University of North Carolina's HARD Track Experiment at TREC 2005

Diane Kelly and Xin Fu
School of Information and Library Science
University of North Carolina
100 Manning Hall, CB#3360
Chapel Hill, NC 27599-3360

[dianek | fu]@email.unc.edu

1. Introduction

In this year's HARD Track, we focused on two aspects related to the elicitation of relevance feedback: the display of document surrogates and features for identifying and selecting terms. We looked at these issues with respect to interactive query expansion (IQE). In typical interactive query expansion scenarios, users mark documents that they find relevant and the system automatically extracts terms from these documents and adds them to users' queries, or suggests potential query terms from these documents and allows users to determine which of these terms are added to their queries. While a large number of studies have been conducted on IQE, results of such studies do not convey a consistent picture of IQE use and effectiveness.

Empirical, laboratory-based studies have led to the general finding that users of experimental interactive IR systems desire IQE features (c.f., Beaulieu, 1997; Belkin, et al., 2001). However, much of the evidence from these studies indicates that relevance feedback features are rarely used and when they are used, they are unlikely to result in retrieval improvements. For instance, some studies have found that users do not select many terms (Beaulieu, 1997; Belkin, et al., 2001), while other studies have found that users select terms, but that these terms do not necessarily improve performance (Anick, 2003). This has been attributed to problems related to the design of relevance feedback interfaces (Ruthven, 2003), task complexity and the user's lack of additional cognitive resources (Belkin, et al., 2001), and the amount of extra time required to use such features. Users in a series of studies by Belkin, et al. (2001) rarely used relevance feedback features and often expressed confusion over suggested terms. In a study of simulated interactive query expansion, Ruthven (2003) demonstrated that users are less likely than systems to select effective terms for query expansion. Ruthven (2003) demonstrated some potential benefit of term relevance feedback when the best terms were used in query expansion, but 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 (2003) found that users made use of a term suggestion feature to expand and refine their queries. However, this did not result in improvements in retrieval performance, which suggests that terms users selected were not particularly good. Conversely, in another study of an operational retrieval system, Efthimiadis (2000) found that users selected about one-third of terms suggested by the system and that, in general, these terms improved retrieval performance. Harman (1988) also demonstrated that IQE led to retrieval improvements.

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. 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. One purpose of the current study is to investigate if an interface that provides term context helps users make better query expansion decisions. Previous research does not provide a clear idea about how term context will affect user behavior and retrieval. 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 discriminant when selecting terms? Consequently, will selected terms improve or worsen retrieval performance? We hypothesize that users will select more terms when they are presented in context than when they are presented in isolation (H1) and that these additional terms will improve retrieval performance (H2).

In this study, we are also interested in investigating users' abilities to suggest terms to add to their queries given appropriate stimulation. It has been suggested in the literature that only through interaction with texts can users come to understand and learn about their information needs (Belkin, 1993). Furthermore, in our previous work, we found that with appropriate probing, users could articulate additional information about their information needs beyond what they articulated in their initial queries (Kelly, Dollu, & Fu, 2005). We propose that interactions with text surrogates can stimulate users' thinking about their information needs and that this stimulation can help users identify additional terms to add to their queries. Specifically, we hypothesize that users will identify more terms using an interface that presents sentence-level document surrogates and elicits free-form text input than an interface that presents these same surrogates with check boxes (H3). We anticipate that sentences will provide users with ideas about terms for query expansion in both a direct fashion (i.e., terms contained within sentences) and an indirect fashion (i.e., via interaction and stimulation, where users think of additional terms not contained within sentences). We further hypothesize that terms suggested by users via the former interface will result in better retrieval performance than those selected via the latter interface (H4).

2. Clarification Forms

We submitted three clarifications forms (CFs), each demonstrating a different method of displaying document surrogates and eliciting query expansion terms. Table 1 summarizes the three methods. The first form 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 form 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 sentences. The final form displayed the same sentences from Form 2, but with a text box for input. 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.

Table 1. Clarification form design

Form	Display	Elicitation
1	Terms	Check-boxes
2	Terms and sentences	Check-boxes
3	Sentences	Text box

Screen shots of the interfaces are displayed below in Figures 1-3. The comparison between Form 1 and Form 2 allowed us to explore hypotheses 1 and 2, while the comparison between Form 1 and Form 3 allowed us to test hypotheses 3 and 4. We designed our forms to look slightly different from one another with respect to style (e.g., font style, background color). Since users would complete all three forms for each of their topics, we hoped to minimize the possibility that users would recognize them as a set and react accordingly.

