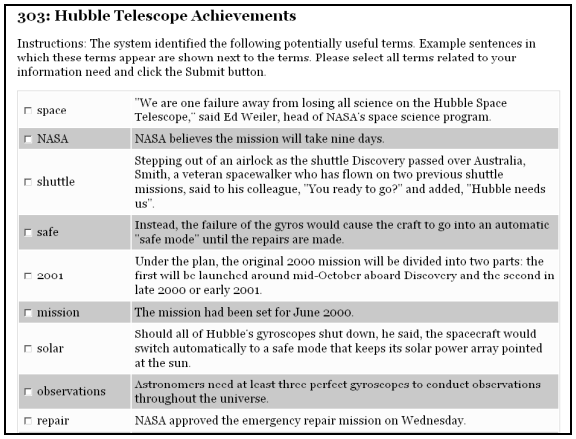

Figure 2. Interface_2: Terms + Context (sentences), and Check-boxes

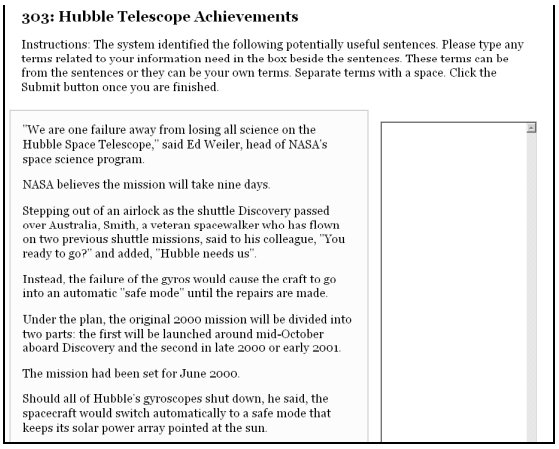

Figure 3. Interface_3: Sentences and Text Box

We used the Lemur IR toolkit (http://www.lemurproject.org) to conduct our experiments, with its basic defaults for indexing and Okapi BM25 for retrieval. Although we made use of a basic stop word and acronym list, we did not use a stemmer.

We used the information contained in the title field of the TREC HARD topics (collection described in more detail in Section 2.3) to populate our interfaces with terms and sentences. We built baseline queries with information contained in the title field for each topic (average query length = 2.50 terms) and conducted pseudo relevance feedback retrieval runs using each query. We set the pseudo relevance feedback parameter to use the top twenty ranking terms from the top ten ranking documents. We modified Lemur's basic retrieval feature (Reteval) so that for each query, terms used for pseudo relevance feedback were printed to a file, along with the document identification numbers from which these terms were extracted. The technique used for selecting pseudo relevance feedback terms is based on Robertson Selection Value (RSV) and described more fully in [19]; this technique is included as part of the Lemur toolkit. We used terms identified via this method to populate interfaces one and two and to identify sentences for interfaces two and three. We also used the retrieval results from this pseudo relevance feedback run as a baseline run in our evaluation.

To identify sentences, we constructed one word queries consisting of terms extracted during pseudo relevance feedback. For each topic, we collected all documents from which terms originated into a directory, parsed documents into sentences so that each sentence was in a unique file, indexed the files, and used the term queries and corresponding sentence files for retrieval. We used the top ranking sentences for each term query to populate interfaces two and three.

2.2 Subjects

Subjects were recruited and compensated in two different ways. Because of our between-subject design, we desired to have at least 15 subjects per condition. In the first recruitment and compensation approach, we sent out a solicitation email to the entire undergraduate population at our university. Subjects choose to participate in one of three study sessions. Interface condition was assigned randomly to each session. As compensation, subjects were offered drinks, snacks, and a small university gift (e.g., key chain). Subjects were also given the

opportunity to win one of three $30.00 USD gift certificates to the university bookstore.

