Developing and Evaluating a Query Recommendation Feature to Assist Users with Online Information Seeking and Retrieval

Diane Kelly
University of North Carolina at Chapel Hill
dianek@email.unc.edu

Final Report
2008 OCLC/ALISE Library and Information Science Research Grant Project

22 July 2011

Suggested citation:
Kelly, Diane. 2011. "Developing and Evaluating a Query Recommendation Feature to Assist Users with Online Information Seeking and Retrieval." 2008 OCLC/ALISE research grant report published electronically by OCLC Research. Available online at: http://www.oclc.org/research/grants/reports/2008/kelly2008.pdf

1 Executive Summary

Query formulation is one of the most difficult aspects of information seeking and it is also one of the most important. Research has shown that users have a hard time translating their information needs into queries for multiple reasons, including lack of knowledge of a particular domain. Two techniques, term relevance feedback and query recommendation, provide methods for helping users reformulate their queries, but each is limited in different ways. This research combined these two techniques, and developed and evaluated a query recommendation feature that was not contingent on the existence of a set of previously posed queries, but instead relied on term relevance feedback for query terms. Two major studies were conducted. The goal of the first was to develop and evaluate techniques for creating queries automatically based on term relevance feedback. The goal of the second was to evaluate the effectiveness and usability of a query recommendation feature based on these techniques. To evaluate our approach, we conducted an interactive information retrieval study with 55 subjects and 20 topics. Each subject completed four topics, half with a term suggestion system and half with a query suggestion system. We also investigated the source of the suggestions: approximately half of all subjects were provided with system-generated suggestions, while half were provided with user-generated suggestions. Results show that subjects used more query suggestions than term suggestions and saved more documents with these suggestions, even though there were no significant differences in performance. Subjects preferred the query suggestion system and rated it higher along a number of dimensions including its ability to help them think of new approaches to searching. Qualitative data provided insight into subjects' usage and ratings, and indicated that subjects often used the suggestions even when they did not click on them.

2 Background and Motivation

Query formulation is one of the most difficult and important aspects of information seeking and retrieval. Many information retrieval (IR) techniques have been developed to assist users formulate and reformulate queries, most notably relevance feedback (RF) (Ruthven & Lalmas, 2003). Examples of RF include users adding terms suggested by a system to their queries (term RF) or indicating to a system passages or documents that are relevant to their information needs. However, much of the evidence from interactive studies of RF indicates that such features are not often used. This has been attributed to problems related to the design of RF interfaces (Ruthven, 2003), task complexity and the user's lack of additional cognitive resources (Beaulieu & Jones, 1998), and the amount of extra time and effort required to use such features (Belkin, et al., 2001).

Techniques to assist users with query formulation seem particularly important in cases where users are searching for topics about which they have little knowledge or familiarity. Users may lack the vocabulary and knowledge to sufficiently cover all aspects of the search topic, and in some cases to even formulate an initial search query. Hsieh-Yee (1993) found that when working with less familiar topics, subjects were more likely to consult a thesaurus for term suggestion. Vakkari (2001) found that as subjects learned more about their topics they began to use a wider and more specific search vocabulary. These results suggest query formulation assistance might be particularly useful when users are unfamiliar with topics.

One problem with current term RF 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, especially for users who are less familiar with topics. The display of terms most often consists of a single list. 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. In a previous study, two interfaces that provided term context were compared to a baseline term RF interface which presented a list of terms to users (Kelly & Fu, 2006). Although queries created with each interface significantly outperformed users' initial queries, there were no differences in retrieval performance among interface conditions. Joho, et al. (2002) presented users with a list and a menu hierarchy display for query expansion terms and found that subjects selected more terms from the menu hierarchy, although there were no differences in retrieval performance. Subjects stated that the menu hierarchies gave them a better idea of the contents of retrieved documents, which seems to suggest that contextual displays are more useful. However, it is still unclear if term context is important and if so, what form it might take.

Another problem with traditional RF interfaces is that the suggested terms are generated based on users' initial queries. If these queries are poor, then it is unlikely that the resulting suggested terms will be useful. Many of these techniques are also dependent on the user feeding relevant documents to the system. Again, retrieval of relevant documents is usually dependent on entering a good query. White, et al. (2005) found that when working on complex tasks, users were less effective with an explicit relevance feedback system and preferred a system that used implicit feedback for these reasons.

One way to add context to suggested terms is to combine them with other terms and present them to users as query suggestions rather than term suggestions. Query suggestions provide alternative ways to help users formulate queries and explore unfamiliar areas (c.f., Smyth, et al., 2004; White, et al., 2007). Most of these techniques work by identifying past queries that are similar to the user's current query and suggesting these to the user. Query suggestion is not new. Some early search engines, such as Lycos, suggested alternative queries and terms to users (Beeferman & Berger, 2000), but these techniques were not adopted by many users. Recently, however, there has been a revival of interest in query and term suggestion features by major search engine companies and researchers.

Smyth, et al. (2004) describe the I-SPY search engine, which incorporates a collaborative ranking function based on similar query-document pairs and suggests related queries to users. Results of several evaluations, including ones with subjects, demonstrated that the techniques were effective at improving retrieval. Freyne, et al. (2007) describe the integration of I-SPY with a social navigation component and reported that subjects found query suggestions useful during browsing. White, et al. (2007) compared the effectiveness and usability of a system that suggested queries with one that suggested destinations (or pages). White, et al. studied two types of tasks: exploratory and known-item, and found an interaction effect according to task type. Subjects provided mixed reviews about the suggestion features, but for known-item tasks the query suggestion feature was rated most positively. Furthermore, subjects who rated the query suggestion system more positively indicated they did so because the system saved them from typing queries and helped them generate new ideas for query reformulation. Those who rated it unfavorably stated they felt the suggested queries were not relevant. Overall, the evidence indicates query suggestion can be a useful feature at least for some types of tasks and users.

A necessary condition for query suggestion to occur is that a set of queries that are similar to the current query exists and that the similarity of these queries to one another (and to the current query) can be determined. A number of techniques have been proposed to determine query similarity, including the use of term overlap [18], query hitting time (Mei, Zhou & Church, 2008) and the examination and comparison of results retrieved in response to queries (Billerbeck, et al., 2003; Fitzpatrick & Dent, 1997; Raghavan & Sever, 1995). Query similarity has been used as a method for identifying terms for automatic query expansion (Billerbeck, et al., 2003; Fitzpatrick & Dent, 1997) and query rewriting Jones, et al., 2006). However, all of these techniques need a sufficiently large number of existing query and/or query-document pairs to work, and their effectiveness at generating query suggestions for user consumption is unclear. Moreover, many queries are unique and occur infrequently. Techniques that rely on the existence of previous queries are unlikely to work well in these situations.

3 Purpose

The work proposed here seeks to address the problems outlined above by combining research in these two areas – term RF and query recommendation. To summarize, the major problems related to term RF features are that users are often reluctant to make use of such features, and often have a difficult time selecting good terms from the list of candidates. The major problems related to query recommendation are that systems do not always have available a set of similar queries and determining the similarity of these queries can be difficult.

