Comparing Query Logs and Pseudo-Relevance Feedback for Web-Search Query Refinement

Ryen W. White
Microsoft Research
One Microsoft Way
Redmond, WA 98052
ryenw@microsoft.com

Charles L. A. Clarke
School of Computer Science
University of Waterloo
Waterloo, Canada
claclark@plg.uwaterloo.ca

Silviu Cucerzan
Microsoft Research
One Microsoft Way
Redmond, WA 98052
silviu@microsoft.com

ABSTRACT

Query logs and pseudo-relevance feedback (PRF) offer ways in which terms to refine Web searchers' queries can be selected, offered to searchers, and used to improve search effectiveness. In this poster we present a study of these techniques that aims to characterize the degree of similarity between them across a set of test queries, and the same set broken out by query type. The results suggest that: (i) similarity increases with the amount of evidence provided to the PRF algorithm, (ii) similarity is higher when titles/snippets are used for PRF than full-text, and (iii) similarity is higher for navigational than informational queries. The findings have implications for the combined usage of query logs and PRF in generating query refinement alternatives.

Categories and Subject Descriptors

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

General Terms

Experimentation, Human Factors.

Keywords

Pseudo-relevance feedback, Web search, query logs.

1. INTRODUCTION

Struggling searchers can benefit from system support to help them resolve their information problems. Pseudo-relevance feedback (PRF) [6] assumes top-ranked retrieved information sources are relevant, and takes terms from those sources and offers them to searchers as query refinement alternatives. PRF algorithms have more knowledge of term distribution statistics than searchers, can provide recommendations representative of the current content of highly-ranked sources, and can generate refinements for all queries for which there are search results. However, query refinement may indicate searcher dissatisfaction with top-ranked search results, and taking terms from these results for use in PRF may not align well with searcher intentions. An alternative is to use the query refinement behavior of many searchers captured in the query logs of large-scale systems such as Web search engines [4]. This has the advantage of representing popular intentions but refinements are chosen based on popularity not discriminatory power and recency is not guaranteed.

Despite the popularity of the two techniques, this poster is the first to report results of a comparison of PRF and query log-based refinement. Knowledge of when the techniques differ and when they are similar is vital in designing query suggestion algorithms that use multiple sources of evidence; a long-term direction for our continuing research in this area.

SIGIR'07, July 23–27, 2007, Ámsterdam, The Netherlands.
ACM 978-1-59593-597-7/07/0007.

2. STUDY

In this section we describe the query refinement techniques, the metric used to compare them, and the study methodology.

2.1 Query Refinement Techniques

Two query refinement techniques were tested that both generate a ranked list of refinements and scores for a given unrefined query.

2.1.1 Popular Query Extensions (QE)

We employed the query logs of a popular commercial search engine and extracted QEs from queries that contained the original query as a prefix, based on the observation that the most important terms in query refinement are in prefix position [4]. For term ranking, we used the overall frequency of the QEs in the logs rather than the frequency with which they followed the original query in search sessions. We did this for simplicity and to avoid introducing an unnecessary search engine bias (as in-session refinements depend heavily on the search results presented).

2.1.2 Pseudo-Relevance Feedback (PRF)

We selected a traditional PRF approach [1] that used different sources (i.e., either the full-text or titles/snippets) from the top-ranked search results to compute terms for query refinement. The source of terms was varied according to the experimental design, as is described later in this poster. Terms that appear in these sources were scored, such that the score for each term t was:

$$PRF_t = C_t . IDF_t$$

where C_t is the number of top-ranked sources that contain t, and IDF_t is the inverse document frequency of t generated across all assumed-relevant documents, for all test queries used in the study.

2.2 Similarity Metric

Through applying QE and PRF, for a given query we obtained a ranked list of QEs with associated frequencies, and a ranked list of PRF terms, with associated scores. QE frequencies resembled a Zipfian distribution and PRF scores decreased at a constant rate. It was therefore problematic to use term weights to compare the techniques as terms were being scored on different scales. To address this and compare the two term lists we use Normalized Cumulative Discounted Gain (NDCG) [1], a measure traditionally used for result set evaluation. NDCG matched our intuition better than other measures of similarity such as the Jaccard coefficient, since the importance of the query refinements (estimated through information gain) is considered. Formally, NDCG for the top n terms between two types of query refinement: query extensions (E) and pseudo-relevance feedback (P) is defined as:

$$Similarity(E, P) = N_E \sum_{i=1}^{n} \frac{2^{r_P(i)} - 1}{log(i+1)}$$

where sum is over all terms in n, $r_P(i)$ is the rating assigned to a term at position i in P (which is set to zero if the term is not in E), and N_E is a normalization constant chosen so that a perfect ordering of the terms (in this case the query extension ordering) will receive a score of one. Since NDCG is not symmetric we needed to compute it twice, once for each refinement technique (i.e., $Similarity(E, P)$ as above and $Similarity(P, E)$, where N_P replaces N_E and $r_E(i)$ replaces $r_P(i)$). In doing so, each technique took a turn at being the "ideal" ordering and the other type was the "test" ordering. Since we did not have any relevance ratings for the terms we had to choose ratings based on our estimation of the information gain for the user of choosing a term at a given rank. We did this based on the frequencies/scores assigned to each term by the refinement techniques we tested. For query extensions $r_E(i) = n/i$ (e.g., 20, 10, 6.67, etc. in line with the Zipfian distribution of the frequencies described earlier), whereas for the PRF $r_P(i) = n - (i - 1)$ (e.g., 20, 19, 18, etc. in line with the constant rate of decrease in scores assigned by PRF).