This first attempt at recruiting subjects only resulted in 35 volunteers, so we developed another approach to recruitment and compensation. In the second approach, we made arrangements with a professor in our university's journalism and mass communication school to offer students in his undergraduate course extra credit as compensation for participation. This resulted in the recruitment of 26 volunteers. No volunteers recruited via the second approach had completed the experiment previously. Subjects recruited via the second approach were given the choice of four study sessions. Assignment of interface condition to session was a function of how many people signed up for each session and how many subjects we needed in each interface condition to create approximately equal groups.

In total, 61 subjects participated in this experiment. Twenty subjects used Interface_1, 21 subjects used Interface_2 and 20 subjects used Interface_3. Forty-six subjects were female and fifteen were male. Subjects had a mean age of 20 years. Subjects' mean search experience was 3.31 (where 1=very inexperienced and 4=very experienced) and subjects indicated that they searched the Web for information very frequently (mean=3.85, where 1=less than monthly and 4=daily).

2.3 Collection

We used the TREC HARD 2005 collection in this study [1]. This collection consists of the AQUAINT corpus, which contains 3 GB of newswire text data in English, drawn from three sources: the Xinhua News Service (1996-2000), the New York Times News Service (1998-2000), and the Associated Press Worldstream News Service (1998-2000).

The HARD 2005 collection consists of 50 standard TREC topics comprised of a title, a description and a narrative all taken from the TREC Robust Track. These topics have all been designated as 'difficult' by the Robust Track coordinators[1]. An example topic can be viewed in Figure 4. Topics were about a wide-range of subjects, including black bear attacks, cult lifestyles, Nobel prize winners, and mental illness drugs. In our experiment, topics were assigned systematically to subjects, such that they were rotated and counter-balanced across subject and across condition.

As part of TREC, binary relevance assessments of documents had been obtained for each topic using standard TREC methods [23]. Subjects in our study did not conduct any relevance assessments. Instead, we used these pre-existing relevance judgments to evaluate retrieval performance.

2.4 Search Topic Form

Subjects completed two major activities in the experiment: reviewing Search Topic Forms and building queries with one of the term relevance feedback interfaces. Subjects did not conduct any searching in this experiment. An example Search Topic Form is displayed in Figure 4. The Search Topic Form presented subjects with title, description and narrative fields from TREC topics, asked subjects to create initial queries for topics, and indicate their familiarity with search topics on a 7-point scale.

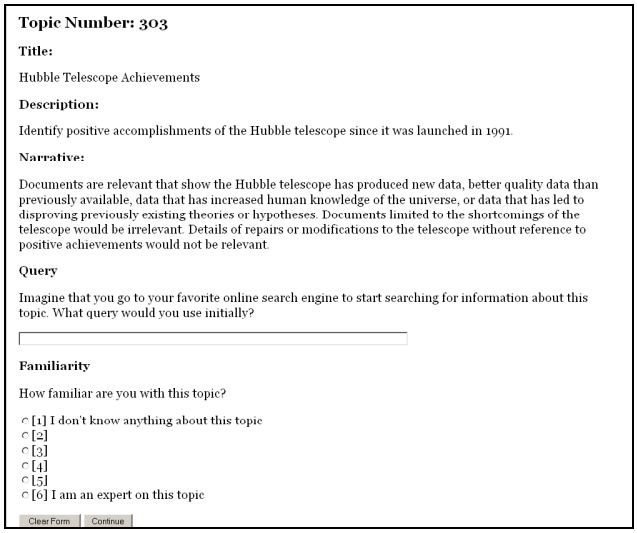

Figure 4. Search Topic Form

2.5 Protocol

At the start of the experiment, subjects read and agreed to a Consent Form and completed a short demographic questionnaire that asked them to indicate their ages, sexes, search experiences and frequencies with which they search the Web. Following this, subjects read a document describing general instructions for the experiment. Subjects were then presented with a Search Topic Form for their first search topics. After this, subjects were presented with a term relevance feedback interface and asked to select or identify terms related to the information need described in the previous Search Topic Form. The process of reviewing a topic and building a query continued for ten topics. When subjects submitted the form for the tenth topic, a thank you and debriefing note appeared on the screen.

