Evaluating Sources of Query Expansion Terms

Xin Fu & Diane Kelly
School of Information & Library Science
University of North Carolina at Chapel Hill, NC USA 27599-3360
+1 919.962.8065
[fu | dianek] @ email.unc.edu

ABSTRACT

This study investigates the effectiveness of retrieval systems and human users in generating terms for query expansion. We compare three sources of terms: system generated terms, terms users select from top-ranked sentences, and user generated terms. Results demonstrate that overall the system generated more effective expansion terms than users, but that users' selection of terms improved precision at the top of the retrieved document list.

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

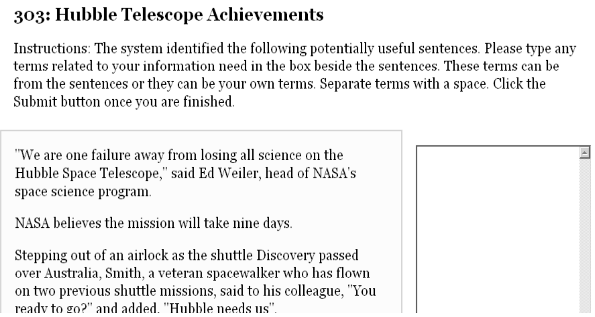
Keywords: Query expansion, query length, source of query term

1. INTRODUCTION

Sources of query expansion terms is an important aspect of many term relevance feedback (RF) studies. Approaches include having the system suggest a list of terms, and automatically adding them to users' queries (automatic RF), allowing users to pick which terms to add (interactive RF), and eliciting new terms from users. Ruthven [3] compared the relative effectiveness of interactive query expansion and automatic query expansion and found that users were less likely than systems to select effective terms for query expansion. Kelly, et al. [1] evaluated an interface that elicited information from users about their information needs beyond simple queries and found that the additional information significantly improved retrieval performance. Finally, in a study of term sources for query expansion during user-intermediary retrieval, Spink [4] found that the most effective query expansion terms came from users. The work presented here extends previous work by investigating the effectiveness of the system and users in suggesting terms for query expansion. Three sources of terms were compared: system generated terms, terms users select from top-ranked sentences, and user generated terms.

2. METHOD

A secondary analysis of data, collected through a study on RF interfaces [2], was employed for the method of this study. In this study, the TREC HARD 2005 collection, which includes 3 GB of news articles, 50 standard TREC topics and binary relevance assessments, was used. The interface used to collect the data is displayed in Figure 1. It displayed twenty sentences for each search topic. Users were asked to enter terms they wanted to add to their queries in the text box. Users were further instructed that terms could be from sentences or their own terms.



303: Hubble Telescope Achievements

Instructions: The system identified the following potentially useful sentences. Please type any terms related to your information need in the box beside the sentences. These terms can be from the sentences or they can be your own terms. Separate terms with a space. Click the Submit button once you are finished.

"We are one failure away from losing all science on the Hubble Space Telescope," said Ed Weiler, head of NASA's space science program.

NASA believes the mission will take nine days.

Stepping out of an airlock as the shuttle Discovery passed over Australia, Smith, a veteran spacewalker who has flown on two previous shuttle missions, said to his colleague, "You ready to go?" and added, "Hubble needs us".

Figure 1. Term Relevance Feedback Interface

We used the Lemur IR toolkit (http://www.lemurproject.org) to conduct our experiments, with its basic defaults for indexing and Okapi BM25 for retrieval. To populate the interface, we used the information contained in the title field of the TREC HARD topics to build queries for each topic and the default pseudo RF technique in Lemur. We used pseudo RF to obtain the top 20 ranking terms from the top 10 ranking documents for each topic (we modified the Lemur retrieval module to output these terms and sentences to a file). Then for each term, we ran one word queries on a collection consisting of all sentences from the corresponding top 10 documents to identify sentences. We then used the top ranking sentences for each term query to populate the interface. Thus, the interface displayed terms identified by pseudo RF techniques nested within example sentences.

Twenty undergraduates, recruited from the university community, spent 1 hour working on 10 topics (the 50 TREC topics were assigned systematically to users). For each topic, users first reviewed a *search topic form* that displayed the topic and asked them to construct initial Web queries for each topic. Then users were presented with the RF interface (Figure 1) and asked to provide terms related to their topics for query expansion. From users' responses we were able to identify (1) terms users selected that the system would have suggested via traditional term RF (system generated terms); (2) terms users selected that were contained within sentences that the system would not have suggested (top-ranked sentences); and (3) terms users identified that were not displayed in the interface (user generated).

3. RUNS AND RESULTS

To prepare the dataset for analysis, we first dropped cases where users did not provide initial queries (n=44) or RF terms (n=4). Thus, the dataset we used for analyses included 152 cases. For each case, we used users' initial queries from search topic forms (*UQ*) to construct a baseline run (the mean length of users' initial

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

queries was 4.15) and a pseudo RF run (*UQps*) where we added all top 20 ranking terms from the RF interface to users' queries.

