A Study of Real-Time Query Expansion Effectiveness

Ryen W. White[1]
Microsoft Research
One Microsoft Way
Redmond, WA 98052
ryen.white@microsoft.com

Gary Marchionini
School of Information and Library Science
University of North Carolina
Chapel Hill, NC 27599
march@ils.unc.edu

ABSTRACT
In this poster, we describe the study of an interface technique that provides a list of suggested additional query terms as a searcher types a search query, in effect offering interactive query expansion (IQE) options while the query is formulated. Analysis of the results shows that offering IQE during query formulation leads to better quality initial queries, and an increased uptake of query expansion. These findings have implications for how IQE should be offered in retrieval interfaces.

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

General Terms
Experimentation, Human Factors

Keywords
Real-time query expansion, query quality

1. INTRODUCTION
Improving the quality of queries has been of great interest in Information Retrieval (IR) research. Techniques like Relevance Feedback (RF) [6] have been proposed as a way in which IR systems can support the iterative development of a search query using examples of relevant information. Interactive query expansion (IQE) uses RF to suggest additional terms for query modification [3]. However, studies of IQE effectiveness have shown that it can be worthwhile but searchers may make poor term selections [5], or rarely use it in operational settings [1]. This could be related to how IQE is presented, and it is therefore important to investigate alternative ways of offering IQE.

Real-Time Query Expansion (RTQE) describes an interface mechanism whereby candidate expansion terms are presented to the searcher as they enter their search query. This approach integrates IQE directly into query formulation, giving help at a stage in the search when it can positively affect query quality, and possibly supporting the development of improved expansion strategies by searchers. Although similar techniques have already been implemented (e.g., Google Suggest), there exists to our knowledge no study of how effective such techniques are for real searchers. In this poster we describe a study of RTQE effectiveness based on the quality of queries it helps generate.

[1] This research was conducted while the first author was employed at the University of Maryland, College Park, MD USA 20742.

SIGIR'06, August 6–11, 2006, Seattle, Washington, USA.
ACM 1-59593-369-7/06/0008.

2. STUDY
A laboratory-based within-subject user study was conducted. Three experimental systems were developed: a search system with no query expansion (*Baseline*), and two systems that used Pseudo-Relevance Feedback (PRF) [4]. *RealTime* implements a variant of RTQE that uses the titles and abstracts of top-ranked results as sources for PRF, and presents query expansion options as a searcher enters their query. *Retrospective* also uses the top ten titles and abstracts as PRF sources, but offers query expansion options after a retrieval operation has been performed. The difference between *RealTime* and *Retrospective* was *when* the expansion support was offered. On each of these two systems the top ten terms are displayed in a "Recommended words" list situated between the query entry box and the search button. To append a term from the list of recommendations to the current query, the searcher can double-click the term in the list with the mouse pointer. Figure 1 illustrates the component in action.

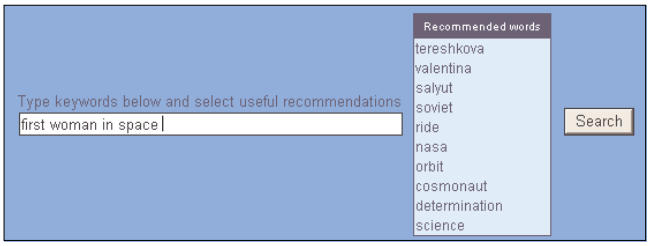

Figure 1. Term suggestions in real-time.

In *RealTime* the terms are presented as a list very shortly (less than two seconds depending on network latency) after the searcher finishes typing the first term of their query, and updates after each term is typed. Searchers may either select a term or ignore the suggestions, and complete their query.

A total of 36 subjects were recruited between the campuses of the University of Maryland at College Park, and the University of North Carolina at Chapel Hill (18 subjects per campus). Subjects were both undergraduate and graduate students from a range of nine different majors. Subjects were run independently. For the study we developed six known-item retrieval type tasks and six open-ended, exploratory type tasks. The exploratory tasks were phrased in the form of simulated work task situations [2], i.e., short search scenarios that were designed to reflect real-life search situations and allow subjects to develop personal assessments of relevance. Each experiment ran for up to two hours. Subjects attempted 12 tasks (two known-item and two exploratory tasks on each system), and completed background, post-system, and exit questionnaires containing semantic differentials, Likert scales, and open-ended questions. Their interaction with the systems was also logged for later analysis.