As a tool for query reformulation, query recommendation may have an advantage over traditional term RF interfaces because of its form and the kind of interaction required by users: users simply click on a query to view new results, rather than selecting one or more terms to add to their queries and re-running retrieval. Overall, the costs involved with selecting a recommended query are less than that of selecting terms to add to one's query. Furthermore, query recommendation provides term context; terms are not presented in isolation but within the context of other query terms. This might help users disambiguate query terms, and more quickly and easily recognize whether a suggestion is relevant to their information needs. It is further proposed that query suggestions constructed from term RF, rather than from previous queries, can be used to address the query sparseness and similarity problems described earlier. Essentially, terms that the system would suggest for term RF would be combined to form queries that could then be used for query recommendation.

The purpose of this research is to evaluate the effectiveness and usability of an automatic technique that combines users' queries with terms generated by RF to create new query suggestions, and to compare differences between term and query suggestion interfaces.

4 Phases of Research

This work had three major phases. The first consisted of identifying techniques to automatically create query suggestions. The research question driving this phase was: How can terms suggested by a traditional term relevance feedback algorithm be used to automatically create queries? The primary activities of this phase consisted of setting-up the IR system and experimenting with different techniques for creating suggestions. The second phase consisted of evaluating the effectiveness and usability of the technique. The research question driving this phase was: How effective and usable is a query recommendation feature based on term suggestion as compared to a traditional term relevance feedback system and to a system that uses human generated suggestions? The primary activities of this phase consisted of designing the study, creating instruments, building the experimental infrastructure, and recruiting and administering the study to subjects. This final phase was data analysis and write-up of the results. The experiments conducted as part of Phase 1 were informal and did not result in formal write-up, so Phase 3 primarily involved the study conducted as part of Phase 2. Table 1 show the phases and the dates in which the work was conducted.

Phase	Task	Conducted
1	Installed Lemur and Indexed Document Collection	January 2008
	Developed Techniques for Constructing Queries	February – March 2008
	Evaluated Techniques (includes data analysis)	March – April 2008
2	Designed and Developed Interfaces	May 2008
	Designed and Developed Instruments	May 2008
	Designed and Developed System Study Infrastructure and Logger	June 2008
	Pilot Testing	July 2008
	Prepared and Submitted Institutional Review Board Application	July 2008
	Recruited Subjects	August 2008
	Ran Study	August – October 2008
3	Analyzed Data	November – January 2008
	Wrote-up Study and Submitted Paper	December 2008- January 2009

5 Method

5.1 Phase 1: Building the Query Suggestion Techniques

The first activities conducted during this phase consisted of installing the Lemur[1] IR toolkit and indexing the documents collection (see Section 5.2.2 for a description of the collection). We used the KL-divergence retrieval model within the Lemur IR toolkit for basic retrieval and built our query and term suggestion features on top of this kernel. After several experiments, we decided to create the query suggestions using a combination of semantic clustering and pseudo-relevance feedback (PRF). In this phase, we also experimented with various parameters and weighting schemes for the retrieval techniques. We decided to use clustering so as to get terms from a more diverse set of documents (rather than just the top 10 or 20 documents).

The query suggestion technique worked as follows. First, when the user issued a query, we retrieved the top 100 results using Lemur's KL-divergence retrieval model (these were also the documents that were displayed to users). We clustered these results using agglomerative clustering and cosine-similarity as a distance metric, using a similarity threshold of 0.75 to maintain a relatively strong inner-cluster document similarity (in our initial testing, this method produced on average between 5 and 10 clusters containing more than one document). We ordered the clusters by size, and picked the 5 largest to generate queries. For each cluster, we

[1] http://www.lemurproject.org/

applied PRF to expand a copy of the user's query, using the documents within the cluster as the feedback documents. We used the default PRF algorithm available in Lemur for the KL divergence method; in essence, the text query is a language model which is updated to incorporate the text of the feedback documents. From this expanded query we drew a set of terms, ordered by descending weight in the query's language model. We kept the top *N* terms from this set, where *N* was equal to the number of terms within the user's original query plus 2. These term sets were then used to populate both query and term suggestions. To create query suggestions, for each cluster, we ordered terms from left to right by descending score and appended these terms to the user's original query; hence, each query was produced by terms which were part of a common cluster and terms from the user's original query. For term suggestion, we used all terms to populate the suggested term list, pruning duplicates and terms which appeared within the original query.

5.2 Phase 2: Evaluating the Query Suggestion Techniques

We designed a user-centered evaluation to evaluate the query suggestion technique. Because term suggestion is the classic method for providing searchers with query formulation assistance, we compare our technique to this method. In addition to this comparison, we compare the queries produced by our automatic technique with those created by humans.

The study design was a 2X2 factorial with one within-subjects variable and one between-subjects variable. The within-subjects variable was *type of suggestion*, with two levels: query suggestion and term suggestion. The presentation order of these types of suggestions was counter-balanced across subjects. The between-subjects variable was *source of suggestions*, with two levels: system-generated or user-generated. In sum, half of the subjects received query and term suggestions produced by human, while half received synthetic suggestions. Each subject used both a query and term suggestion system.

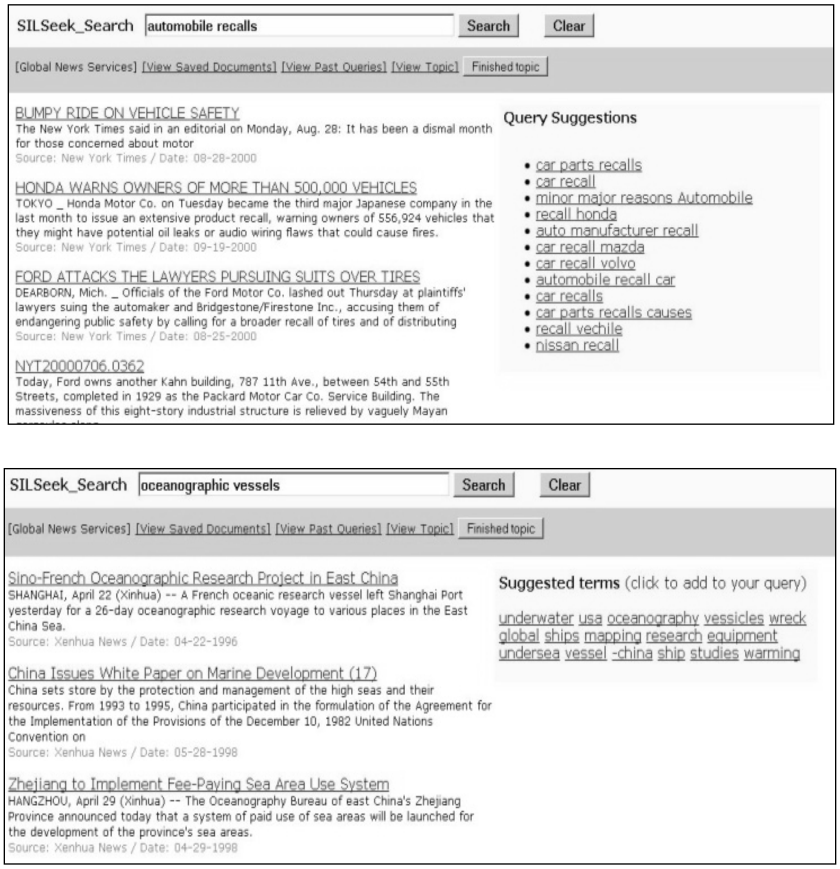
5.2.1 Obtaining User-generated Queries and Terms

To obtain queries and terms generated by other humans and to test our research infrastructure, we conducted a separate study with 35 subjects who participated in a remote evaluation of the automatic techniques. The procedures for the remote experiment were the same as those used in the experiment described in this paper except that subjects completed the study remotely and all subjects received system-generated suggestions. The purposes of the remote study were to generate queries for use in the main evaluation. We do not report results of this study here, but we later compare the data from the remote study with a subset of subjects who participated in the main study (Kelly & Gyllstrom, 2011). The subjects in this study were undergraduates at our university who responded to an email solicitation sent to a university mailing list. Subjects were paid $10.00 USD for their participation which lasted about 1 hour.