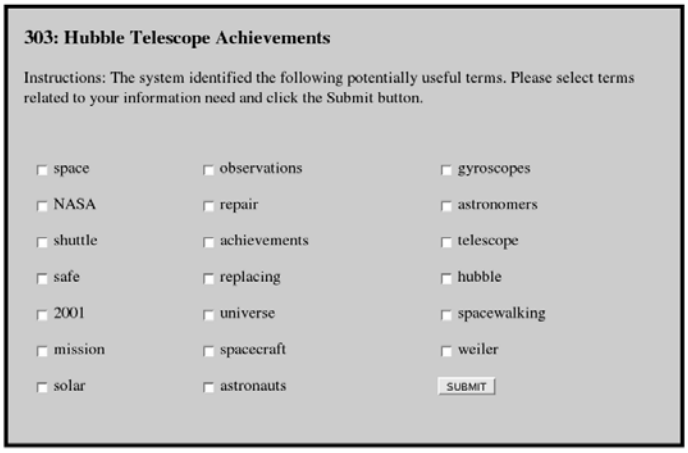

Figure 1. Clarification Form 1

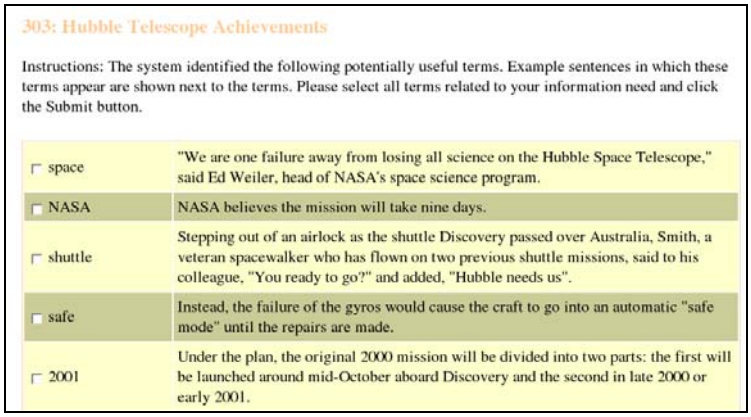

Figure 2. Clarification Form 2

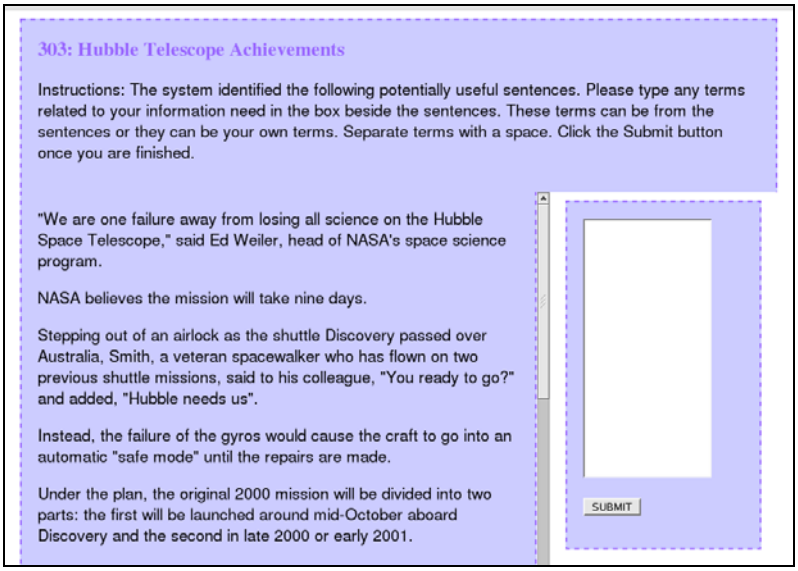

Figure 3. Clarification Form 3

3. Retrieval System

We used the Lemur IR toolkit (http://www.lemurproject.org) to conduct our experiments, with its basic defaults for indexing (BuildIndex function), and Okapi BM25 for retrieval. Although we made use of a basic stop word and acronym list, we did not use a stemmer. Our baseline run consisted of the *title* and *description* for each topic. Our experimental runs consisted of adding selected or suggested terms to baseline queries.

