

# Re-Ranking Search Results Using Query Logs

Ziming Zhuang  
The Pennsylvania State University  
University Park, PA 16802, USA  
[rickzhuang@psu.edu](mailto:rickzhuang@psu.edu)

Silviu Cucerzan  
Microsoft Research  
Redmond, WA 98052, USA  
[silviu@microsoft.com](mailto:silviu@microsoft.com)

## ABSTRACT

This work addresses two common problems in search, frequently occurring with underspecified user queries: the top-ranked results for such queries may not contain documents relevant to the user's search intent, and fresh and relevant pages may not get high ranks for an underspecified query due to their freshness and to the large number of pages that match the query, despite the fact that a large number of users have searched for parts of their content recently. We propose a novel method, *Q-Rank*, to effectively refine the ranking of search results for any given query by constructing the query context from search query logs. Evaluation results show that *Q-Rank* gains a considerable advantage over the current ranking system of a large-scale commercial Web search engine, being able to improve the relevance of search results for 82% of the queries.

## Categories and Subject Descriptors

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

## General Terms

Algorithms, Design, Experimentation.

## Keywords

Ranking, query logs, search engine, relevance feedback.

## 1. INTRODUCTION AND PRIOR WORK

The sheer amount of Web pages and the exponential growth of the Web suggest that users are becoming more and more dependent on the search engines' ranking schemes to discover information relevant to their needs. Typically, users expect to find such information in the top-ranked results, and more often than not they only look at the document snippets in the first few result pages [2] and then they give up or reformulate the query. This can introduce a significant bias to their information finding process and calls for ranking schemes that take into account not only the overall page quality and relevance to the query, but also the match with the users' real search intent when they formulate the query.

Query logs of large-scale search engines contain the queries issued by a huge number of users, and are consequently an implicit source of collective endorsement about "what typical users are looking for" in any particular time frame. We propose a novel re-ranking algorithm, *Q-Rank*, which uses distributional information about the query context as extracted from search logs, to leverage implicit knowledge about the users' search intents and apply such knowledge to effectively refine the ranking of web search results.

This study has strong ties with previous work on query expansion,

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

ACM CIKM '06, November 6–11, 2006, Arlington, VA, USA

Copyright 2006 ACM 1-58113-000-0/00/0004...\$5.00.

which has been shown to be an effective method to bridge the gap between the lexical choices made by users and the lexical content of the Web (e.g. [1, 5]) and, more general, the users' information needs and their expression of those needs. Recently, there have been developments that exploit query logs for search results refinement. For example, in [7], past sequence of queries are used to complement the current query in estimating document relevance. In [6], previous queries are selected based on the similarity between their search results and those retrieved by the target query and then used to suggest an extended document list.

## 2. RE-RANKING ALGORITHM

*Q-Rank* is based on a straight-forward yet very effective rationale, that the most frequently seen *query extensions* of a target query (terms extracted from queries that contain the target query as an affix) and *adjacent queries* (queries that immediately precede or follow a query in a user search session) provide important hints about users' search intents. Intuitively, the distribution over query reformulations in search engine query logs at any point in time can be regarded as a snapshot of the typical user interests related to the concept in a target query; thus, when a user submits the target query, it can be assumed that she may be interested in a collection of documents that closely match this distribution [4].

We first formalize the definition of query context extracted from the query logs as follows. Let  $Q$  denote the set of queries in a search engine query log for a given time frame. We use the notation  $Q_{next}(q)$  for the set of queries that were seen following a target query  $q$  in user search sessions, and  $Q_{prev}(q)$  for queries that were seen preceding it. The union of these two sets will be referred to as the *adjacent queries*:  $Q_{adj}(q)$ .

We define the *user-based expansions* of a query  $q$  as being all logged queries that contain  $q$  as a prefix and the *query extensions* as being all the suffixes obtained by removing  $q$  from such expansions. Formally,  $Q_{ext}(q) = \{q_{ext}|q." ".q_{ext} \in Q\}$ , where " " denotes an empty space and "." denotes the operation of concatenating strings.

Let  $D(q)$  (or simply  $D$  when no ambiguity arises from omitting the target query) denote a set of candidate documents for ranking. In this work, we assume that  $D$  contains the top-ranked  $n$  documents by a search engine for the target query. We assign a re-ranking score for each document  $d$  in  $D$  based on its lexical overlap with a set of most popular query extensions and adjacent queries to  $q$ , as in Definition 1. The numerator of this formula has two terms, which correspond to query extensions and adjacent queries, respectively. Each of the terms is weighted by the dampen factor, summed, and then divided by the original rank of the document in order to account for the initial ranking calculated by the search engine. Finally, the documents in  $D$  are re-ranked in descending order of their corresponding re-ranking scores.

$$RS(d, q) \doteq \frac{\gamma \cdot \sum_{i=1}^{|Q_{ext}|} tf(q_i, d) \cdot \ln \frac{|D|}{|D_{q_i}|} \cdot \ln \frac{qf(q_i)}{\sum_{j=1}^{|Q_{ext}|} qf(q_j)}}{R(d)} + \frac{(1-\gamma) \cdot \sum_{i=1}^{|Q_{adj}|} tf(q_i, d) \cdot \ln \frac{|D|}{|D_{q_i}|} \cdot \ln \frac{qf(q_i)}{\sum_{j=1}^{|Q_{adj}|} qf(q_j)}}{R(d)}$$

**Definition 1.  $Q$ -Rank re-ranking score.**  $Q_{ext}$  and  $Q_{adj}$  denote the query context sets.  $\gamma \in [0, 1]$  leverages the contribution of each type of query context.  $tf(q_i, d)$  denotes the frequency of the query context  $q_i$  in document  $d$ .  $D_{q_i}$  contains all documents  $d$  for which  $tf(q_i, d) > 0$ .  $qf(q_j)$  denotes the logged frequency of query  $q_j$ .  $R(d)$  denotes the initial rank of  $d$ .

