

The Loquacious User: A Document-Independent Source of Terms for Query Expansion

Diane Kelly, Vijay Deepak Dollu, & Xin Fu
School of Information & Library Science
University of North Carolina at Chapel Hill
Chapel Hill, NC USA 27599-3360
+1 919.962.8065

[dianek | vijayd | fu] @ email.unc.edu

ABSTRACT

In this paper we investigate the effectiveness of a document-independent technique for eliciting feedback from users about their information problems. We propose that such a technique can be used to elicit terms from users for use in query expansion and as a follow-up when ambiguous queries are initially posed by users. We design a feedback form to obtain additional information from users, administer the form to users after initial querying, and create a series of experimental runs based on the information that we obtained from the form. Results demonstrate that the form was successful at eliciting more information from users and that this additional information significantly improved retrieval performance. Our results further demonstrate a strong relationship between query length and performance.

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, elicitation, clarification form, query length, user feedback, information need, problem description, polyrepresentation

1. INTRODUCTION

It is well-known that users often have a difficult time articulating their information needs [5] and that users typically pose very short queries, usually between two and three words in length [18]. While it might be the case that the difficulty that users have with articulating their information needs causes their queries to be short, it has recently been argued that one reason users pose short queries is because traditional search interfaces encourage them to do so [8]. It is also well-known that in best-match systems longer queries usually result in better retrieval performance at least as

explored in batch-mode information retrieval (IR) experiments [c.f. 10]. Thus, there is an apparent mismatch between what interfaces encourage users to do, what users are doing and what has been demonstrated to result in good retrieval.

Query expansion is an effective technique for increasing the number of terms contained in users' queries and improving retrieval performance [13]. The underlying assumption behind query expansion is that more terms are better; that is, that query length is associated positively with retrieval performance. Query expansion techniques can assist users with increasing the length of their queries through automatic and interactive techniques. Automatic query expansion (AQE) occurs when the system selects appropriate terms for use in query expansion and automatically adds these terms to users' queries. In most cases, terms that the system suggests come from documents that the system believes have a high probability of being relevant, as in the case of pseudo relevance feedback, or from additional tools such as thesauri [20]. Ruthven [25] observes, "one argument in favor of AQE is that the system has access to more statistical information on the relative utility of expansion terms and can make a better selection of which terms to add to the user's query".

Conversely, interactive query expansion (IQE) gives the user control over which of a set of system suggested terms are added to the query, and in some cases, which sources are used to automatically generate this set of terms. The user can mark documents that he or she finds relevant and the system can then suggest potential query terms from these documents. In this situation, the selection of sources for terms and the final selection of terms (i.e. the terms that are actually added to the query) are under the user's control. In a variation of this technique, the system automatically suggests terms that it believes are potentially useful for query expansion. While users have the final say with respect to which terms are added to the query, the control that users exert over the process is less than that described in the first interactive technique since they do not select the source from which terms are taken.

IQE is thought advantageous since users have control over what gets added to their queries [7, 22, 25]. Interestingly, at its inception, relevance feedback was thought of as a technique for assisting users arrive at an *ideal* query. While it was acknowledged that users had a difficult time articulating queries, the predominant viewpoint was that by providing users with terms used to index documents, they would be equipped with a more appropriate vocabulary with which to formulate queries; all users needed to do was select the most appropriate terms from the display.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SIGIR '05, August 15–19, 2005, Salvador, Brazil.

Copyright 2005 ACM 1-59593-034-5/05/0008...\$5.00.

Empirical studies have led to the general finding that users of interactive IR systems desire explicit relevance feedback features and, in particular, term suggestion features [4, 7]. However, much of the evidence from laboratory studies has indicated that relevance feedback features are not used. While users often report a desire for relevance feedback and term suggestion, they do not actually use these features during their searching activities. This, in part, has led to the general belief that users are unwilling to engage in explicit relevance feedback. However, things may be changing; recently Anick [3] demonstrated in a web-based study, that users made use of a term suggestion feature to expand and refine their queries. Thus, it may be the case that users are becoming more comfortable with providing feedback when such techniques are integrated effectively into searching. Unfortunately though, Anick [3] did not demonstrate any improvements in performance, which suggest that terms users select (or that the system suggests) may not be optimal.

Researchers have also investigated users' ability to select good terms for query expansion [15, 23, 25]. In a study of simulated interactive query expansion, Ruthven [25] demonstrated that users are less likely than systems to select effective terms for query expansion. Ruthven [25] used a range of query expansion terms from 1 to 15, and found that providing the system with more query expansion terms did not necessarily improve retrieval performance. Ruthven [25] demonstrated some potential benefit of IQE if 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. Thus, while batch-mode experiments evaluating the effectiveness of automatic query expansion have been favorable, experiments involving users have had mixed results.