2.6 Retrieval Runs

We conducted a series of runs based on data subjects provided during the experiment. We created baseline runs using subjects' initial queries from the Search Topic Forms, and experimental runs for each interface condition by adding terms identified by subjects with the experimental interfaces to their baseline queries. In total, there were three pairs of baseline and experimental runs, each corresponding to the three experimental conditions.

We included two pseudo relevance feedback runs as additional baselines. The rationale for this is that if performance results of experimental runs were comparable to runs using pseudo relevance feedback, then the value of our experimental interfaces, which require subjects to expend extra effort, would be questionable. The first pseudo relevance feedback run was described above in Section 3. It consisted of adding the 20 terms identified using pseudo relevance feedback to TREC baseline queries which were constructed using the title field. This run was equivalent to a user selecting all terms listed on Interface_1. The second pseudo relevance feedback run was constructed by using subjects' initial queries from the Search Topic Form. The pseudo relevance feedback parameters were identical to those used for the TREC-based pseudo relevance feedback run, where the top 20

[1] The Robust TREC Track focuses on ad hoc retrieval tasks, with an emphasis on individual topic performance rather than average topic performance. As a result, this Track identifies and studies topics that have performed poorly in past TRECs. These topics constitute the 'difficult' set.

ranking terms from the top 10 ranking documents were added automatically to subjects' initial queries. A summary of runs performed in this study is displayed in Table 1.

Table 1. Baseline and Experimental Runs

Source	Baselines	Pseudo-RF	Experimental
TREC	titles (1 run)	titles + 20 terms (1 run)	-
Subjects	initial queries from Search Topic Form (3 runs – one per interface condition)	initial queries + 20 terms (3 runs – one per interface condition)	initial queries + interface term (3 runs – one per interface condition)

2.7 Evaluation Measures

We used R-precision and precision-at-10, both standard TREC evaluation measures, to evaluate our results [23]. R-precision is precision at R documents retrieved, where R is the number of known relevant documents in the corpus. Thus, R-precision makes some use of recall in its computation. Precision-at-10 is precision calculated over the first 10 retrieved documents. We included precision at 10 as a measure since it most closely captures the situation that has been found to occur in real life retrieval settings, namely that the majority of users only view the first page of search results, which is usually about ten items [22].

3. RESULTS

In total, subject contributed 505 baseline queries and 610 experimental queries. As a reminder, subject baseline queries were derived from Search Topic Forms, while experimental queries consisted of subjects' baseline queries plus all terms identified via experimental interfaces. Ideally, we would have liked to have had 610 data points for both initial queries and experimental queries, since a total of 61 subjects completed this experiment. The discrepancy between baseline and experimental queries is a result of some subjects not entering initial queries (n=83). In addition, one subject misunderstood the question eliciting the initial query and indicated that she would "search Google" for all ten of her initial queries. Another subject entered "Google" as an initial query for two of ten topics she completed. For our analysis we dropped all cases where there were no initial queries (n=95). Thus, the dataset we used for analyses includes 505 cases. The distribution of cases according to experimental interface is as follows: term interface (n=182), term context interface (n=167), and sentence text-box interface (n=156).

3.1 Query Length (H1 and H3)

Figure 5 displays mean lengths of subjects' queries for each type of run/interface combination. Means and standard deviations for each pair (baseline, experimental) were Interface_1: 3.20 (1.13), 10.53 (4.19); Interface_2: 3.90 (1.77), 8.92 (3.77); Interface_3: 4.02 (1.84), 21.21 (7.76). Overall, queries created using the sentence and text-box interface (Interface_3) resulted in the longest queries, queries created using the term interface (Interface_1) resulted in the second longest and queries created using the term context interface (Interface_2) resulted in the shortest queries.