To identify which user queries from the remote study we would use to populate our user-generated suggestion system, we divided the 35 subjects' queries (n=669) into three stages: those entered during the first third of the search (early), those entered during the second third (mid) and those entered during the final third (late) and then sampled queries equally from the mid and late groups. The reason for doing this was because we believed that these queries would be more likely to be unique and helpful; at the start of the search, users entered similar queries, but as they progressed through the search, their queries became more unique. We randomly selected queries from each of these bins for each assigned search topic (described in Section 5.2.2). We excluded single-word queries and limited the number of suggestions to 15. To populate the term suggestion interface for the user-generated suggestion condition, we divided the query suggestions into single terms and removed duplicates.

5.2.1 Search Interfaces

The search interfaces for the query and term suggestion systems are displayed in Figure 1. Each of these interfaces allowed subjects to view the documents they had saved, their past queries and the search topic. Clicking on the title of a search result replaced the search results list with the full text of the document (the other features remained in position). From the full text view, subjects could save the document. Clicking on a suggested query replaced the current query with the new query (and produced new results). With the term suggestion interface, subjects could click individual terms to move them to the query box. These terms were appended to the current query unless the subject deleted the query. We did not provide subjects with a tutorial or call these features to their attention as we wanted to avoid biasing their behaviors. While we provided short instruction on the term suggestion interface ('click to add to your query'), we felt that it was reasonable to assume that most subjects would understand that the traditionally styled hyperlinks were clickable.


SILSeek_Search
Search
Clear
[Global News Services] [View Saved Documents] [View Past Queries] [View Topic]
Finished topic

SILSeek_Search: automobile recalls

BUMPY RIDE ON VEHICLE SAFETY
The New York Times said in an editorial on Monday, Aug. 28: It has been a dismal month for those concerned about motor
Source: New York Times / Date: 08-28-2000

HONDA WARNS OWNERS OF MORE THAN 500,000 VEHICLES
TOKYO _ Honda Motor Co. on Tuesday became the third major Japanese company in the last month to issue an extensive product recall, warning owners of 556,924 vehicles that they might have potential oil leaks or audio wiring flaws that could cause fires.
Source: New York Times / Date: 09-19-2000

FORD ATTACKS THE LAWYERS PURSUING SUITS OVER TIRES
DEARBORN, Mich. _ Officials of the Ford Motor Co. lashed out Thursday at plaintiffs' lawyers suing the automaker and Bridgestone/Firestone Inc., accusing them of endangering public safety by calling for a broader recall of tires and of distributing
Source: New York Times / Date: 08-25-2000

NYT20000706.0362
Today, Ford owns another Kahn building, 787 11th Ave., between 54th and 55th Streets, completed in 1929 as the Packard Motor Car Co. Service Building. The massiveness of this eight-story industrial structure is relieved by vaguely Mayan

Query Suggestions

- car parts recalls
- car recall
- minor major reasons Automobile
- recall honda
- auto manufacturer recall
- car recall mazda
- car recall volvo
- automobile recall car
- car recalls
- car parts recalls causes
- recall vechile
- nissan recall

SILSeek_Search: oceanographic vessels

Sino-French Oceanographic Research Project in East China
SHANGHAI, April 22 (Xinhua) -- A French oceanic research vessel left Shanghai Port yesterday for a 26-day oceanographic research voyage to various places in the East China Sea.
Source: Xenhua News / Date: 04-22-1996

China Issues White Paper on Marine Development (17)
China sets store by the protection and management of the high seas and their resources. From 1993 to 1995, China participated in the formulation of the Agreement for the Implementation of the Provisions of the December 10, 1982 United Nations Convention on
Source: Xenhua News / Date: 05-28-1998

Zhejiang to Implement Fee-Paying Sea Area Use System
HANGZHOU, April 29 (Xinhua) -- The Oceanography Bureau of east China's Zhejiang Province announced today that a system of paid use of sea areas will be launched for the development of the province's sea areas.
Source: Xenhua News / Date: 04-29-1998

Suggested terms (click to add to your query)

underwater usa oceanography vessicles wreck global ships mapping research equipment undersea vessel -china ship studies warming

Figure 1. Interfaces for Query and Term Suggestion Systems

5.2.2 Corpus and Topics

The TREC Robust Track collection (Voorhees, 2006) was used in this study. This collection contains the 3GB AQUAINT corpus of newswire text in English, 50 topics and a set of relevance judgments. Over one million documents are included in the corpus from Xinhua News (1996-2000), New York Times News (1998-2000), and Associated Press Worldstream News (1998-2000).

Twenty topics were used in this study. We were interested in selecting topics of varying levels of difficulty to investigate the relationship between topic difficulty and use of suggestions. To select topics, we examined the performances of queries written by subjects from another study that used all 50 topics from this collection (424 queries contributed by 61 subjects) (Kelly & Fu, 2006). We computed normalized discounted cumulated gain (nDCG) (Järvelin & Kekäläinen, 2002) at depth 50 for each query, averaged these values for each topic and sorted topics into four quartiles according to nDCG: easy, medium, moderate and difficult. We then manually selected five topics from each bin by considering the number of relevant documents in the corpus for each topic and whether we thought the topic would be of interest to our target subjects (undergraduate students). Each subject completed one topic from each difficulty bin (4 total). The basic objective of the search task was to find documents relevant to the information described by the topic. Subjects were limited to 15 minutes per topic. Topics were rotated and counter-balanced across subjects and systems.

5.2.3 Use of Suggestions

Subjects' interactions with the system were logged. The log was examined to identify use of suggestions, or the number of clicks users made on suggested queries and terms.