Another approach to assisting users build better queries is to encourage them to provide longer descriptions of their information needs at the time of initial querying, or eliciting 'enhanced' queries [8, 11]. For instance, Belkin, et al. [8] compared an experimental query interface, which was designed to elicit longer queries from users, with the standard query interface found on most retrieval systems. Although Belkin, et al. [8] found that users entered longer queries using the experimental interface than they did when using the standard interface and that query length was associated positively with user satisfaction, there was no correlation between query length and performance. Thus, the finding from batch-mode experiments demonstrating a positive relationship between query length and retrieval performance has not been replicated in experiments with users.

Finally, Ingwersen's theory of polyrepresentation [17] suggests that obtaining multiple representations of a single information need is a better approach to representing user needs than solitary, isolated queries. Empirical work by Belkin et al. [6] examined the impact of different query representations on retrieval and found that a variety of representations led to better performance. This work suggests that eliciting a variety of information need representations from users, and using this information in different combinations, is likely to improve retrieval.

In this paper, we argue that users should be considered as useful sources of terms for query expansion, independent of any information from the system. While we accept that in some situations users are unable to clearly articulate their information needs because they lack the knowledge or vocabulary to describe

their needs, we propose that users have not been probed in the most effective way; that is, that traditional interfaces for querying and for relevance feedback are not optimal for eliciting the most robust and useful description from users of their information needs. Thus, we are interested in taking greater advantage of the user as a source of terms for query expansion and incorporating the idea of polyrepresentation into feedback.

Our interest was motivated by the idea that traditional query expansion techniques, which typically present top-ranked documents or keywords to users for feedback, are unlikely to work well in all retrieval situations, especially when ambiguous queries are posed, because there is a large chance that documents retrieved in response to such queries will be irrelevant. Furthermore, as previously described, users often have a difficult time selecting the best terms for query expansion even when they are willing to do so. In many cases, users do not understand why certain terms have been suggested and in many other cases, the terms which the system suggests are not necessarily the best. Rather than force users to interact with system suggested terms or documents, we were interested in investigating users as sources of terms for query expansion, independent of any information from the system. We were further interested in examining the relationship between query length and performance in a situation where queries were strictly user generated.

Based on the research described previously, we developed and evaluated a generic, document-independent feedback form that could be used in multiple information-seeking situations. We accomplished this by creating an online form and presenting this form to users after initial querying in hopes of eliciting more complete descriptions of their information needs. We used the information that we obtained from the form as a source of terms for query expansion, created a series of experimental runs based on this information and evaluated the performance of each run.

2. METHOD

This experiment was carried out as part of the 2004 TREC High Accuracy Retrieval from Documents Track (HARD) [2]. The Track had three objectives: (1) to determine if metadata about the query, user or search context could be used to improve retrieval; (2) to determine if a single, highly focused interaction with the user could be used to improve retrieval; and (3) to determine if passage retrieval could be used to further improve retrieval.

TREC participants were free to select any of these objectives to explore in their experiments. The experiment presented in this paper focuses only on objective (2), since we were interested in investigating various techniques for eliciting additional information from users about their information problems. Track participants who explored (2) created *clarification forms*, which constituted the "highly focused interaction." In the next section, we discuss the experimental protocol of the HARD track, which is necessarily part of our experimental protocol. This is followed by a discussion of our clarification form.

2.1 HARD Track Experimental Protocol

The HARD track was conducted in collaboration with the Linguistic Data Consortium (LDC) at the University of Pennsylvania. The LDC recruited and managed users, and mediated all interactions that occurred between users and Track participants. The experimental protocol of the HARD track

differed a bit from traditional interactive IR user studies. While users created topics and evaluated documents, they did not conduct any interactive searching, or directly interact with any Track participants.

2.1.1 Users & Topics

The LDC recruited thirteen users to participate in the project. In total, these thirteen users contributed 50 topics. However, only 45 topics were used in this study because for five of the fifty topics, no relevant documents were retrieved by any system participating in the Track. Most participants were in their early to mid-twenties; several were undergraduate students, others had already earned an undergraduate degree and several others held both undergraduate and graduate degrees.

Users were presented with an online Topic and Metadata Creation Form. With this form, users created TREC-style topics [26], which contained a title, description and topic-narrative.

2.1.2 Metadata

As part of the HARD track, users also provided metadata, such as desired genre of document, and familiarity with topic, when they created their topics. Metadata items were determined jointly by Track participants and the LDC before the time of topic creation. Metadata are not described in detail in this paper since they were not used in the study.