To populate our clarification forms, we modified Lemur's basic feature (Reteval) so that for each topic, terms identified by the system to use for pseudo relevance feedback were printed to a file, along with document identification numbers from which these terms were extracted. We set the pseudo relevance feedback parameter to use the top twenty ranking terms from the top ten ranking documents. The technique used for selecting terms is based on Robertson Selection Value (RSV) and described more fully in Robertson, Walker, Jones, & Hancock-Beaulieu (1995); this technique is included as part of the Lemur toolkit.

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 one word queries and corresponding sentence level documents for retrieval. We used the top result for each query to populate Form 2 and Form 3.

4. Results and Discussion

In this section, we first present the responses that we received from each of the three clarification forms and the statistical tests of H1 and H3. This is followed by a presentation and comparison of retrieval results for each technique to test H2 and H4.

4.1 Clarification form responses

Table 2 lists the mean number of terms that users marked as relevant on Form 1 and Form 2 as well as the mean number of terms that they entered on Form 3. The figure for Form 3 is calculated based on terms that the retrieval system actually used in the experimental run, after breaking hyphen connected terms into two (e.g., "family-planning" to "family planning") and removing stop words. When raw data is considered, Form 3 elicited an average of 11.12 terms with a standard deviation of 8.559. Overall, Form 3 elicited the most terms from users and Form 2 the least. Paired sample t-tests revealed significant differences between Form 1 and Form 2 [$t(49)=3.404$, $p<0.05$] and between Form 2 and Form 3 [$t(49)=-2.255$, $p<0.05$], but not between Form 1 and Form 3 [$t(49)=-0.535$, $p=0.595$].

Table 2. Means and standard deviations for the number of terms users selected or entered on CFs and the amount of time in seconds spent accomplishing this

	Terms	**Time**
Form 1	9.94 (3.835)	45.38 (31.07)
Form 2	8.06 (4.723)	141.92 (50.57)
Form 3	10.48 (7.762)	148.50 (47.25)

Table 2 also lists the mean number of seconds users spent marking or identifying terms with each type of form. Recall that users were limited to 180 seconds (3 minutes) per form. Users spent the least amount of time on Form 1, which is no surprise given that this form contained terms and checkboxes only. The frequency distribution for time reveals that most users (n=48) spent less than 78 seconds completing Form 1. Users spent an average of 141.92 seconds on Form 2, which contained sentences and checkboxes. It is likely that many users did not have time

to complete this form. The frequency distribution for time for Form 2 indicates that 13 users stopped at 180 seconds and 13 users stopped at 181, 182 and 183 seconds. Assuming that users were stopped automatically at the end of 180 seconds, it is likely that these 26 users did not evaluate the entire form. This may explain why users identified the fewest terms with Form 2. Finally, users spent an average of 148.50 seconds completing Form 3 which presented sentences and a free-form textbox. Although on average, users spent the most time completing this form, fewer users were stopped by the time limit when completing this form than when completing Form 2. The distribution shows that the time for 12 users was greater than or equal to 180 seconds. It is interesting to note that the maximum time for Form 3 (267 seconds) greatly exceeded that for Form 2 (183 seconds). We are unsure if this represents a processing delay or some other problem.

4.1.1 Form 1 vs. Form 2 (H1)

The significant difference between Form 1 and Form 2 does not provide support for hypothesis H1 since it is in the opposite direction than what we hypothesized. In a topic level analysis of the number of selected terms, we noted that users selected equal number of terms from the two forms in nine cases, more terms from Form 1 in 30 cases and more terms from Form 2 in 11 cases. We are mindful that time may have impacted these results.

After a further examination of the selected terms, we noted that of the 50 topics only five received identical term judgments from both Form 1 and 2. For 16 topics, the selected term set from Form 2 was a subset of the set from Form 1, for 5 topics the selected term set from Form 1 was a subset of Form 2, and for 24 topics there was little overlap between terms. Interestingly, for one of these topics the user marked six terms as relevant on Form 1 and two terms as relevant on Form 2, but these were all different terms. In five cases, users did not mark any terms, once when responding to Form 1 and four times when responding to Form 2. Surprisingly, for these four cases, users selected 2, 4, 11 and 14 relevant terms from the corresponding Form 1.