5.2.4 Performance

Two performance measures were computed using documents subjects' saved: number saved and session-based discounted cumulated gain (sDCG) (Järvelin, Price & Delcambre, 2002). To compute sDCG, first DCG is computed for each query (Järvelin & Kekäläinen, 2002). sDCG is a session-based measure that accommodates interactive search situations where a user enters multiple queries during a single session (where a session is considered in this case a temporal instance of a user searching for information about a single topic). sDCG includes a discount function for each query issued by a user that is based on the sequence number of the query within the session. Essentially, the total value of search results returned by queries issued later in a search session is worth less than those returned by queries issued earlier in the session. sDCG for a query is defined as:

$$\text{sDCG}(q) = (1 + \log_{bq} q)^{-1} * \text{DCG}$$

where *bq* is the logarithm base for the query discount, *q* is the position of the query in the session and DCG is the vector for the search results returned by the query. We normalized DCG scores, so the resulting sDCG also represents a normalized value (snDCG). Järvelin, et al. (2002) describe *bq* as a bounded parameter (1 < *bq* < 1000) which can be set to model varying types of users (e.g., *bq*=2 for an impatient user and *bq*=10 for a patient user). We set *bq* to 10 since our users were experimental subjects who were tasked with evaluating the system. It is also necessary to set two parameters for the nDCG vector: *b* which is the logarithm base for the result discount and *d* which is the depth of the search results examined. We set *b* = 2 and *d* = 100. Once we computed snDCG for each query in a session, we averaged these values to arrive at a composite measure for each session, which represents a subject's performance on a particular topic.

5.2.5 Perceptions & Preferences

Subjects evaluated the topics and the systems according to effectiveness, satisfaction and preference using an Exit Questionnaire. Results regarding the topic evaluations are described in Bailey, Kelly and Gyllstrom (2009). The main part of the questionnaire presented the systems together (query and term suggestion) and asked subjects to rate the two along several items (see Table 5 for items). We presented the systems together to facilitate relative comparisons and the presentation order of the systems was based on how subjects experienced them. Subjects used a 5-point scale to respond to all items, where 1=strongly disagree and 5=strongly agree. This questionnaire included 7 items that assessed the effectiveness of the suggestions, 3 items that assessed the effectiveness of the system, and 1 general satisfaction item. Subjects were asked three preference items, each of which was followed with open-ended questions about their preferences.

5.2.6 Stimulated Recall

A portion of subjects completed stimulated recall using two of their searches so that we could better understand their decision-making and perceptions of the suggestion features. In total, 24 subjects completed stimulated recall. While these subjects searched, their second and fourth searches were recorded using screen-recording software. We selected the second and fourth searches because we wanted to sample one search for each system (term suggestion and query suggestion). We believed the second searches subjects completed with each system would be the most illustrious since subjects had a chance to acclimate to the systems and the experiment. During the stimulated recall, the recordings were played back to subjects and they were asked to describe their thoughts, actions and decision-making.

5.2.7 Procedures

Study sessions took place in a private laboratory, during individual sessions. Upon arrival to the lab, subjects logged in to the system and completed the consent form, demographic questionnaire and search experience questionnaire. No tutorial was given to subjects. Before subjects started searching, they were shown a pre-topic questionnaire, which displayed the topic and asked them several questions about their knowledge of, and experience with, the topic. Subjects had up to 15 minutes to find and save documents relevant to the information described in the topic. After completing four search topics, subjects completed the Exit Questionnaire. Most subjects completed their participation at this time, except for those who participated in the stimulated recall, which was done following the Exit Questionnaire. Sessions lasted 1-1.5 hours.

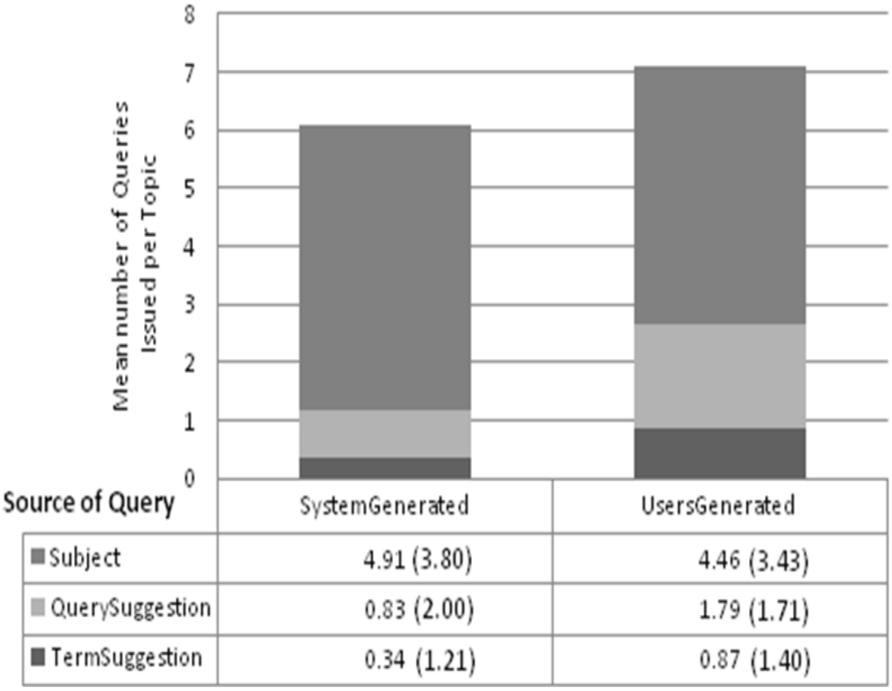
5.2.8 Subjects

Subjects were recruited via email solicitation to the undergraduate mailing list at our university. Fifty-five subjects participated in this study (33 females and 22 males). Subjects' mean age was 21 years (SD=3.5). Twenty percent of the subjects were humanities majors, 33% were social science majors, 18% were science majors, 24% were in a professional school and 5% were undecided. Subjects rated their skills and abilities to locate information on the Web as slightly above average (*M*=5.24, *SD*=.98) (1=poor; 7=outstanding). Subjects were given a $20 USD honorarium. Subjects were assigned randomly to condition.

6 Findings & Discussion

6.1 Use of Suggestions

Figure 2 displays the average number of queries submitted by subjects for each topic according to whether they received system-generated suggestions or user-generated suggestions. This figure displays the average number of queries that were generated by subjects, the average number of query suggestions taken by subjects and the average number of queries that were modified or created with the term suggestions. On average, subjects who received user-generated suggestions submitted around 7 queries per topic, while those who received system-generated suggestions submitted around 6 per topic. Overall, the majority of queries originated with the subject. Subjects who received user-generated suggestions used significantly more suggestions than those who received system-generated suggestions [*F*(1, 218)=11.60, *p*<.01]. Of these suggestions, subjects used significantly more of the query suggestions than the term suggestions [*F*(1, 218)=10.45, *p*<.01]. There were no significant interaction effects between suggestion type and source of terms.



Source of Query	SystemGenerated	UsersGenerated
■ Subject	4.91 (3.80)	4.46 (3.43)
■ QuerySuggestion	0.83 (2.00)	1.79 (1.71)
■ TermSuggestion	0.34 (1.21)	0.87 (1.40)

Figure 2. Mean (standard deviation) number of queries submitted.

Table 2. Mean number of queries (standard deviation) submitted by subjects according to topic difficulty.