2.1.3 Clarification Forms

Once topics had been created, they were distributed to all Track participants without related metadata. Each Track participant submitted a baseline run to NIST and, several days after this, 1 to 3 clarification forms per topic to the LDC (if they were investigating objective (2) described earlier). For each topic, the clarification form was a Web page that elicited information about the topic or the user (e.g., disambiguating words in the topic or finding out more about the user's interests).

To better understand clarification forms, those used in the 2003 HARD Track are described briefly [1]. There were two general approaches to generating clarification forms in the 2003 HARD Track. The first of these was to use some sort of document surrogate, which had been retrieved in response to users' baseline queries, to populate the clarification form. Users were shown these surrogates, which typically consisted of terms/phrases, sentences and passages (including headlines), and asked to mark them in some way. The second approach to generating clarification forms was to present users with questionnaire-type items whose content was not generated from initial search results. These items included those related to users' previous searching experiences and general preferences, as well as those that asked users to enter additional key terms describing their topics. The information elicited from the clarification forms was used for a variety of experimental techniques, although most often, query expansion and document re-ranking.

Several mandatory guidelines were provided for constructing clarification forms. Clarification forms had to be self-contained HTML Web pages and had to display correctly on a 16-inch monitor with 1152 X 900 resolution using Netscape v4.78. Guidelines also dictated that interactive scripting could not be used. Permissible form fields included text boxes, radio buttons, check boxes, and drop-down menu selectors. Forms from all Track participants who submitted clarification forms were

presented to users in random order. Users were allowed to spend up to three minutes completing each form. Once the clarification forms were completed, this data, along with the metadata, were distributed to Track participants for use in experimental runs.

2.1.4 Corpus

The corpus used in the 2004 HARD track experiments was approximately 1.5 GB in size and contained about 650,000 English-only newswire documents. In addition to these documents, the corpus contained 3,134 documents from Salon.com. All documents were from the year 2003.

2.1.5 Relevance Judgments

Users assessed the relevance of documents retrieved in response to their topics. The determination of which documents to show to users was made using the standard TREC pooling method [25]. Users judged documents as *off-topic*, *on-topic*, and *relevant*, where *off-topic* documents were those that were not about the topic, *on-topic* documents were those that were about the topic, but that did not satisfy one or more metadata requirements, and *relevant* documents were those that were about the topic and satisfied all metadata requirements. Since we did not investigate metadata in this study, we merged *on-topic* and *relevant* documents to arrive at the total set of relevant documents.

2.2 Experimental Clarification Form

Based on our review of previous approaches to clarification forms in the 2003 HARD Track and our particular research interests, we designed a clarification form that consisted of four questions and could be used for all topics without modification. This clarification form is displayed in Figure 1.

HARD-401 Bass Amps

[1] How many times have you searched for information about this topic in the past?

Never
 1 or 2 times
 3 or 4 times
 5 or more times

[2] Describe what you already know about the topic.

[3] Why do you want to know about this topic?

[4] Please input any additional keywords that describe your topic.

submit

Figure 1. Clarification Form

The first question that we presented to users was a familiarity question, which asked users to indicate how many times they had searched for information about their topics in the past. We do not report on users' responses to this question in this paper, but will mention our motivation for including this question as part of our clarification form. As part of the metadata, users were required to respond to another question designed to assess topic familiarity;

we took advantage of this to investigate an alternative measure of familiarity, and to see how well these measures correlated.

Questions 2, 3 and 4 were designed to elicit information from users about their topics for query expansion. In designing clarification form features to elicit this information, we were careful to use large text boxes that allowed users to view the entirety of their responses and hopefully, as found in previous studies [8, 19], encourage them to type in longer responses than they would if presented with a short line. Questions 2 and 3 were open-ended questions (although 2 is presented as a statement), and encouraged users to respond in natural language. Question 2 asked users to describe what they already know about the topic, and Question 3 asked users to indicate why they want to know about the topic. Our goal in using these questions was to encourage users to talk more about their topics, and hopefully in doing so, have them provide additional information that might prove useful in retrieval. Our selection of these two questions was based on an examination of previous research on face-to-face reference interviews [c.f. 16], reference textbooks describing best practice [c.f. 20], and the notion of “polyrepresentative extraction” of information needs [17]. Polyrepresentative extraction suggests eliciting a “what” (i.e. what is currently known about the topic), and a “why” (i.e. why a user wants to know about topic) from the user to use in retrieval.