4.1.2 Form 2 vs. Form 3 (H3)

The difference between Forms 2 and 3 supports the hypothesis that the free-form input box on Form 3 would elicit significantly more terms from users than the check boxes on Form 2. A topic level analysis indicated that users entered at least one term on Form 3 in all 50 cases, and for 32 topics, users entered an equal or greater number of terms on Form 3 than they selected from Form 2. Hypothesis H3 is supported by this data. Again, we caution the reader that time may have contributed to these results.

A comparison of results on a term-by-term basis leads to interesting findings. In no case were terms identical across the two forms for any topic and for six cases, terms selected or entered on Forms 2 and 3 by the same user were exclusive; that is, there was no overlap in these sets of terms. We further examined sources of terms entered on Form 3. In this analysis, all entered terms are considered without applying the stop word list. The average number of terms elicited by Form 3 was 11.12 (std=8.56) terms; 3.26 (std=3.10) of these terms were identified by users when using From 2 while 7.86 (std=6.75) were new. If duplicate terms are removed, the mean drops slightly to 7.38 (std=6.07). Thus, on average, Form 3 elicited seven additional terms from users. Among the 7.86 new terms, 1.22 (std=2.14) were suggested terms from Form 2, 4.56 (std=5.12) were contained within displayed sentences, and 2.08 (std=3.63) were user-generated.

Interestingly, when considering all terms identified with Form 3, approximately 9 terms came from displayed sentences and only 2.08 were user-generated. Of these 9 terms, about 4.48 were terms the system suggested via term suggestion and used to populate Forms 1 and 2, and 4.56 were terms that the system had access to (i.e., terms contained within the suggested sentences), but did not suggest.

Users entered at least one new term for almost half (23) of the topics. For five topics, terms on Form 3 were all user-generated. Often times, additional terms were synonyms or

antonyms of terms in sentences, or subordinates of some general term that was displayed. For example, in topic 354, "Journalist Risks," the user entered several terms corresponding to specific forms of risk ("harassed," "detained," "killed," "arrested," "injured," "beaten," etc.). Some of these terms were mentioned in the "description" field of the topic ["Identify instances where a journalist has been put at risk (e.g., killed, arrested or taken hostage) in the performance of his work."].

In general, these results demonstrate the overall success of using sentences as stimulators and providing free-form text input for users to identify additional terms to add to their queries. At the same time, however, we wish to point out that the effectiveness of this interaction technique may vary between users. There were significant individual differences among the six users in this study with respect to the number of user-generated terms. Both ANOVAs investigating user differences with respect to the number of user-generated terms and the ratio of user-generated terms to the total number of terms led to significant results [$F(5, 44)=3.36$, $p<0.05$; $F(5, 44)=7.68$, $p<0.01$]. In particular, we observed that User D entered a significantly larger number (71.7%) of user-generated terms than any other user (the second highest is only 30.5%). On the contrary, User E only entered one user-generated term for one topic on which he or she worked and none for his or her other topics.

4.2 Retrieval results

Table 3 shows the R-precision, mean average precision (MAP), and precision at ten scores (with standard deviations) for our baseline and the three experimental runs. final_1, final_2 and final_3 runs consist of baseline queries plus query expansion using terms obtained from Forms 1, 2, and 3, respectively. We include one pseudo relevance feedback run as an additional baseline. This run is equivalent to adding all terms from Form 1 to baseline queries.

Table 3. Means and standard deviations for R-precision, MAP and Precision at 10 for each run

	R-precision	**MAP**	**Precision @ 10**
Baseline	0.218 (0.160)	0.160 (0.162)	0.346 (0.294)
Pseudo Relevance	0.264 (0.204)	0.224 (0.216)	0.422 (0.361)
Final_1	0.283 (0.192)	0.235 (0.203)	0.436 (0.361)
Final_2	0.268 (0.194)	0.219 (0.209)	0.426 (0.343)
Final_3	0.279 (0.199)	0.228 (0.205)	0.468 (0.353)

Since R-precision is the official measure of the Track, we used this measure to determine if significant differences existed between runs. The paired sample t-tests indicated significant improvements in all three final runs over the baseline [final_1: $t(49)=4.784$, final_2: $t(49)=3.350$, final_3: $t(49)=3.289$, all $p<0.05$]. Table 4 shows the comparison of our final runs to our baseline run at the topic level. All three techniques improved the average precision for around 66% of the topics. These results provide some evidence for the benefit of the relevance feedback techniques used in this study.