Users on average provided 17.5 terms through the RF interface. We divided each set of terms into three groups: terms that the system would have suggested (i.e., top 20 ranking terms) for RF; terms that were contained within sentences, but that the system would not have suggested for RF; and terms that were strictly user generated. The mean number (and standard deviation) of terms in each group was 7.1 (4.9), 3.9 (3.4) and 6.4 (5.0), respectively.

We added each group of terms to users' initial queries separately and in combination, which resulted in 7 experimental runs: *UQt*, *UQs*, *UQu*, *UQts*, *UQtu*, *UQsu*, and *UQtsu* (t=term, s=sentence, u=user, tu=term+user, etc.). Table 1 displays the mean R-precision and P@10 for the different runs, ordered by R-precision. Note the different ordering when ranked by P@10.

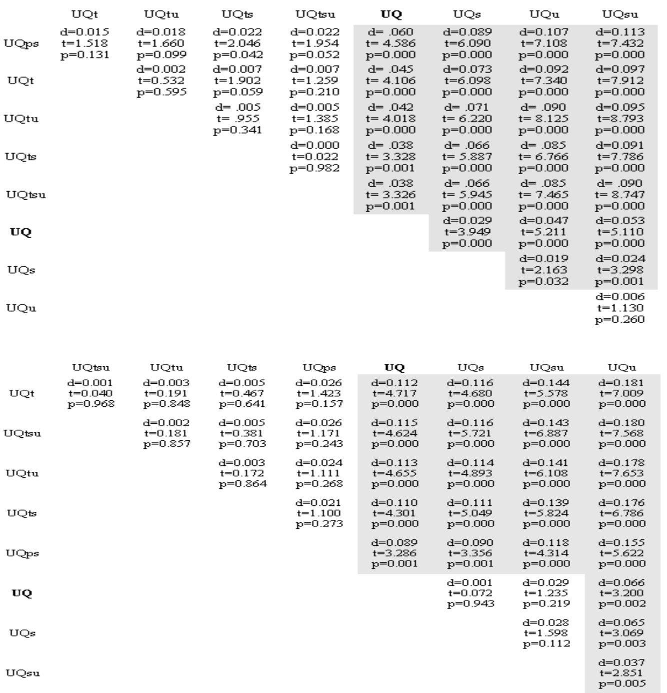
Table 1. Mean performance, number of queries improved, and query length (baseline shaded)

RunII	R-Precision		Precision at 10		Mean Query Length
	Mean (Std. Dev.)	Queries mproved	Mean (Std. Dev.)	Queries mproved	
UQps	.260 (.201)	97	.428 (.359)	76	24.15
UQt	.244 (.191)	94	.455 (.357)	77	11.28
UQtu	.242 (.183)	94	.452 (.350)	79	17.67
UQts	.237 (.191)	88	.449 (.360)	80	15.24
UQtsu	.237 (.185)	87	.454 (.356)	77	21.64
UQ	.199 (.172)	N/A	.339 (.297)	N/A	4.15
UQs	.171 (.156)	35	.338 (.313)	42	8.09
UQu	.152 (.139)	34	.274 (.276)	32	10.51
UQsu	.147 (.137)	44	.311 (.308)	45	14.48

We conducted paired-sample t-tests of runs for both measures. The results suggest that runs fell into two groups: runs that contained system generated terms (*UQps*, *UQt*, *UQts*, *UQtu*, *UQtsu*) and those that did not (*UQ, UQs, UQu, UQsu*). We found statistically significant differences in all pairs across groups, but no statistically significant differences in pairs within the first group, and few statistically significant differences in pairs within the second group (see Figure 2). When compared to the baseline run, runs containing system suggested terms significantly improved retrieval performance. Runs adding only sentence terms and/or user terms to the baseline queries performed worse than the baseline (although not significantly so in all cases). The overall results suggest that the system was more successful at identifying good query expansion terms for retrieval than users.

We also counted the number of queries in each run where retrieval performance improved. Results are displayed in the third and fifth columns in Table 1. A natural break occurred again between runs with and without system suggested terms. Runs with system suggested terms improved R-precision over the baseline run in about 60% of cases and P@10 in more than half. This provided another piece of evidence for the effectiveness of the system in suggesting good terms for query expansion.

Notably, the pseudo RF run (*UQps*) was the best performing run in terms of R-precision, but only ranked fifth in terms of P@10. Considering the different emphasis of the two measures and the mean query length for each run (displayed in the last column in Table 1), the result seems to suggest that the pseudo RF technique was not particularly effective in improving precision at the top of the retrieved document list and that human jurisdiction over system suggested terms is needed for more precise results.