Question 4 asked users to list any additional keywords describing their topics. A number of participants from last year’s TREC used a question like this, some with quite successful results [c.f. 14]. Thus, we included this on our form with hopes that it would again provide some useful data. It was also the case that the majority of clarification forms from last year asked users to make a selection of good terms from a list of terms from top-ranking documents. Thus, we further hoped that this question would allow us to take advantage of the priming that users might receive by being exposed to such lists of terms before they completed our form in the experimental rotation. The assumption, of course, is that if users see a good term on another clarification form, then there is a possibility that they will remember and enter it when they reach our form. We certainly recognize this as potentially confounding the results of our own experiment, but the protocol of the HARD Track makes this unavoidable anyway.

2.3 Baseline & Experimental Runs

We used the Lemur IR toolkit (<http://www.lemurproject.org>) to conduct our retrieval 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.

Our baseline run consisted of the *title* and *description* for each topic. We used this information for our baseline run because we felt that it most closely approximated the length of queries posed by users in online searching environments [18]. Using text from both of these fields created queries that were longer than what is reported in [18], but we did not want our baseline run to produce particularly poor results either.

Our experimental runs were constructed from the information that we obtained from users with our clarification form (Q2, Q3 and Q4). These various runs are displayed in Table 1. In all cases, we used the additional terms elicited from users by each of these techniques as a source of terms for query expansion. For runs where the same term appeared in multiple places (e.g. in the title,

Q2 and Q3), the initial weight of the term was multiplied by the number of times the term appeared. Each of these runs consisted of the baseline query (title + description) plus any additional text provided by the clarification form question.

Table 1. Experimental runs

| Source of Terms | RUN ID |
|--|---|
| Baseline (title + description) | baseline |
| Baseline (BL) + Pseudo Relevance Feedback | pseudo05, pseudo10, pseudo20, pseudo50 |
| BL + Relevance Feedback with Relevant Documents | rfrel05, rfrel10, rfrel20, rfrel50 |
| BL + Clarification Form Q2 | Q2 |
| BL + Clarification Form Q3 | Q3 |
| BL + Clarification Form Q4 | Q4 |
| BL + Combination of Clarification Form Questions | Q3Q4 |
| | Q2Q3 |
| | Q2Q4 |
| | Q234 |

We included two other types of runs as baselines, one using pseudo relevance feedback and another using what we consider as the upper bounds for relevance feedback. We created a set of pseudo relevance feedback runs, each of which extracted a varying number of terms from the top 10 retrieved documents for each topic from our baseline. The four pseudo relevance feedback runs used the top 5, 10, 20 and 50 terms for query expansion. We created a set of upper bound relevance feedback runs, each of which extracted query expansion terms from 10 randomly selected relevant documents. The four runs based on relevant documents used the top 5, 10, 20 and 50 terms for expansion. The technique used for selecting terms for query expansion in both sets of runs was based on Robertson Selection Value (RSV) and is described more fully in [24]; this technique is included as part of the Lemur toolkit.

These additional runs were included as baselines so that we would have some way to evaluate the use of terms provided by users with clarification forms for query expansion, against the use of terms generated using automatic techniques. If runs using terms from the clarification form were comparable in performance to runs using pseudo relevance feedback, then the value of our clarification form, which requires users to expend extra effort, would be questionable. Although our upper bound run would not be possible in a real retrieval situation, we felt that it would either give us an idea of the upper bound if it performed well or, if it performed poorly, provide additional support for the use of our document-independent clarification form as a technique for identifying terms for query expansion. The number of terms selected for use in each of the baseline relevance feedback runs (5, 10, 20 and 50) was based on the range of terms that we elicited with our clarification form questions.

2.4 Evaluation Measure

We used mean average precision (MAP), a standard TREC evaluation measure, to evaluate our results [26]. Mean average precision is the average of all topic-level average precision scores (essentially an average of averages). With respect to each topic, average precision is the mean precision after each relevant document is retrieved (zero is used for relevant documents not retrieved). Thus, a MAP score makes some use of recall in its computation; because of this we do not report recall values.

3. RESULTS

The mean number of terms elicited from users with each clarification form question is displayed in Figure 2. This figure also includes the mean number of terms contained in the baseline queries (mean = 9.33; standard deviation = 4.38). On average (with standard deviation), users provided 16.18 (11.66) terms in their responses to Q2, 10.67 (7.22) in their responses to Q3, and 2.33 (4.30) in their responses to Q4. Each of these numbers reflects the average number of terms that the system used for query expansion, rather than the average raw number of terms contained in users' responses. These means assisted us with identifying an appropriate number of terms to use in our relevance feedback runs, since we wanted the number of terms used for query expansion to be comparable to what was elicited with the clarification form.