Table 4. Performance of final runs versus baseline run according to number of topics that were improved, worsened or stayed the same

	Better	**Same**	**Worse**
Final_1	36	3	11
Final_2	29	7	14
Final_3	29	6	15

4.2.1 Form 1 vs. Form 2 (H2) and Form 2 vs. Form 3 (H4)

Recall that in H2 and H4, we hypothesized that providing context for term selection and providing document surrogates and a free-form term input interface would not only result in the elicitation of more terms, but that these additional terms would lead to improved retrieval results. The R-precision scores reported in Table 3 do not support H2. Even though the R-precision of final_1 was higher than that of final_2, the paired sample t-test led to a non-significant result [$t(49)=1.338$, $p=0.187$]. The R-precision of final_3 was higher than final_2, but again, no statistically significant difference in means was found [$t(49)=-0.529$, $p=0.599$].

4.3 Cross site comparison

In this subsection, we briefly report a comparison between our results and results of other sites. As Table 5 shows, our baseline run falls below the median baseline for most topics, but all three of our experimental runs position above the median for over half of the topics.

Table 5. UNC's topic performance versus other sites

	Best		Median		Worst
Baseline	0	17	4	28	1
Final1	1	26	4	16	3
Final2	2	25	3	18	2
Final3	2	25	2	18	3

5. Conclusions

In this project, we focused on two aspects related to the elicitation of relevance feedback: the display of document surrogates and features for identifying and selecting terms. We compared three forms for eliciting relevance feedback. The first form 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 form 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. The final form displayed the same sentences from Form 2, but with a text box for input.

We hypothesized that users would select more terms when they were presented in context (Form 2) than when they were presented in isolation (Form 1) and that these additional terms would improve retrieval performance (H1 and H2, respectively). Statistical tests did not support either hypothesis. In fact, users identified significantly fewer terms with Form 2 than with Form 1, which was contrary to our expectations. It might have been the case that term context allowed users to be more selective and discriminating, which was why fewer terms on average were identified with Form 2. However, the R-precision score for Form 1 was higher than the R-precision score for Form 2 (although not significantly so), which suggests that the quality of terms identified on Form 1 was not necessarily poor. Thus, a more likely explanation for this is that users simply did not have enough time to complete Form 2. Given more time, users may have selected more terms from Form 2. Our analysis of terms selected by each user with Form 1 and Form 2 indicated no consistent pattern across forms with respect to term selection.

We also hypothesized that users would identify more terms using an interface that presented sentence-level document surrogates and elicited free-form text input (Form 3) than an interface that presented these same surrogates with check boxes (Form 2) (H3). We further hypothesized that terms suggested by users via the former interface would result in better retrieval performance than those selected via the latter interface (H4). Results demonstrated that users identified significantly more terms with Form 3 than with Form 2, which supported H3. In fact,

users identified the most terms with Form 3 despite it being a bit more labor-intensive. However, as with the differences between Forms 1 and 2, the differences between Forms 2 and 3 may be attributed to time, or lack thereof.

When considering all terms identified with Form 3, approximately 9 terms came from displayed sentences and only 2.08 were user-generated. Of these 9 terms, about 4.48 were system suggested terms from Forms 1 and 2, and 4.56 were terms that the system had access to (i.e., terms contained within the suggested sentences), but did not suggest. These results demonstrate that using sentences as stimulators and providing free-form input leads users to identify significantly more terms than when they are presented with suggested terms, sentences and check-boxes. Performance results were in the general direction of H4, that is, the R-precision score for Form 3 was higher than the R-precision score for Form 2, but this difference was not statistically significant.

Finally, we found significant differences for many measures according to user. This is not surprising given the experimental setup of the study. Specifically, six users were responsible for assessing fifty topics and responding to a large number of clarification forms, so individual user differences are likely to be present in the data. We are currently conducting a between-subjects follow-up study with approximately 60 users to further test our three experimental forms and research hypotheses.

6. References

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.

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

Belkin, N. J. (1993). Interaction with texts: Information retrieval as information-seeking behavior. In: Information retrieval '93. Von der Modellierung zur Anwendung. Konstanz: Universitaetsverlag Konstanz, 55-66.

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.

Efthimiadis, E. (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.

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.

Kelly, D., Dollu, V. J., & 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)*, Salvador, Brazil, 457-464.

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.

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.