3. FINDINGS

We use the data derived from the study to assess the quality of the queries generated for known-item and exploratory searches. Query quality is a complex construct that is dependent on many factors such as the searcher's knowledge about the need, search experience, system experience, and the mapping between the need and the information source. As an estimate of query quality we employed a panel of two judges who independently assessed the quality of every query expressed for all subjects using a 5-point scale. The judges met with one of the experimenters and discussed ways to assign values. The basic agreement was to examine the task, conduct a search, and then identify the key concepts in the task to use as basis for judging the subject queries. The judges then coded queries for one task together to establish a common rating scheme. They then independently assessed the queries for each of the tasks. To minimize differences in quality estimates by the judges, the mean of the two judgments was taken as the overall query quality for each query. The mean average query quality ratings for the initial query alone (since it allowed us to isolate the RTQE) and all queries, are shown in Table 1.

Table 1. Mean average query quality (lower = better).

Measure	Baseline		RealTime		Retrospective	
	Initial Query	All Queries	Initial Query	All Queries	Initial Query	All Queries
Known-item	2.14	1.88	1.86	1.86	2.07	1.82
Exploratory	2.01	1.75	1.70	1.72	1.99	1.65
Overall	2.07	1.81	1.78	1.77	2.03	1.73

We applied two-way ANOVA for the "Initial Query" and for "All Queries" separately. Statistical analysis is conducted at $p < .05$.

Initial Query: There was a significant effect of system on initial query quality $F(2,142) = 12.37$, $p < .001$. Post hoc comparisons using a Tukey Test reveal that *RealTime* led to significantly higher initial query quality than other systems for each of the task types. There were no statistically significant differences in initial query quality between the two task types $F(1,71) = 3.26$, ns.

All Queries: Across all queries submitted to the three systems there were no statistically significant differences in query quality $F(2,142) = .95$, ns. However, there were statistically significant differences between the task types $F(1,71) = 4.98$, $p = .029$, with lower means on the known-item tasks. Across all queries, known-item searches, which have a more defined task description, appear to result in higher query quality. One explanation is that the space of high quality queries for these tasks is more constrained, making it perhaps easier to generate good queries.

We also analyzed the composition of queries in two further ways: query iterations, and unique query terms used per query *and* per search task. The analysis of iterations suggests that there were significantly more iterations for exploratory tasks $F(1,71) = 12.53$, $p < .001$, but no difference between search systems $F(2,142) = 1.06$, ns. There was no significant effect of tasks $F(1,1152) = .09$, ns or systems $F(2,1152) = 1.18$, ns on the number of unique terms *per query*. However, *per task*, there are fewer unique terms for known-item tasks $F(1,71) = 18.65$, $p < .001$, perhaps because the tasks were well-defined and fewer result sets were viewed.

4. DISCUSSION AND IMPLICATIONS

An analysis of query quality showed that RTQE improved the quality of initial queries for both known-item and exploratory tasks, making it potentially useful during the initiation of a search, when searchers may be in most need of support. If RTQE is capable of enhancing the quality of some queries, and does not having a detrimental effect on other aspects of search performance, then there is a case for it to be implemented as a feature of all search systems. A promising characteristic of RTQE is that it does not force searchers to use it, or indeed do anything radically beyond the scope of their normal search activities. Additional analysis of the findings, presented in [7], shows that compared to post-retrieval query expansion, RTQE lowers task completion times, increases searcher engagement, and increases the uptake of IQE (*RealTime*: 44% of queries used expansion, *Retrospective*: 28% of queries used expansion).

The nature of the query expansion support offered did not appear to affect the number of query terms, or the number of query iterations. In fact, it was *Retrospective* that led on average to the highest query quality across all queries. This may be because the system provided two types of support: searchers were shown the ten query expansion terms, and they were shown the titles, abstracts, and URLs of the documents from which those terms were derived. The presence of this information may provide an additional source from which to choose terms, but perhaps more importantly, give practiced, motivated searchers a sense of the type of documents that their query retrieved, and a sense for the context within which query modification terms occur in the collection. To this end, in future work we will implement an alternate interface mechanism embedding query expansion terms in the context they appear in their source documents. The future of IQE may well reside in techniques that couple expansion more closely with searchers' normal information-seeking behaviors.

5. REFERENCES

[1] Anick, P. Using terminological feedback for web search refinement: a log-based study. In *Proceedings of the 26th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 88-95. 2003.

[2] Borlund, P. Experimental components for the evaluation of interactive retrieval systems. *Journal of Documentation*, 56(1): 71-90. 2000.

[3] Efthimiadis, E.N. Query expansion. *Annual Review of Information Systems and Technology*, 31: 121-187. 1996.

[4] Jinxi, X. and Croft, W.B. Query expansion using local and global document analysis. In *Proceedings of the 19th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 4-11. 1996.

[5] Ruthven, I. Re-examining the potential effectiveness of interactive query expansion. In *Proceedings of the 26th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 213-220. 2003.

[6] Salton, G. and Buckley, C. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4): 288-297. 1990.

[7] White, R.W. and Marchionini, G. Examining the effectiveness of real-time query expansion. *Information Processing and Management* (in submission).