2.3 Methodology

We chose the top 20 QE and PRF refinements for each of 636 test queries which were obtained by randomly sampling by frequency a one month query log of the Windows Live search engine (i.e., each query had a chance of being selected proportional with its frequency). The size of this set was reduced from 1000 as we needed QE for each query. We removed multi-term extensions from QE as they were not supported by our PRF algorithm. We also removed stop words, and performed rudimentary stemming to remove plurals. We were concerned that the number and type of sources used may influence the quality of terms selected for PRF. To address this, our study included six PRF models that varied term source (the full-text of Web pages or search engine titles/snippets) and the number of results used as feedback (5, 10, or 20). For consistency, PRF terms were generated from result scrapes in the same time frame as query logs were extracted.

Prior to the study, the test queries were hand-classified by two trained judges into three categories: (i) *navigational* (i.e., user goal is to get to a particular, known Web site), (ii) *informational* (i.e., user goal is to acquire information about the query topic), and *resource / transactional* (i.e., user goal is to obtain something other than information e.g., service or entertainment). Initially judges worked independently, and used the description in [5] as the basis for classification. The Cohen's Kappa value for ratings emerging from this portion of activity was .72, signifying a "good" inter-rater agreement. To resolve discrepancies in the ratings, assessors met, and for each query, discussed the rationale behind their rating, and selected a final classification. Test queries comprised 43.0% *navigational*, 44.7% *informational*, and 12.3% *resource/transactional*, roughly in agreement with [5].

3. FINDINGS

We examine the similarity between QE and PRF. In Table 1 we show the similarity for each "term source" – "document number" pair for the top 20 terms, across all queries and three query types. (E, P) is used to denote the use of the QE as ideal ranking and PRF as test ranking in the NDCG computation, and (P, E) is used to denote the use of PRF as ideal ranking and QE as test ranking.

Table 1. Similarity between QE and PRF.

Query Type	Term source	Number of documents used for PRF					
		5		10		20	
		(*E*,*P*)	(*P*,*E*)	(*P*,*E*)	(*E*,*P*)	(*P*,*E*)	(*E*,*P*)
All	Titles/Snippets	.256	.228	.289	.267	.314	.325
	Full text	.108	.104	.142	.137	.222	.196
Nav	Titles/Snippets	.277	.256	.321	.300	.354	.366
	Full text	.139	.135	.168	.160	.263	.233
Inf	Titles/Snippets	.239	.208	.250	.234	.285	.285
	Full text	.081	.080	.120	.111	.197	.155
Res	Titles/Snippets	.244	.199	.282	.272	.308	.328
	Full text	.103	.077	.128	.149	.167	.216

A number of observations can be made from Table 1: (i) (E, P) and (P, E) seem closely related so could perhaps be combined into a single measure, (ii) the similarity between QE and PRF is higher for titles/snippets than full-text, perhaps because snippets may contain a higher concentration of terms that appear in similar. contexts to the unrefined query, so there is less scope for deviation from the query, (iii) the similarity between QE and PRF increases with the number of documents used for PRF, perhaps since more evidence is provided to the PRF algorithm from documents that are ranked lower for the unrefined query, and (iv) QE and PRF are significantly more similar for navigational queries than informational queries.[1] QE and PRF suggestions for navigational queries were topically coherent (e.g., "mapquest" extensions were mainly variants of maps and driving directions). However, for informational queries QE and PRF refinements covered many aspects, with less intersection. For example, for "academy awards" QE refinements were about winners, fashion, pictures, predictions, gossip, and after-show events, whereas PRF refinements were about statuettes, the venue, actors, and ballots.

4. DISCUSSION AND IMPLICATIONS

Our findings show that source, amount of feedback, and query type affect the similarity between QE and PRF. The differences attributable to query type are most interesting. Conceivably, both techniques could be deployed in parallel and refinements offered based on query classification. For example, the techniques appear interchangeable for navigational queries, but complementary for informational queries, and PRF is better able than QE to serve rare queries. In addition, when QE and PRF were least similar, queries seemed ambiguous (e.g., "globe", "woman"), whereas when most similar, queries seemed specific (e.g., "baby names", "cnn"). Query classification based on QE-PRF similarity may enhance existing approaches to query difficulty assessment e.g., [2].

5. REFERENCES

[1] Buckley, C., Salton, G., and Allan, J. (1992). Automatic retrieval with locality information using SMART. *Proc. TREC-1*, 59-72.

[2] Cronen-Towsend, S., Zhou, Y., and Croft, W.B. (2002). Predicting query performance. *Proc. ACM SIGIR*, 299-306.

[3] Järvelin, K. and Kekäläinen, J. (2000). Information retrieval evaluation methods for retrieving highly relevant documents. *Proc. ACM SIGIR*, 41-48.

[4] Jones, R., Rey, B., Madani, O., and Greiner, W. (2006). Generating query substitutions. *Proc. WWW 2006,* 387-396.

[5] Rose, D.E. and Levinson, D. (2004). Understanding user goals in Web search. *Proc. WWW 2004*, 13-19.

[6] Xu, J. and Croft, W.B. (1996). Query expansion using local and global document analysis. *Proc. ACM SIGIR*, 4-11.

[1] One-tailed independent measures t-tests ($\underline{t}(556) \geq 2.58$, $\underline{p} \leq .01$ $\alpha = .017$)