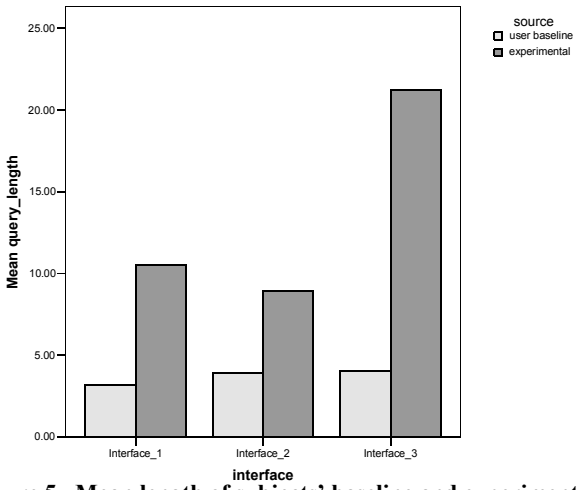

Figure 5. Mean length of subjects' baseline and experimental queries in each interface condition

Paired-sample t-tests were conducted to see if differences in mean query length between subjects' baseline and experimental queries were statistically significant. Results of these tests demonstrate statistically significant differences for all pairs across all conditions at the $p<.000$ level: Interface_1: $t(181) = -25.27$; Interface_2: $t(166) = -19.87$; and Interface_3: $t(155) = -27.88$. Thus, subjects created queries that were significantly longer than their baseline queries when using the experimental interfaces.

To test H1 and H3, we conducted independent sample t-tests between mean query length for Interface_1 and Interface_2 (H1) and Interface_2 and Interface_3 (H3). We found statistically significant results with each test, although in the first case it was not in the direction that we predicted. For H1, the t-test demonstrated that subjects entered significantly longer queries with Interface_1 than Interface_2, $t(347) = 3.75$, $p<.000$. Given that the mean query length for Interface_1 was higher than Interface_2, we did not expect to find support for H1 with the statistical test. It appears that term context might have helped subjects be more selective about which terms to add to their queries. The additional information provided by term context appeared to have provided evidence indicating which terms should be excluded rather than included in queries. It is important to note that in general, it takes longer to interact with Interface_2 than Interface_1, and this may have potentially impacted the results. However, subjects were not timed in this study.

For H3, the t-test demonstrated that subjects entered significantly longer queries with Interface_3 than Interface_2, $t(321) = -18.29$, $p<.000$. These results suggest that using a free form text-box to elicit term relevance feedback leads to longer queries. In this study, the text box and sentences provided subjects with an opportunity to identify useful terms from sentences, as well as an opportunity to generate more terms themselves. Furthermore, it seems to be the case that sentences stimulated subjects to identify more terms. Although we cannot be sure about this without conducting a study of this interface where one condition has sentences and another does not, these results seem to suggest that this occurred in this experiment.

From where do terms identified by subjects in the Interface_3 condition come? We analyzed all queries created using Interface_3 to understand the source of terms. We initially

mapped terms to one of two sources: sentences displayed on the interface and sentences not displayed on the interface (user generated). We further sub-divided terms that were from sentences displayed on the interface along the following dimensions: terms that the system would have suggested via automatic techniques (i.e., terms displayed on the term interface) and terms that would not have been suggested by the system via automatic techniques. Results are displayed in Figure 6. An average of 10.84 terms came from the interface and 6.23 terms came from users. Of the 10.84 terms from the interface, 6.97 were terms that the system would have suggested and 3.87 were terms that were contained within sentences, but that the system would not have suggested as part of term relevance feedback.