	Topic Difficulty			
	Easy	**Medium**	**Moderate**	**Difficult**
Subject Generated	4.13 (2.88)	3.80 (2.48)	4.87 (3.68)	6.00 (4.73)
System Assisted	0.73 (1.21)	0.76 (1.12)	1.04 (2.02)	1.22 (2.12)
Term Suggestion	0.29 (0.83)	0.24 (0.64)	0.18 (0.51)	0.47 (1.57)
Query Suggestion	0.44 (1.01)	0.53 (1.05)	0.85 (2.03)	0.75 (1.66)
Total Queries Issued	4.85 (3.46)	4.56 (2.71)	5.91 (4.48)	7.22 (5.47)

Table 2 displays the mean number of queries submitted according to topic difficulty. This includes the mean number of queries that originated with subjects and the mean number that were created with the help of a suggestion. This latter category is further broken down into mean number of query suggestions taken and mean number of queries created using term suggestions. Overall, subjects issued the most queries for hard topics and the second most for moderate topics. Subjects entered slightly more queries for easy topics than medium topics. One-way ANOVA showed that there was a significant difference in these means [*F*(3, 219)=4.60, *p*<.01]. Bonferroni post-hoc tests showed that the difference was between hard and easy topics and hard and medium topics. Results also showed that as topics became more difficult, subjects used more suggestions. However, the differences in these means was not significant [*F*(2, 219)=1.06, *p*=.368]. Of the two, subjects used more query suggestions than term suggestions when issuing new queries, and in general, the number of each of these increased as topic difficulty increased. However, mean number of query and term suggestions did not differ significantly according to topic difficulty [Query Suggestions: *F*(3, 219)=.906, *p*=.439; Term Suggestions: *F*(3, 219)=.915, *p*=.435].

6.2 Performance

Table 3 displays the mean number of documents subjects saved per topic for different kinds of queries. On average, subjects saved close to 9 documents per topic with the queries they created themselves. Subjects who received user-generated suggestions saved significantly more documents with system assisted queries than subjects who received system-generated suggestions [*F*(1, 218)=14.10, *p*<.01]. Subjects also saved significantly more documents with query suggestions than with queries created with term suggestions [*F*(1, 218)=3.19, *p*<.05]. There were no significant interaction effects between suggestion type and source of terms, and no significant difference in the total number saved in the system-generated and user-generated suggestion systems.

Table 3. Mean (standard deviation) number of documents saved per query type (SGS=system-generated suggestions; UGS=user-generated suggestions; TS=term suggestion; QS=query suggestion).

	SGS	UGS	Total
Subject Generated Queries	8.97 (6.60)	8.79 (7.82)	8.88 (7.81)
System Assisted Queries	0.41 (1.27)	1.48 (2.80)	0.92 (2.19)
TS Queries	0.34 (1.18)	0.90 (2.70)	0.61 (2.05)
QS Queries	0.48 (1.35)	2.06 (2.81)	1.23 (2.29)
Total Documents Saved	9.26 (6.53)	10.45 (7.72)	9.80 (7.12)

Table 4 displays the mean snDCG scores according to source of suggestions and suggestion type. Overall, subjects who received user-generated query suggestions performed the best. Subjects who received user-generated suggestions outperformed those who received system-generated suggestions although this difference was not significant [*F*(3, 217)=3.266, *p*=.072]. Although the differences were not significant, there seems to be an interaction effect: subjects who received system-generated suggestions performed slightly better with term suggestions, while those who received user-generated suggestions performed slightly better with query suggestions. Overall, there was little difference between subjects' performances in the term and query conditions.

Table 4. Mean snDCG of queries issued by subjects for each topic (where topic=session).

Source	Suggestion Type	Mean	Stand. Dev.
System Generated Suggestions	Term	.379	.170
	Query	.367	.157
	Total	.373	.163
User Generated Suggestions	Term	.409	.157
	Query	.418	.169
	Total	.413	.162
Total	Term	.393	.164
	Query	.390	.164
	Total	.391	.164

6.3 Perceptions

Subjects' mean responses to the perceived effectiveness and usability items from the Exit Questionnaire are presented in Table 5. Table 6 shows results of ANOVA tests of the differences between these means. Overall, subjects rated the query suggestion system higher than the term suggestion system on 7 of the 11 items. Two of the items concerned the suggestions' abilities to help the subject think of new approaches to searching and better understand the topics. For both of these items, query suggestions were rated significantly higher than term suggestions (see Table 6, Items 3 and 4). Subjects' ratings also indicated that they found it significantly easier to find relevant documents with the query suggestion system and that overall, the query suggestion system was significantly more effective in helping them complete the search tasks (Table 6, Items 8 and 10). Although the log data showed that subjects did not use query suggestions frequently, they felt that the query suggestion system helped them in a variety of ways, some of which are not detectable from the log.

For four items, subjects rated the term suggestion system higher than the query suggestion system, although this difference was only significant for one item. Two of these items were related to the execution of the suggestions: one asked subjects to rate how helpful the suggestions were in modifying their queries and another asked about how easy the suggestions were to use. The responses to the first item might be an artifact of the item itself and how subjects' perceived of the suggestions – the terms allowed subjects to modify their current queries, while the queries replaced their original queries. Although both actions could be considered query modification, it seems that subjects distinguished between modifying and issuing a new query. The latter item perhaps gets at the idea that the terms can be used independently, while the queries need to be used as unit. Subjects' responses to another of these items where term suggestion was rated higher than query suggestion were also unexpected – despite query suggestions getting higher ratings on most items, subjects indicated that it was easier to understand how the suggested terms related to the search topics.

When comparing subjects' responses who received system-generated suggestions with user-generated suggestions, there was an equal split in ratings: for five of the items, the system which displayed system-generated suggestions was rated higher, for five of the items the situation was reverse, and for one item the scores were nearly identical. Most notably, the quality of the suggestions was rated significantly higher by those receiving system-generated suggestion (Table 6, Item 7) and these suggestions were also rated as easier to use and of greater help with query modification (Table 6, Items 5 and 2). The items that were rated significantly higher by those who received user-generated suggestions were primarily about the suggestions' abilities to stimulate their thinking about their searches and the topics (Table 6, Items 3 and 4). These results are encouraging, since they show that subjects did not perceive large, consistent differences in the suggestions generated by our automatic techniques and those generated by humans. Finally, there were no significant interaction effects between system type and source of suggestions.

Table 4. Mean (standard deviation) for Exit Questionnaire items according to condition (TS=term suggestions; QS=query suggestions; SGS=system-generated suggestions; UGS=user-generated suggestions) (higher values in bold).