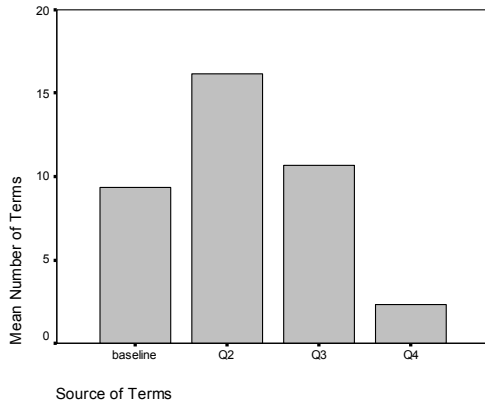


Figure 2. Mean number of terms provided by users according to source

From the means and standard deviations, it is clear that the length of users' responses varied considerably according to question, and that the length of users' responses varied within each question. For the 45 topics, users provided some response to Q2 for 40 topics, Q3 for 42 topics, and Q4 for only about half (20) of the topics. We noted a large number of spelling errors in users' raw responses to Q2 and Q3 (18 and 10), which we corrected.

We were a bit surprised by the results of Q4, by both the actual number of users who responded and by the average number of terms that these users provided. Given the success of some groups from the 2003 HARD Track using a similar question, we expected to elicit more terms with this question than with Q2 or Q3. This was especially true since we expected users to be primed to respond to this question based on their interactions with clarification forms provided by other Track participants. The low response rate to Q4 might be a result of users' preferences for communicating in natural language rather than keywords. These results might further provide some support that Q2 and Q3 from

our clarification form are better techniques for eliciting information from users about their information problems than Q4. However, these results might also be explained by an order effect. Unfortunately, we did not set up our clarification form to explicitly compare the differences in these questions. Instead, questions were always presented in the exact same order, rather than rotated. We purposely choose to list the questions in this order because we felt that the questions made the most logical sense in this order. Thus, it might be the case that users were just more fatigued by the time they reached Q4, or out of time, or did not feel that they had anything new to add.

3.1 Overall Performance

The overall performance of each run is displayed in Figure 3. The MAP score for our baseline run was 0.2843. As can be seen in the figure, experimental runs out-performed the baseline run and all pseudo relevance feedback runs.

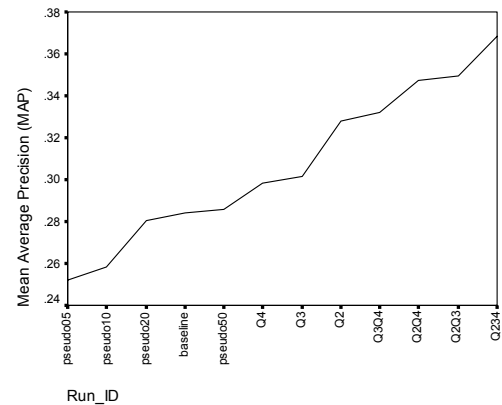


Figure 3. MAP scores

The pseudo relevance feedback runs only outperformed the baseline in one case, when 50 terms were used for query expansion. However, in no case did these runs outperform runs constructed using terms from our clarification form. We did three follow-up pseudo relevance feedback runs using 100, 150 and 200 terms, and found that performance was nearly identical to runs using 50 terms (0.2857, 0.2878, 0.2843). Thus, in this case, performance did not continue to increase with additional terms. These results seem to indicate that the pseudo relevance feedback techniques are not particularly effective in this retrieval situation. This could be because the terms that are used from these documents are not particularly good or discriminating, or because the documents themselves are not particularly good for pseudo relevance feedback. It might be the case that the additional terms that are added for the run using 20 terms for feedback are responsible for the very slight increase in performance over the baseline and the pseudo relevance feedback runs using 05 and 10 terms. Increasing the number of documents used for pseudo relevance feedback might be a more effective way to improve performance in this situation, but we leave this for future work. Overall, these results suggest that the user is a better source of terms for query expansion, rather than retrieved documents.

The relevance feedback runs using known relevant documents (our upper bound runs) performed very well and are not included in Figure 2. As one might expect, all of these runs performed well, with the 50-term run performing the best. The MAP for each run using 05, 10, 20 and 50 terms for query expansion was .4367,

.5284, .5743 and .6129, respectively. When there are known relevant documents, it appears that the system is quite effective at selecting terms for query expansion. Clearly, if our system had retrieved these ten documents first, then the pseudo relevance feedback runs would have performed better. However, if users' initial queries are ambiguous, then this is a very unlikely event. Thus, there still exists a need for a document-independent source of terms for query expansion.