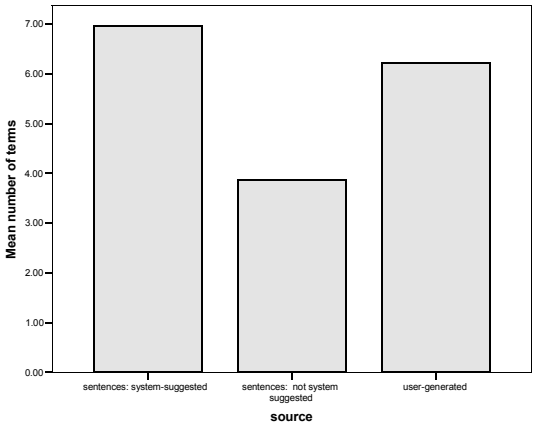

Figure 6. Source of terms from Interface_3

Interestingly, more terms came from the interface than were created by subjects. Since we have yet to examine these terms, we cannot characterize them. For instance, it might be the case that terms selected from the sentences were synonyms or variations of initial query terms. We leave a more thorough examination of these terms to future work.

3.2 Performance (H2 and H4)

Mean r-precision and precision-at-10 scores for baseline, experimental and pseudo RF runs for each interface condition and for TREC queries are displayed in Table 2. The best performing runs according to each measure are highlighted. Interestingly, baseline runs created from TREC topics (TREC Baseline) outperformed baseline runs created from subjects' initial queries.

Table 2. Mean (standard deviation) performance of baseline, experimental and pseudo relevance feedback runs

	R-precision		
	Baseline	Experimental	Pseudo RF
TREC	.2184 (.0226)	-	.2643 (.2042)
Interface_1	.1922 (.1682)	.2642 (.1736)	.2613 (.1736)
Interface_2	.1972 (.1681)	.2577 (.1838)	.2642 (.1930)
Interface_3	.1996 (.1735)	.2364 (.1862)	.2569 (.2032)
	Precision at 10		
	Baseline	Experimental	Pseudo RF
TREC	.3460 (.0415)	-	.4220 (.3610)
Interface_1	.3330 (.3079)	.4310 (.3376)	.4250 (.3524)
Interface_2	.3380 (.3007)	.4280 (.3361)	.4230 (.3571)
Interface_3	.3370 (.2958)	.4480 (.3551)	.4240 (.3624)

We conducted several analyses to compare these results. We conducted paired-sample t-tests of each pair of baseline and experimental runs for each interface to see if queries created in experimental conditions performed better than baseline queries. We found statistically significant differences in baseline query performance and experimental query performance for all measures across all three conditions at a $p<.000$ level, r-precision [Interface_1: $t(181) = -7.72$; Interface_2: $t(166) = -5.76$; Interface_3: $t(155) = -4.62$] and precision-at-10 [Interface_1: $t(181) = -4.33$; Interface_2: $t(166) = -4.42$; Interface_3: $t(155) = -4.62$. Even though there were large variations in the lengths of queries created in each interface condition, all experimental runs outperformed baseline runs.

It is clear from Table 2 that in most cases the pseudo relevance feedback runs outperformed both the baseline and experimental runs according to r-precision, although an experimental run and a pseudo relevance feedback run tied for best performing run. This suggests that in this situation the amount of effort subjects' spent creating queries was unnecessary, since gains from pseudo relevance feedback techniques were nearly equivalent, and sometimes slightly better, than gains from experimental runs. In contrast, the Interface_3 experimental run outperformed all runs, including the pseudo relevance feedback run according to precision-at-10. These results suggest that the additional terms provided by subjects with Interface_3 worked more to improve the precision of search results, and specifically, precision at the top of the retrieved document list, but not necessarily the total recall. It is unclear if terms generated by subjects, rather than terms contained in displayed sentences, were responsible for the improvements in precision-at-10. Future analyses will allow us to determine this.

We conducted paired-sample t-tests of each pair of baseline and pseudo relevance feedback runs for each interface condition, and found statistically significant differences in performance among all pairs of runs for both performance measures at the $p<.000$ level, r-precision [Interface_1: $t(181) = -7.72$; Interface_2: $t(166) = -5.76$; Interface_3: $t(155) = -4.62$] and precision-at-10 [Interface_1: $t(181) = -4.33$; Interface_2: $t(166) = -4.42$; Interface_3: $t(155) = -4.62$. Differences between experimental runs and pseudo relevance feedback runs were not statistically significant, although in some cases experimental runs outperformed the pseudo relevance feedback runs.