Questionnaire Item	**TS**	**QS**	**SGS**	**UGS**
1. The suggestions made things I wanted to accomplish easier to do.	3.24 (1.26)	**3.58** **(1.20)**	3.20 (1.26)	**3.66** **(1.17)**
2. The suggestions helped me modify my queries.	**2.75** **(1.22)**	2.67 (1.23)	**3.17** **(1.14)**	2.16 (1.10)
3. The suggestions helped me think of new approaches to searching.	2.49 (1.20)	**3.24** **(1.36)**	2.42 (1.29)	**3.40** **(1.18)**
4. The suggestions helped me better understand the topics.	2.65 (1.24)	**3.29** **(1.29)**	2.72 (1.32)	**3.28** **(1.21)**
5. The suggestions were easy to use.	**3.02** **(0.95)**	2.75 (1.21)	**3.23** **(1.08)**	2.46 (0.95)
6. It was easy to understand how the suggestions related to the search topics.	**3.40** **(1.13)**	2.75 (1.19)	**3.08** **(1.17)**	3.06 (1.25)
7. The quality of the suggestions was good.	**3.51** **(1.05)**	3.27 (1.18)	**3.63** **(1.09)**	3.10 (1.09)
8. It was easy to find relevant documents with the system.	2.91 (1.08)	**3.51** **(1.01)**	3.13 (1.07)	**3.30** **(1.11)**
9. It was easy to understand why some documents were retrieved in response to my queries.	2.44 (0.88)	**2.76** **(1.02)**	2.55 (0.93)	**2.66** **(1.00)**
10. Overall, the system was effective in helping me complete the search tasks.	2.91 (1.04)	**3.31** **(1.25)**	**3.15** **(1.21)**	3.06 (1.11)
11. Overall, I was satisfied with my performance.	3.13 (1.04)	**3.44** **(0.99)**	**3.43** **(1.05)**	3.10 (0.97)

Table 5. Results of ANOVA tests of the differences in means for items reported in Table 4.

Item	**Type (Terms or Queries)**	**Source (System or Users)**
1	*F*(1,108) = 2.29, *p*=.133, ns	*F*(1,108) = 3.88, *p*<.05, **U>S**
2	*F*(1,108) = .115, *p*=.735, ns	*F*(1,108) = 21.72, *p*<.01, **S>U**
3	*F*(1,108) = 10.75, *p*<.01, **Q>T**	*F*(1,108) = 18.48, *p*<.01, **U>S**
4	*F*(1,108) = 7.34, *p*<.01, **Q>T**	*F*(1,108) = 5.63, *p*<.05, **U>S**
5	*F*(1,108) = 2.04, *p*=.156, ns	*F*(1,108) = 15.59, *p*<.01, **S>U**
6	*F*(1,108) = 9.45, *p*<.01, **T>Q**	*F*(1,108) = .011, *p*=.917, ns
7	*F*(1,108) = 1.31, *p*=.254, ns	*F*(1,108) = 6.48, *p*<.05, **S>U**
8	*F*(1,108) = 10.18, *p*<.01, **Q>T**	*F*(1,108) = .705, *p*=.403, ns
9	*F*(1,108) = 3.38, *p*=.069, ns	*F*(1,108) = .361, *p*=.549, ns
10	*F*(1,108) = 3.84, *p*<.05, **Q>T**	*F*(1,108) = .168, *p*=.682, ns
11	*F*(1,108) = 3.17, *p*=.078, ns	*F*(1,108) = 3.05, *p*=.084, ns

6.4 Preferences

Subjects were asked three preference questions, each followed by an open-ended question to elicit explanations for their choices. Table 7 displays these questions and subjects' preferences. Overall, subjects found the suggested queries more useful than the suggested terms regardless of whether the suggestions were system or user generated. However, over twice as many subjects who received system-generated suggestions indicated that both types of suggestions were equally useless. Fewer subjects who received user-generated suggestions selected this option. Very few subjects said the suggestions were equally useful. With respect to ease of operating and integration into searching, again, more subjects picked query suggestions over term suggestions and this was relatively consistent regardless of whether they received system-generated or user-generated suggestions. These results are somewhat inconsistent with subjects' numeric responses to Item 5, where term suggestion was rated higher (these differences were not statistically significant). Overall, more subjects preferred query suggestion, although the preference difference was most pronounced for subjects who received user-generated suggestions. Subjects were nearly equally likely to pick term, query or no preference in the system-generated condition. Despite some differences in subjects' responses to these three questions, Chi-square tests showed no significant differences in the distributions.

The general message seems to be that subjects prefer query suggestions and that those who received system-generated suggestions did not always prefer one type of suggestion over the other. Follow-up responses provided insight into subjects' preferences. Generally, subjects' comments who received system-generated suggestions did not differ much from those who received user-generated suggestions. However, there was more mention of random suggestions in the system-generated group. Several subjects mentioned that they preferred their own ideas to suggestions for either system. Those who preferred the query suggestions liked the all in one approach and the ability to click once to get results. They also liked the specificity and focus of the suggestions, and commented that the queries presented whole ideas. A subject commented, "Query suggestion was faster, easier and provided a better understanding of how I should perform my search. There may be a slight bias because I'm not used to term suggestion but I found it a bit uncomfortable to use."

Many subjects saw the terms as being jumbled together or taking too much effort to execute. Those who preferred the term suggestions thought that the terms provided more flexibility and that the suggested queries were too similar. They liked being able to refine their existing queries instead of starting new ones. Most subjects seemed to view the terms as a way to modify their existing queries, although it was possible to create new queries with the terms. A subject stated, "With the terms, if you used them, it added them on to your already listed search which sometimes caused the search results not to make sense." Although the presentation of terms was standard, these comments show that there is more work to be done on integrating term suggestion into search routines and communicating their functionality to subjects. However, we note that our instructions on the term suggestion interface ("Click to add to your query") may have misled subjects to believe that terms could only be added to an existing query rather than used to create a new query.

Table 7. Distribution of subjects' preferences.

		Source of Suggestions		
		System	Users	Total
Which were more useful to you while you searched?	Suggested Terms	6	4	10
	Suggested Queries	11	13	24
	Equally Useful	1	2	3
	Equally Useless	10	4	14
	Total	28	23	51
Which were easier to operate and integrate into your searching?	Suggested Terms	12	7	19
	Suggested Queries	16	15	31
	Total	28	22	50
Overall, which of the systems did you prefer?	Term Suggestion	9	5	14
	Query Suggestion	11	13	24
	No Preference	8	3	11
	Total	28	21	49

6.5 Stimulated Recall

Data from the stimulated recall was similar to the qualitative data elicited via the open-ended items. Some subjects using the system-generated term suggestions complained that suggested terms were "random terms that had nothing to do with what I was trying to find," and "inapplicable." Users of the user-generated term suggestion system called the terms "elementary" and wanted a "more specific vocabulary for the topic." One subject commented that the suggested terms were really broad and the queries had better keywords. Other subjects liked adding terms to their queries and modifying their own queries.

Subjects made several comments about the query suggestions. Some did not use query suggestions because they did not feel that they needed the help. Others wanted to use only part of the query suggestion. Others stated that even though they did not click on the suggested queries, the queries gave them ideas for manually changing the keywords in their queries. One subject commented that the query suggestions often anticipated what he was thinking and another subject stated that he used the query suggestions because he did not know how to start his search. One subject commented that he liked the query suggestion because it “gave you what you needed right there” and “I don’t think I even typed anything.”

Overall, subjects’ comments centered on several themes: control, quality and assistance. Subjects who preferred the term suggestions liked the increased control it gave them over query modification. Those who preferred the query suggestions liked the ease with which they could execute a new query. Subjects also made many comments about the quality of the suggestions. Those who received user-generated suggestions commented that they did not always see how the term suggestions related to their searches and wanted a more specific vocabulary. These suggested terms were from queries other users had created and by themselves, the terms may not have been particularly useful. Those commenting about query suggestions indicated noticing some redundancy, perhaps because in the system-generated condition new terms were appended to subjects’ queries. Finally, many subjects who preferred query suggestions commented on how the queries helped them get started and think of other avenues for searching. This was also reflected in their response to the closed items in the Exit Questionnaire.