The terms that we elicited from users for query expansion improved retrieval performance in all cases. The worst performing experimental run was the run that used the terms elicited by Q4 (MAP=0.2985) of the clarification form ("Please input any additional keywords that describe your topic"). Recall that this question elicited the fewest number of terms (2.11) from users, for the fewest number of topics (n=20).

Q3 (MAP=0.3018) increased performance only marginally over Q4, even though users responded to this question for a larger number of topics (n=42) than Q4 and the average length of these responses was longer (10.73). Recall that Q3 asked users to indicate why they wanted to know about the topic. It might be the case that this information alone is insufficient in improving retrieval. For instance, one user responded, "I am female and have been studying martial arts since I was 10 years old. I am now practicing Muay Thai boxing, Brazilian Jiu Jitsu and mixed martial arts. I want to know more about other women's experiences in these sports, and also what the best competitions are." This response is somewhat informative in its own right. However, consider another response, "To see the combination of tradition and modernity and locate where that mixing occurs." This response is obviously generated within the context of some other known information (i.e. that provided in the TREC-topic or in response to the preceding clarification form question, Q2), and is ambiguous on its own.

Q2 performed better than any other single question from the clarification form (MAP=0.3279). This question asked users to describe what they already know about the topic, and resulted in the lengthiest responses from users with respect to the clarification form. We are not able to say definitively that *this* question caused the lengthiest and most effective responses from users, since it might be the case that users' response behavior is a result of the location of this question on the clarification form rather than Q2's inherent "goodness." However, the information that users provided in response to this question typically contained background and contextual information, which is important information that is usually not provided by users in real-world information-seeking situations. This, we feel, is strong evidence for the value of this particular question.

Our combination runs were the most effective in this experiment, with the combination of Q2, Q3 and Q4 performing the best. The order in which these questions performed in combination are consistent with how they performed alone: Q4 (0.2985) < Q3 (0.3018) < Q2 (0.3279) and Q3Q4 (0.3325) < Q2Q4 (0.3474) < Q2Q3 (0.3495) < Q234 (0.3685). Although none of these differences were statistically significant, these results do not tell the whole story since they include analysis of all topics, regardless of whether users provided any responses to the particular clarification form question.

3.2 Paired Sample Performance

To further understand our results, we conducted paired samples t-tests between baseline and experimental runs. In these paired tests, we only included topics for which users provided some response to the clarification form question. For instance, the comparison between the baseline run and the Q2 run consists of 40 topics, while the comparison between the baseline and Q4 runs consists of 20 topics. For combination runs, we only included topics for which there were responses to all questions. Results of these paired samples t-tests are displayed in Table 2. For each pair, the table shows the number of topics included in analysis, MAP scores (means and standard deviations), and *p*-values of t-tests (ns=not significant).

Table 2. Results of Paired Samples T-tests

| Pair | N | M (SD) | <i>p</i> -value |
|----------|----|------------------------------|-----------------|
| BL, Q2 | 40 | .3007 (.2547), .3498 (.2575) | .039 |
| BL, Q3 | 42 | .2844 (.2515), .3032 (.2598) | ns |
| BL, Q4 | 20 | .2975 (.2568), .3295 (.2224) | ns |
| BL, Q2Q3 | 37 | .3022 (.2590), .3847 (.2609) | .006 |
| BL, Q2Q4 | 18 | .2975 (.2709), .3912 (.2313) | .009 |
| BL, Q3Q4 | 19 | .3127 (.2544), .3680 (.2316) | .035 |
| BL, Q234 | 17 | .3145 (.2692), .3993 (.2440) | .023 |

For all but two pairs, there were statistically significant improvements in retrieval performance of the experimental runs over the baseline runs. It is difficult to compare these results since the number of topics differs in each case. Although the largest difference in performance between baseline and experimental runs occurred for BL-Q2Q4, this difference did not result in the strongest *p*-value. However, it is clear that all MAP scores are higher in this analysis than the previous analysis and that Q234 was the highest performing run overall. Contrary to the overall results, Q3 rather than Q4 is the lowest performing run which resulted in Q2Q4 performing a little better than Q2Q3.

These results corroborate the general findings in overall performance reported in section 3.1 and provide strong evidence for the effectiveness of runs comprised of a combination of representations of the user's information need, or a polyrepresentation. Taken together, these results suggest that each question is eliciting somewhat different information, which, when used in combination, present the most useful evidence for query expansion. As noted earlier, responses to Q3 often followed from, or were within the context of, responses to Q2. This suggests that probing users with a number of different, but related questions might elicit the most robust and useful problem descriptions. Furthermore, eliciting keywords from users can be helpful, but only in combination with other information. Indeed, responding to questions using natural language might prime users for providing better keywords. In cases where keywords duplicate terms from previous questions, this information can be used for indirect, but user-specified, term weighting.