We conducted Mann-Whitney tests to explore differences in performance between the TREC pseudo relevance feedback run and each experimental run. Again, the TREC pseudo relevance feedback run is equivalent to subjects adding all suggested terms from Interface_1 to their queries. We conducted a Mann-Whitney test because the two groups were of very unequal size (50 queries in the TREC run and over 150 queries in each of the experimental runs). Results showed that differences between all pairs of runs were non-significant. Since the pseudo relevance feedback run required no additional effort on the part of the subject and the experimental runs did, the pseudo relevance feedback run definitely has the advantage in this situation.

H2 and H4 predicted that queries created with Interface_2 would perform significantly better than queries created with Interface_1, and that queries created with Interface_3 would perform significantly better than queries created with Interface_2. To test these hypotheses, we conducted a one-way ANOVA using interface condition as the independent variable and the two performance measures as dependent variables. We conducted two

ANOVAs, one with subjects' baseline query performance and one with subjects' experimental query performance. The purpose of the first ANOVA with subjects' baseline queries was to establish that the three groups were similar; this would allow us to eliminate the possibility that pre-existing differences in baseline query performance affected differences in experimental query performance. Results of the first ANOVA were non-significant, suggesting that the three experimental groups were homogenous with respect to the quality of their baseline queries. Results for the second ANOVA were also non-significant, demonstrating no performance differences across each of the experimental conditions. Thus, we found no support for H2 and H4.

Finally, previous work [12] has demonstrated a strong statistically significant relationship between query length and precision-based performance metrics, and we wanted to investigate this with our data. Since the mean query length of each experimental run differed, we conducted correlations between query length and performance for each experimental condition. Results of these correlations are displayed in Table 3. For Interfaces 1 and 3, there were no statistically significant correlations between query length and performance, but for Interface 2 there was. This is interesting because subjects' queries were much shorter in the Interface 2 condition, than in the other interface conditions. This data suggests that perhaps there is a positive correlation between query length and performance up to certain query length threshold, but after that gains for additional query terms are negligible. This is consistent with the performance results, where we see similar results across condition despite differences in mean query length. The difference in results between this study and those reported in [12] might be because the source of terms differed in these studies. In [12], all query terms were user-generated, while only a small subset of query terms in this experiment was user-generated.

Table 3. Correlations between query length and performance for each condition (*correlation significant at the 0.01 level)

		Performance Measure	
		R-precision	Precision-at-10
Interface	1	.101	.103
	2	.325*	.332*
	3	.153	.092

3.3 Familiarity

Finally, we looked at subjects' familiarity with topics to see if there were differences in familiarity across experimental conditions. We wanted to eliminate the possibility that familiarity was a confounding variable with respect to the results presented above. Subjects indicated that they were very unfamiliar with most topics. Subjects' mean familiarity with topics in each condition were as follows: Interface_1: 1.90 (.99), Interface_2: 1.81 (1.05), and Interface_3: 1.83 (1.02). We performed an ANOVA with the familiarity data to determine if mean familiarity was similar across the three experimental conditions. Results demonstrated no statistically significant differences in these means according to condition.

4. CONCLUSIONS

In this study, we found significant differences in subjects' mean query length according to term relevance feedback interface. Furthermore, we found that queries created while using these interfaces led to statistically significant improvements in retrieval performance over baseline queries according to r-precision and precision-at-10. However, we found no significant differences in retrieval performance across interfaces. We also found that pseudo relevance feedback runs performed just as well as experimental runs according to r-precision, but not as well according to precision-at-10. With respect to our hypotheses, we found support for Hypothesis 3, but no support for any of our other hypotheses. For Hypotheses 1, we found significant results in the direction opposite of what we predicted.