### 3. EXPERIMENTATION

We were granted access to a dataset comprising several thousand queries associated with several million web search results from a popular commercial search engine (MSN Search). Each query and search result pair was scored by editors on a scale from 0 to 5, reflecting the page’s relevancy to the corresponding query: 0 being completely irrelevant and 5 being extremely relevant. From this dataset, we randomly selected two sets of 1,000 queries and the associated search results as our development datasets. Another 2,000 queries were randomly selected from the remaining data for evaluation. We also had access to a two-month query log of the same search engine.

We measured the ranking quality using the discounted cumulative gain (DCG) metric [3]. A higher DCG reflects a better ranking of the results. DCG for the top  $n$  results generated by MSN Search and  $Q$ -Rank were computed and compared.

We first investigated three parameter settings, in which only  $Q_{ext}$  were used, only  $Q_{adj}$  were used, and both type of query contexts were used. When adjacent queries were used, we employed an equal number of preceding and subsequent queries. In these experiments, using adjacent queries alone consistently produced the best ranking on the development sets.

Next, we investigated the performance of  $Q$ -Rank with various ranking ranges ( $n=10, 15, 20$ ) and number of re-rank candidates ( $c=20, 30, 40$ ). We observed that the largest improvements were measured for  $n=10$ , partly because many lower ranked documents had no qualitative judgment and  $Q$ -Rank achieved the best ranking performance for  $c=30$  for all types of query contexts.

While  $Q$ -Rank exhibited a tendency to push up high quality and relevant documents that were originally ranked lower, we also observed that keeping the top  $u=2$  search results unchanged produced better performance. We hypothesize that this was at least in part due to the fact that the major commercial search engines often employ lists of *definitives* to be shown as the top results.

We also ran a series of experiments to determine a good empirical value of the interpolation parameter ( $\gamma$ ). When varying  $\gamma$ , on average,  $Q$ -Rank improved the rankings for 75.8% of the re-ranked queries. The percentile peaked at  $\gamma=0$  (78.5%), which was consistent with our previous findings that using adjacent queries alone achieves the best results. For  $\gamma=0.5$ , the DCG scores were increased by an average 6.81% for 76.3% of the re-ranked queries,

representing more than half (53.8%) of the queries in the development set.

Finally, we tested  $Q$ -Rank on the evaluation dataset, using the set of parameters ( $n=10, u=2, c=30, \gamma=0.5$ ) that achieved the best performance on the development data. Results showed an improvement on 81.8% of the re-ranked queries with an average increase in the DCG scores of 8.99%. The characteristics of these results were very consistent with those from the development dataset. Figure 1 plots the percentage of queries with improved rankings for various query lengths. There was no consistent pattern in the performance changes that correlates with query length. Interestingly, for long queries (four words or more), the adjacent queries worked the best.

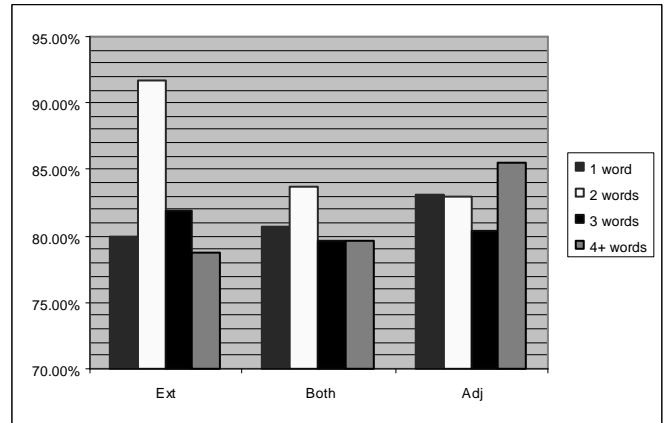


Figure 1. Queries with increased DCG scores in the test set

### 4. CONCLUSION

We propose  $Q$ -Rank as a novel method to effectively refine the Web search results ranking by using distributional information about query extensions and session-adjacent queries extracted from search engine logs. We conducted extensive experiments to determine the impact of various factors such as query length, re-rank range, interpolation coefficient between different types of query contexts, etc. Empirically,  $Q$ -rank was applicable to the majority of the real-world queries sampled from the logs of a large-scale commercial search engine, and outperformed the baseline ranker. Because  $Q$ -Rank is computationally efficient and independent of the underlying ranking algorithm, it can be easily integrated with any existing Web search system.

### 5. REFERENCES

- [1] Cui, H., Wen, J., Nie, J., and Ma, W. Probabilistic Query Expansion Using Query Logs. In *Proceedings WWW 2002*.
- [2] Jansen, J., and Spink, A. An Analysis of Web Documents Retrieved and Viewed”, In *Proceedings of ICOMP 2003*.
- [3] Jarvelin, K., and Kekalainen, J. IR Evaluation Methods for Retrieving Highly Relevant Documents. In *SIGIR 2000*.
- [4] Lau, T. and Horvitz, E. Patterns of search: Analyzing and modeling web query refinement. In *Proceedings of UM 1999*.
- [5] Kraft, R. and Zien, J. Mining Anchor Text for Query Refinement. In *Proceedings of WWW 2004*.
- [6] Nambiar, U., and Kambhampati, S. Providing Ranked Relevant Results for Web Database Queries. In *WWW 2004*.
- [7] Shen X., and Zhai, C. Exploiting Query History for Document Ranking in Interactive Information Retrieval. In *SIGIR 2003*.