	UQt	UQtu	UQts	UQtsu	**UQ**	UQs	UQu	UQsu
UQps	d=0.015 t=1.518 p=0.131	d=0.018 t=1.660 p=0.099	d=0.022 t=2.046 p=0.042	d=0.022 t=1.954 p=0.052	d= .060 t= 4.586 p=0.000	d=0.089 t=6.090 p=0.000	d=0.107 t=7.108 p=0.000	d=0.113 t=7.432 p=0.000
UQt		d=0.002 t=0.532 p=0.595	d=0.007 t=1.902 p=0.059	d=0.007 t=1.259 p=0.210	d= .045 t= 4.106 p=0.000	d=0.073 t=6.098 p=0.000	d=0.092 t=7.340 p=0.000	d=0.097 t=7.912 p=0.000
UQtu			d= .005 t= .955 p=0.341	d=0.005 t=1.385 p=0.168	d= .042 t= 4.018 p=0.000	d= .071 t= 6.220 p=0.000	d= .090 t= 8.125 p=0.000	d=0.095 t=8.793 p=0.000
UQts				d=0.000 t=0.022 p=0.982	d= .038 t= 3.328 p=0.001	d= .066 t= 5.887 p=0.000	d= .085 t= 6.766 p=0.000	d=0.091 t=7.786 p=0.000
UQtsu					d= .038 t= 3.326 p=0.001	d= .066 t= 5.945 p=0.000	d= .085 t= 7.465 p=0.000	d= .090 t= 8.747 p=0.000
UQ						d=0.029 t=3.949 p=0.000	d=0.047 t=5.211 p=0.000	d=0.053 t=5.110 p=0.000
UQs							d=0.019 t=2.163 p=0.032	d=0.024 t=3.298 p=0.001
UQu								d=0.006 t=1.130 p=0.260

	UQtsu	UQtu	UQts	UQps	**UQ**	UQs	UQsu	UQu
UQt	d=0.001 t=0.040 p=0.968	d=0.003 t=0.191 p=0.848	d=0.005 t=0.467 p=0.641	d=0.026 t=1.423 p=0.157	d=0.112 t=4.717 p=0.000	d=0.116 t=4.680 p=0.000	d=0.144 t=5.578 p=0.000	d=0.181 t=7.009 p=0.000
UQtsu		d=0.002 t=0.181 p=0.857	d=0.005 t=0.381 p=0.703	d=0.026 t=1.171 p=0.243	d=0.115 t=4.624 p=0.000	d=0.116 t=5.721 p=0.000	d=0.143 t=6.887 p=0.000	d=0.180 t=7.568 p=0.000
UQtu			d=0.003 t=0.172 p=0.864	d=0.024 t=1.111 p=0.268	d=0.113 t=4.655 p=0.000	d=0.114 t=4.893 p=0.000	d=0.141 t=6.108 p=0.000	d=0.178 t=7.653 p=0.000
UQts				d=0.021 t=1.100 p=0.273	d=0.110 t=4.301 p=0.000	d=0.111 t=5.049 p=0.000	d=0.139 t=5.824 p=0.000	d=0.176 t=6.786 p=0.000
UQps					d=0.089 t=3.286 p=0.001	d=0.090 t=3.356 p=0.001	d=0.118 t=4.314 p=0.000	d=0.155 t=5.622 p=0.000
UQ						d=0.001 t=0.072 p=0.943	d=0.029 t=1.235 p=0.219	d=0.066 t=3.200 p=0.002
UQs							d=0.028 t=1.598 p=0.112	d=0.065 t=3.069 p=0.003
UQsu								d=0.037 t=2.851 p=0.005

Figure 2. T statistics from Pair-wise comparisons of R-Precision (top) and P@10 (bottom). (d = difference between two runs; for all tests degree of freedom = 151; shaded cells significant at 0.05)

4. CONCLUSIONS

In this study, we found that the pseudo RF run outperformed runs that consisted of terms that were user-selected and -generated according to R-precision. However, we also found that among terms that users selected from the interface, terms that would have been suggested by the system for term relevance feedback were more effective at improving precision-oriented performance than other terms contained within sentences and terms that users generated on their own. This demonstrates the system's ability to identify good terms for query expansion, but also suggests that human intervention is important if precision-oriented results are desired. Whether this intervention is cost effective with respect to user effort and performance gain is a question for future research.

5. REFERENCES

[1] Kelly, D., Dollu, V. D. & Fu, X. (2005). The loquacious user: A document-independent source of terms for query expansion. In *Proceedings of SIGIR '05,* Salvador, Brazil, 457-464.

[2] Kelly, D. & Fu, X. (2006). Elicitation of Term Relevance Feedback: An Investigation of Term Source and Context. In *Proceedings of SIGIR '06*, Seattle, WA.

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

[4] Spink, A. (1995). Term relevance feedback and mediated database searching: Implications for information retrieval practice and systems design. *Information Processing and Management*, 31(2), 161-172.