Instead of finding support for H1, we found that subjects identified significantly more terms with the term interface than the term context interface. At the start of the study, we were unclear about the direction of this hypothesis. Ultimately, the measure of term goodness is how effective it is with respect to retrieving relevant documents. It is likely the case that two very good terms can outperform six mediocre terms. While we found large differences in the lengths of subjects' queries across all three interface conditions, ultimately all queries, regardless of length, performed similarly. This suggests that perhaps in this experiment at least, there was a 'ceiling effect' with respect to query length. That is, once a certain number of query terms were identified, gains made from additional terms was negligible.

We found support for Hypotheses 3 and our examination of the sources of query terms identified by subjects using Interface_3 demonstrated that the majority of terms were from displayed sentences. Subjects generated approximately six of their own terms using this interface. Overall, term context in the form of sentences provided subjects with an opportunity to identify more useful terms, as well as an opportunity to generate more terms themselves. Although there were no significant differences in mean performance, runs constructed using terms identified with Interface_3 outperformed all other runs with respect to precision-at-10. These results suggest the additional terms provided by subjects with Interface_3 worked to improve the precision of search results, and specifically, precision at the top of the retrieved document list. Future studies might compare variations on this interface, including variations in the size of the box and the type of term context provided.

In this study we found that pseudo relevance feedback runs created from subjects' baseline queries outperformed most experimental runs according to r-precision. This finding in particular motivates several new questions about term relevance feedback interfaces. There is an important question regarding the amount of gain one might expect from term relevance feedback interfaces. How much gain is necessary in order to offset the cost (in terms of effort) to users? In ad-hoc studies of term relevance feedback, gains in retrieval performance are usually quite small and it is unlikely that many of these gains, despite being statistically significant, would make a real difference to users. Accordingly one might ask how much gain is necessary in order for users to perceive a difference in performance. Finally, does the process of interacting with terms and/or document surrogates contribute in other important ways to the information-seeking situation and the resolution of users' information needs? While we have not managed to resolve the discrepancies in previous interactive term relevance feedback research with our study (an unlikely feat for any study), we believe that our results make an important contribution to this body of research and motivate further research.

5. REFERENCES

[1] Allan, J. (2006). HARD Track overview in TREC 2005 high accuracy retrieval from documents. In E. M. Voorhees & L. P. Buckland (Eds.), *TREC-2005, Proceedings of the Fourteenth Text Retrieval Conference*. Washington, D.C.: Government Printing Office.

[2] Anick, P. (2003). Using terminological feedback for web search refinement: A log based study. In *Proceedings of the 26th Annual ACM International Conference on Research and Development in Information Retrieval (SIGIR '03)*, Toronto, CA, 88-95.

[3] Beaulieu, M. (1997). Experiments on interfaces to support query expansion. *Journal of Documentation, 53*(1), 8-19.

[4] Belkin, N. J. (1993). Interaction with texts: Information retrieval as information-seeking behavior. In *Information Retrieval '93*, Germany, 55-66.

[5] Belkin, N. J., Cool, C., Kelly, D., Lin, S. J., Park, S. Y., Perez-Carballo, J., & Sikora, C. (2001). Iterative exploration, design and evaluation of support for query reformulation in interactive information retrieval. *Information Processing & Management, 37*(3), 404-434.

[6] Belkin, N. J., Cool, C., Kelly, D., Lee, H.-J., Muresan, G., Tang, M.-C., & Yuan, X.-J. (2003). Query length in interactive information retrieval. In *Proceedings of the 26th Annual ACM International Conference on Research and Development in Information Retrieval (SIGIR '03)*, Canada, 205-212.

[7] Croft, W. B. & Das, R. (1990). Experiments with query acquisition and use in document retrieval systems. In *Proceedings of the 13th Annual ACM International Conference on Research and Development in Information Retrieval (SIGIR '90)*, Brussels, 349-368.