3.3 Query Length and Performance

The overall performance results seem to suggest a strong relationship between query length and performance for our experimental techniques. Previous research has demonstrated a

positive relationship between query length and performance [10] and we were interested in seeing if this finding held true in our study. A scatter-plot of performance according to query length for our experimental runs is displayed in Figure 4.

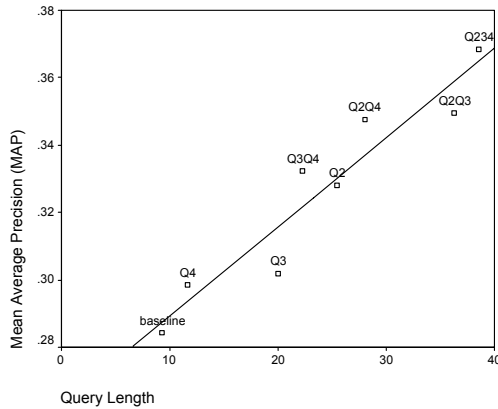


Figure 4. MAP and Query Length for Experimental Runs

From the figure, it is clear that there is a strong linear relationship between query length and performance. A regression analysis provided further evidence for this, $MAP = 0.263 + .000265(\text{query length})$, $p=.000$, $r^2=0.9068$ (adjusted, .891), suggesting that query length is a very good predictor of MAP. This result is quite important since it demonstrates a statistically significant, strong positive relationship between query length and performance even in situations where the query is composed solely of user-generated terms.

4. CONCLUSIONS

We found large differences in the length of users' responses to each of the elicitation questions that we used on our feedback form, and we were excited to see that users were willing to provide such lengthy responses to some of our questions. One of our original motivations for this experiment was to identify a technique that could demonstrate users' potential as sources of terms for query expansion. Our techniques were successful at eliciting information from users, information that, in turn, improved retrieval performance. This research further provides empirical support for Ingwersen's [17] notion of polyrepresentation, and more specifically polyrepresentative extraction of information needs.

Q2 was the most successful single question, both in the amount of information it elicited and increase in performance. While it was the case that Q2 was the first question that users encountered, the background and contextual information that users provided in their responses to this question provide support for its goodness as a question, regardless of order effects. Indeed, the experimental run of users' responses to all three of our feedback form questions out-performed all other runs, which suggests that probing users with a number of different, but related questions might elicit the most robust and useful problem descriptions.

As previous research has demonstrated [25], users often have a difficult time making good expansion decisions, if they are even willing to do so. It can be difficult for users to choose terms since terms are often presented to users out of context. Users may also be reluctant to select terms if they don't understand why they were suggested or from where they came. Further, users may be

censoring themselves based on how they think a retrieval system works and may be overly critical in their determination of which system suggested terms to select. In this study, pseudo relevance feedback runs did not perform particularly well. These results provide some evidence that the user is perhaps a better source of terms for query expansion, rather than retrieved documents and that if properly probed, users will provide a large quantity of quality feedback information.

In this experiment, we also demonstrated a significant relationship between query length and performance, using only user-generated terms. This result corroborates what others have found in batch-mode experiments and suggests that we should be working harder to elicit lengthier queries from users to improve search precision.

Although we do not present a topic level analysis of performance in this paper, we are currently working to see if our performance improvements are due to improvements in many different topics or only a few. Recent work has demonstrated that topic-level analysis is important for understanding differences in retrieval performance [9], and it is likely the case that there are differences in performance across topics in this experiment too. For our future work, we would also like to compute and compare the query clarity score [12] of each initial query with the query clarity after the information from each clarification form question has been added. Examining the query clarity score of each query might allow us to predict which topics are most likely to benefit from a follow-up technique such as the one described in this paper.

We cannot conclude this paper without acknowledging some limitations to this study, which potentially have some impact on the generalizability of the results. This study was carried out as part of the TREC HARD track and because of this, followed an unusual experimental protocol. Two of the more important aspects of this protocol which might have impacted our results are the delay between the time that users defined their topics and the time that they actually completed our clarification form, and the fact that users completed a number of clarification forms in a row. Further, only a small number of users ($n=13$) participated in this experiment. Although the potential impact of the experimental protocol on the validity and reliability of our results cannot be ignored, we feel, nevertheless, that our results are important and make an essential contribution to the research on query expansion and polyrepresentative extraction of information needs.