Query suggestions seem to have an advantage when subjects face a cold-start problem and when they exhaust their own ideas for searches. Term suggestion seems more useful when increased control over query modification is desired. The results suggest that a hybrid presentation of queries and terms – where subjects can execute the whole query or select individual query terms – might provide more useful and flexible assistance. The query suggestions helped subjects better understand the terms and topics, generate new ideas for searching, and engage in more effortless searching. If terms within these queries could be manipulated, this would give subjects greater control over query formulation and allow them to be more discriminating and use their judgment to guide the assistance.

7 Summary and Conclusion

We proposed and evaluated a feature which used pseudo-relevance feedback and clustering to generate query suggestions. We compared the query suggestions to term suggestions and examined differences between automatically generated suggestions and those generated by humans in a previous experiment. We found that subjects did not use the suggestion features that often, but when they did, more query suggestions were used. We also found that subjects submitted more queries and used more suggestions for difficult topics. Subjects who received user-generated suggestions saved more documents found through suggestions; the most were saved for query suggestions. snDCG showed that the best performance was achieved by those who received user-generated query suggestions and that there appeared to be an interaction with source of suggestions and suggestion type: those who received user-generated suggestions did better with query suggestions, while those with system-generated terms did better with term suggestions. Query suggestions were rated higher for the majority of the perceived effectiveness and satisfaction items and more subjects preferred them to the term suggestions. There was no clear preference for user-generated suggestions over system-generated system which was encouraging since these results show that our automatic techniques were perceived to be at least as good as user-generated suggestions.

8 Main Publication and Presentations

[1] Kelly, D., Gyllstrom, K., & Bailey, E. W. (2009). A comparison of term and query suggestion features for interactive searching. *Proceedings of the 32th Annual ACM International Conference on Research and Development in Information Retrieval (SIGIR '09)*, Boston, Massachusetts, 371-378.

Additional Presentations of Work

Developing and Evaluating a Query Recommendation Feature to Assist Users with Online Information Seeking and Retrieval, Association for Library and Information Science Education Conference (ALISE), January 23, 2009.

Providing Assistance for Query Formulation and Idea Generation in Interactive Search, Department of Library and Information Science International Symposium on Recent Trends in Library and Information Science Research, National Taiwan University, Taipei, Taiwan, September 10-11, 2009.

9 Enabled Research and Publications

Several follow-up projects based on the initial research and/or using the infrastructure developed in the initial research were conducted. The term *initial* is used to refer to the project that was conducted as part of the OCLC Grant. Because this report is so late and we have been able to use the data, infrastructure and results of the OCLC project in many additional projects, this distinction is made to differentiate the project *directly* supported by the OCLC grant from those that were *indirectly* supported. All projects are included in this report to demonstrate how much this grant has benefited the PI, students and query suggestion research. The titles and abstracts of these follow-up projects/papers are presented below along with a description of how they extended and/or were enabled by the initial research. All publications with the exception of the Master's paper were referred publications.

[2] Kelly, D., Cushing, A., Dostert, M., Niu, X., & Gyllstrom, K. (2010). Effects of popularity and quality on the usage of query suggestions during information search. *Proceeding of the 28th ACM Conference on Human Factors in Computing Systems (CHI '10)*, Atlanta, GA, 45-54.

Abstract: Many search systems provide users with recommended queries during online information seeking. Although usage statistics are often used to recommend queries, this information is usually not displayed to the user. In this study, we investigate how the presentation of this information impacts use of query suggestions. Twenty-three subjects used an experimental search system to find documents about four topics. Eight query suggestions were provided for each topic: four were high quality queries and four were low quality queries. Fake usage information indicating how many other people used the queries was also provided. For half the queries this information was high and for the other half this information was low. Results showed that subjects could distinguish between high and low quality queries and were not influenced by the usage information. Qualitative data revealed that subjects felt favorable about the suggestions, but the usage information was less important for the search task used in this study.

Relation to Initial Work: We used the basic experimental infrastructure developed for the initial study in this study. This study was a follow-up to the initial study in that we examined (1) if subjects could distinguish between high and low quality query suggestions and (2) if subjects' selections were influenced by usage information.

[3] Kelly, D. & Gyllstrom, K. (2011). An examination of two delivery modes for interactive search system experiments: Remote and laboratory. *Proceedings of the 29th ACM Conference on Human Factors in Computing Systems (CHI '11)*, Vancouver, Canada.

Abstract: We compare two delivery modes for interactive search system (ISS) experiments: remote and laboratory. Our study was completed by two groups of subjects from the same population. The first group completed the study remotely and the second group completed the study in the laboratory. We compare differences in participants, participation behaviors, search behaviors and evaluation behaviors. Overall, for most measures no significant differences were found, but there were some notable differences. Greater variance was observed in time taken and number of documents opened and saved by remote subjects. Lab subjects provided more favorable responses to exit questionnaire items and reported significantly higher satisfaction. Lab subjects also provided significantly longer responses to open questions, while remote subjects provided more null responses. These results suggest that many behaviors do not change significantly according to study mode and that results from remote ISS experiments are similar to those from laboratory experiments.

Relation to Initial Work: We used data collected in the initial study from the 35 subjects in the study which was conducted to obtain queries and terms to populate the user-generated suggestion system. We compared data generated by these subjects with a subset of subjects from the initial study (the subjects who used the system-generated suggestions) to see if and how study mode (remote vs. laboratory) impacted search and evaluation behaviors.

[4] Niu, X. & Kelly, D. (under review). The use of query suggestions as idea tactics during information search. Submitted to the *Journal of the American Society for Information Science and Technology*.

Abstract: Query suggestion is a common feature of many information search systems. While much research has been conducted about how to generate suggestions, fewer studies have been conducted about how people interact with and use suggestions. The purpose of this paper is to investigate how and when people integrate query suggestions into their searches and the outcome of this usage. The paper further investigates the relationships between search expertise, topic difficulty and task stage and query suggestion usage. A controlled laboratory study was conducted with 23 subjects who used an experimental search system with query suggestions to conduct four searches. Results showed that subjects integrated the suggestions into their searching fairly quickly and that subjects with less search expertise used more suggestions and saved more documents. Subjects also used more suggestions during the latter stages of search and when searching for more difficult topics. These results show that query suggestion can provide support in situations where people have less search expertise, greater difficulty searching and at specific stages during the search.

Relation to Initial Work: This work was a secondary analysis of data collected in [2] above. In both the initial study [1] and study [2], subjects told us during interviews and stimulated recall that the query suggestions gave them ideas for searching and helped them get started with their searches. They further stated that the suggestions helped them think of other avenues for searching when they exhausted their own ideas and when they were uncertain about how to proceed. This study was conducted to better understand when and how these subjects used query suggestions. The study also examined contextual factors (search experience, topic difficulty and search stage) that might impact usage of query suggestions.