[8] Efthimiadis, E. N. (2000). Interactive query expansion: A user-based evaluation in a relevance feedback environment. *Journal of the American Society for Information Science & Technology, 51*(11), 989-1003.

[9] Efthimiadis, E. N. (1996). Query expansion. *Annual Review of Information Science & Technology, 31*.

[10] Harman, D. (1988). Towards interactive query expansion. In *Proceedings of the 11th Annual ACM International Conference on Research and Development in Information Retrieval (SIGIR '88)*, Grenoble, 321-333.

[11] Joho, H., Coverson, C., Sanderson, M., & Beaulieu, M. (2002). Hierarchical presentation of expansion terms. In *Proceedings of the 17th Annual ACM Symposium on Applied Computing (SAC '02)*, Madrid, Spain, 645-649.

[12] Kelly, D., Dollu, V. D. & Fu, X. (2005). The loquacious user: A document-independent source of terms for query expansion. In *Proceedings of the 28th Annual ACM International Conference on Research and Development in Information Retrieval (SIGIR '05),* Brazil, 457-464.

[13] Kelly, D. & Fu, X. (2006). University of North Carolina's HARD Track Experiments at TREC 2005. In E. M. Voorhees & L. P. Buckland (Eds.), *TREC-2006, Proceedings of the Fourteenth Text Retrieval Conference*. Washington, D.C.: Government Printing Office.

[14] Koenemann, J., & Belkin, N. J. (1996). A case for interaction: A study of interactive information retrieval behavior and effectiveness. In *Proceedings of the SIGCHI Conference (CHI'96)*, Canada, 205-212.

[15] Larson, R. R. (2001). TREC interactive with Cheshire II. *Information Processing & Management, 37*(3), 485-505.

[16] Magennis, M. & van Rijsbergen, C. J. (1997). The potential and actual effectiveness of interactive query expansion. In *Proceedings of the 20th Annual ACM International Conference on Research and Development in Information Retrieval (SIGIR '97)*, Philadelphia PA, USA, 324-332.

[17] Nameth, Y., Shapira, B., & Taeib-Maimon, M. (2004). Evaluation of the real and perceived value of automatic and interactive query expansion. In *Proceedings of the 27th Annual ACM International Conference on Research and Development in Information Retrieval (SIGIR '04)*, Sheffield, UK, 526-527.

[18] Pennanen, M., & Vakkari, P. (2003). Students' conceptual structure, search process, and outcome while preparing a research proposal: A longitudinal case study. *Journal of the American Society for Information Science & Technology, 54*(8), 759-770.

[19] Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., Gatford, M. (1995). Okapi at TREC-3. In D. Harman (Ed.), *TREC-3, Proceedings of the Third Text Retrieval Conference*. Washington, D.C.: Government Printing Office.

[20] Ruthven, I. (2003). Re-examining the potential effectiveness of interactive query expansion. In *Proceedings of the 26th Annual ACM International Conference on Research and Development in Information Retrieval (SIGIR '03)*, Toronto, CA, 213-220.

[21] Spink, A. (1994). Term relevance feedback and query expansion: Relation to design. In *Proceedings of the 17th Annual ACM International Conference on Research and Development in Information Retrieval (SIGIR '94)*, Dublin, Ireland, 81-90.

[22] Spink, A., & Jansen, B. J. (2004). *Web search: Public searching of the Web*. Kluwer Academic Publishers.

[23] Voorhees, E. M. (2006). Overview of TREC 2006. In E. M. Voorhees & L. P. Buckland (Eds.), *TREC-2006, Proceedings of the Fourteenth Text Retrieval Conference*. Washington, D.C.: Government Printing Office.

[24] Wu, M., Fuller, M., & Wilkinson, R. (2001). Using clustering and classification approaches in interactive retrieval. *Information Processing & Management, 37*(3), 459-484.

[25] Yang, K., Maglaughlin, K. L., & Newby, G. B. (2001). Passage feedback with IRIS. *Information Processing & Management, 37*(3), 521-541.