After reading this paper, one might be left wondering just how 'loquacious' can we expect users to be in a real information-seeking situation? Users should not be discounted as sources of terms for query expansion. Instead, we should work to design clever and creative techniques for encouraging users to be loquacious rather than reticent, both during their initial querying and during follow-up interactions. While it is still the case that we are often "helping people find what they don't already know" [5], we should also strive to help people articulate what they don't know is important.

5. ACKNOWLEDGMENTS

We would like to acknowledge Stephanie Strassel and Meghan Glenn from the Linguistic Data Consortium at the University of Pennsylvania for their assistance with the TREC HARD Track, and James Allan at the University of Massachusetts, Amherst for his efforts organizing and managing the Track.

6. REFERENCES

- [1] Allan, J. (2004). Hard Track overview in TREC 2003 high accuracy retrieval from documents. In E. Voorhees & L. P. Buckland (Eds.), *TREC-2003, Proceedings of the Twelfth Text Retrieval Conference*. Washington, D. C.: Government Printing Office.
- [2] Allan, J. (2005). HARD Track overview in TREC 2004 high accuracy retrieval from documents. In E. M. Voorhees & L. P. Buckland (Eds.), *TREC-2004, Proceedings of the Thirteenth Text Retrieval Conference*. Washington, D.C.: Government Printing Office.
- [3] 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.
- [4] Beaulieu, M. (1997). Experiments on interfaces to support query expansion. *Journal of Documentation*, 53(1), 8-19.
- [5] Belkin, N. J. (2000). Helping people find what they don't know. *Communications of the ACM*, 43(8), 58-61.
- [6] Belkin, N. J., Cool, C., Croft, W. B., & Callan, J. P. (1993). The effect of multiple query representations on information retrieval system performance. In *Proceedings of the 16th Annual ACM International Conference on Research and Development in Information Retrieval (SIGIR '93)*, Pittsburgh PA, USA, 339-346.
- [7] 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.
- [8] 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)*, Toronto, CA, 205-212.
- [9] Buckley, C. (2004). Why current IR engines fail. In *Proceedings of the 27th Annual ACM International Conference on Research and Development in Information Retrieval (SIGIR '04)*, Sheffield, UK, 584-585.
- [10] Buckley, C., Salton, G., & Allan, J. (1994). The effect of adding relevance information in a relevance feedback environment. In *Proceedings of the 17th Annual ACM International Conference on Research and Development in Information Retrieval (SIGIR '94)*, Dublin, UK, 292-300.
- [11] 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.
- [12] Cronen-Townsend, S., Zhou, Y., and Croft, W. B. (2002). Predicting query performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (SIGIR '02)*, Tampere, Finland, 299-306.
- [13] Efthimiadis, E. N. (1996). Query expansion. *Annual Review of Information Science & Technology*, 31.
- [14] Grunfeld, L., Kwok, K. L., Dinstl, N., & Deng, P. (2004). TREC2003 Robust, HARD and QA track experiments using PIRCS. In E. Voorhees & L. P. Buckland (Eds.), *TREC-2003, Proceedings of the Twelfth Text Retrieval Conference*. Washington, D. C.: Government Printing Office.
- [15] 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.
- [16] Ingwersen, P. (1982). Search procedures in the library analyzed from the cognitive point of view. *Journal of Documentation*, 38, 165-191.
- [17] Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation*, 52, 3-50.
- [18] Jansen, B. J., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing and Management*, 36, 207-227.
- [19] Kalgren, J. & Franzen, K. (1997). Verbosity and interface design. Retrieved on 23 May 2005 at <http://www.ling.su.se/staff/franzen/irinterface.html>.
- [20] Katz, W. A. (2002). *Introduction to reference work: reference services and reference processes, volume 2 (8th Edition)*. NY: McGraw-Hill.
- [21] Kekalainen, J. & Jarvelin, K. (1998). The impact of query structure and query expansion on retrieval performance. In *Proceedings of the 21th Annual ACM International Conference on Research and Development in Information Retrieval (SIGIR '98)*, Melbourne, 130-137.
- [22] 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 on Human Factors in Computing Systems (CHI '96)*, Canada, 205-212.
- [23] 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.
- [24] 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.
- [25] 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.
- [26] Voorhees, E. M. (2005). Overview of TREC 2004. In E. M. Voorhees & L. P. Buckland (Eds.), *TREC-2004, Proceedings of the Thirteenth Text Retrieval Conference*. Washington, D.C.: Government Printing Office.