[5] Bailey, E. W., Kelly, D., & Gyllstrom, K. (2009). Undergraduates' evaluations of assigned search topics. *Proceedings of the 32th Annual ACM International Conference on Research and Development in Information Retrieval (SIGIR '09)*, Boston, Massachusetts, 812-813. [Poster]

Abstract: This paper evaluates undergraduate students’ knowledge, interests and experiences with 20 topics from the TREC Robust Track collection. The goal is to characterize these topics along several dimensions to help researchers make more informed decisions about which topics are most appropriate to use in experimental IIR evaluations with undergraduate student subjects.

Relation to Initial Work: This work is a secondary analysis of the data collected during the initial study. In the initial study, subjects evaluated the search topics along several dimensions including interest, personal relevance, familiarity and difficulty. The major objective of this work was to provide a characterization of these topics so that researchers could have more information when trying to select which topics to use from the collection in their studies.

[6] Mack, Adam (2011). *The Power of Suggestion: An Evaluation of the Positive Effects of Source and Position on the Use of Query Suggestions*. Completed in partial fulfillment of the MS in Information Science, School of Information and Library Science, University of North Carolina, Chapel Hill, NC.

Abstract: This study examines whether the perceived source of query suggestions and their location within the search interface affects users’ likelihood of clicking on the suggested queries. Participants were asked to complete three search tasks using the search system provided. The query suggestions were placed on either left or right side of the page, and were labeled as either system-generated or as having come from other users of the system. Participants also evaluated their engagement with the search tasks and the quality and usefulness of the query suggestions. Results indicated that users presented with query suggestions on the left scored significantly higher on two measures of engagement. While the effects of source did not meet tests of statistical significance, participants who believed the query suggestions came from other users had higher mean scores for three of the four search engagement scales and the query suggestion rating scale.

Relation to Initial Work: This study used the research infrastructure developed during the initial study as well as the queries generated by subjects. It extended the initial study, by investigating whether the location of the query suggestions impacted use and whether the explicitly stated source of the suggestion (computer or human) impacted use.

10 Summary of Student Involvement

Six students have benefited from this project. Karl Gyllstrom, UNC Computer Science Ph.D. student, and Earl Bailey, UNC SILS Ph.D. student, were involved with the initial study. Karl was paid with the grant funds and Earl was funded through the SILS. Karl was involved with all aspects of the project and in particular, developing the system, query suggestion features and other research infrastructure. Karl was involved with the entire life-cycle of the project (January 2008-December 2009). Earl joined the team in Fall 2008 and primarily helped with administering the study to subjects and data analysis.

The infrastructure developed in this project was later used in two other studies. One study was a collaborative effort between the PI and three students (two SILS Ph.D. and one Master's student) [2] and another was a Master's student project which the PI advised [6]. Although no longer being paid and having finished his Ph.D., Karl Gyllstrom continued to be involved with all of these projects primarily through the provision of technical assistance.

11 Future Research

In the last year, we have completed two new studies exploring query suggestions. The first was a collaboration between the PI and a usability engineer at MITRE. In this study, we examined query suggestions in the context of intranet search and also added a star-based recommendation feature to the interface so that subjects could rate the different query suggestions. We also asked professional searchers to create the query suggestions we provided to subjects and investigated if knowledge of the source of the queries impacted their usage (in other words would subjects be more likely to take a suggestion made by one of their own information professionals). The second study was led by a Ph.D. student at SILS. In this study, we further investigated the idea of allowing people to recommend queries by associating star-ratings with them. The basic goal was to better understand how people would approach the task of rating queries and how result quality related to the ratings. Data from both of these studies is currently being analyzed.

12 References

Beaulieu, M. & Jones, S. (1998). Interactive searching and interface issues in the Okapi best match probabilistic retrieval system. *Interacting with Computers, 10*, 237-248.

Beerferman, D., & Berger, A. (2000). Agglomerative clustering of a search engine query log. *Proceedings ACM SIGKDD*, Boston, MA, 407-416.

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.

Billerbeck, B., Scholer, F., Williams, H. E., & Zobel, J. (2003). Query expansion using associated queries. *Proceedings CIKM '03*, New Orleans, LA, 2-9.

Fitzpatrick, L., & Dent, M. (1997). Automatic feedback using past queries: Social searching? *Proceedings of SIGIR '97*, Philadelphia, PA, 306-313.

Freyne, J., Farzan, R., Brusilovsky, P., Smyth, B., & Coyle, M. (2007). Collecting community wisdom: Integrating social search & social navigation. *Proceedings of IUI '07*, Honolulu, HI, 52-61.

Hsieh-Yee, I. (1993).Effects of search experience and subject knowledge on online search behavior: Measuring the search tactics of novice and experienced searchers. *Journal of the American Society for Information Science 44*, 161-174.

Järvelin, K. & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS), 20*, 422-446.

Järvelin, K., Price, S. L, Delcambre, L. M. L., & Nielsen, M. L. (2008). Discounted cumulated gain based evaluation of multiple-query IR sessions. *Proceedings of ECIR '08,* Glasgow, Scotland.

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.

Jones, R., Rey, B., Madani, O., & Greiner, W. (2006). Generating query substitutions. *Proceedings of the WWW Conference*, Edinburgh, Scotland, 387-396.

Kelly, D., & Fu, X. (2006). Elicitation of term relevance feedback: An investigation of term source and context. *Proceedings of SIGIR '06*, 453-460.

Kelly, D. & Gyllstrom, K. (2011). An examination of two delivery modes for interactive search system experiments: Remote and laboratory. *Proceedings of the 29th ACM Conference on Human Factors in Computing Systems (CHI '11)*, Vancouver, Canada.

Mei, Q., Zhou, D., & Church, K. (2008). Query suggestion using hitting time. *Proceedings CIKM*, Napa Valley, CA, 469-477.

Raghavan, V. V., & Sever, H. (1995). On the reuse of past optimal queries. *Proceedings of SIGIR '95*, Seattle, WA, 344-350.

Ruthven, I. (2003). Re-examining the potential effectiveness of interactive query expansion. *Proceedings of SIGIR '03,* Toronto, CA, 213-220.

Ruthven, I., & Lalmas, M. (2003). A survey on the use of relevance feedback for information access systems. *Knowledge Engineering Review, 18*(2), 95-145.

Smyth, B., Balfe, E., Freyne, J., Briggs, P., Coyle, M., & Boydell, O. (2004). Exploiting query repetition and regularity in an adaptive community-based Web search Engine. *User Modeling and User-Adapted Interaction, 14*(5), 382-423.

White, R. W., Bilenko, M., & Cucerzan, S. (2007). Studying the use of popular destinations to enhance web search interaction. *Proceedings of SIGIR '07*, Amsterdam, The Netherlands, 159-166.

White, R. W., Ruthven, I. & Jose, J. M. (2005). A study of factors affecting the utility of implicit relevance

feedback. *Proceedings for the 28th Annual International ACM SIGIR Conference on Research and*

Development in Information Retrieval (SIGIR 2005), 35-42.

Vakkari, P. (2001). Changes in search tactics and relevance judgments when preparing a research proposal: A summary of the findings of a longitudinal study. *Information Retrieval, 4*(3-4), 295-310.

Voorhees, E. M. (2006). Overview of the TREC 2005 Robust Retrieval Track. *Proceedings of TREC-14*.