

Using the Current Browsing Context to Improve Search Relevance

Mandar Rahurkar
University of Illinois-Urbana Champaign
405 North Mathews Avenue
Urbana, IL 61820
rahurkar@uiuc.edu

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

ABSTRACT

In this paper, we investigate the problem of improving the relevance of a Web search engine by adapting it to the dynamic needs of the user. We examine a representative case of sudden information need change, namely the exposure of the user to news data. In our earlier work we showed that the majority of queries submitted by users after browsing documents in the news domain are related to the most recently browsed document. We explore several methods of biasing the search by performing query expansion and re-ranking of the search results of a major search engine for queries identified as good candidates for contextualization. We show that these methods highly increase the similarity between the obtained top 10 search results and the most recently browsed document.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Algorithms, Experimentation, Measurement.

Keywords: Contextualization, Web search, query logs, query expansion.

1. INTRODUCTION

The importance of search personalization and contextualization has been well understood by the information retrieval community [3, 4]. Several previous studies have explored the use of query *context* to improve the relevance of the search results, by modeling the context as the history of executed searches, the search results clicked by a user, the documents on the user's computer, or the documents browsed by the user in the past. In practice, to avoid retrieving such personal data from the users, Web search engines employ approaches in which users build profiles for themselves by selecting categories of interest from a predefined list (e.g., *Google Personal*).

However, solutions that employ data aggregated over a long period of time or that take into account pre-specified categories of interests cannot adapt to the immediate needs of the user, which might change abruptly. Thus, using such aggregated/general data to bias the search may actually be detrimental to the quality of the search results, especially in situations in which the user's informa-

tion needs are highly dynamic and the aggregated user data does not capture the current intent of a user. One such situation, which we investigate in this work, is that of users querying a search engine while browsing the news.

2. DATA COLLECTION

We analyzed the query logs of the Live Search engine and retrieved those instances in which queries were submitted by users immediately after browsing a page in the FOX News domain. The choice of this news service was purely motivated by technical reasons (document archival and access methods employed by the news provider). 10,668 unique pairs of URLs and queries were extracted from search engine logs over a period of several months, of which 6,149 contained URLs that were indexed by the search engine employed at the time of running the experiments. In order to facilitate testing of (re-)ranking methods using that instance of the Live Search engine, only this restricted set of 6,149 pairs was used for experimentation.

3. QUERY CONTEXTUALIZATION

In our earlier work [2], we developed an algorithm for predicting whether a query is relevant to a previously browsed document, which achieved 96% precision at 90% recall (F-measure of 0.93).

In this work, we focus on improving the relevance of the search results for user queries which are related to a browsed document by leveraging the information contained in that document. Because the browsed documents are in the news domain, we decided that an important category of terms to be used for expanding the user query is constituted by the *named entities* mentioned in these documents, such as geopolitical entities, organizations, events, and people.

To overcome the absence of contextualized relevance judgements, we employed the lexical similarity of the documents retrieved to the originally browsed document as a measure of relevance of a retrieved set of documents. Formally, the similarity between the target document d corresponding to query q and the search results S_j , $j = 1..m$, is computed as the cosine similarity score of their corresponding vectors \vec{V} and \vec{V}_i . We observe both the distribution of similarity values obtained and the average similarity across all query document pairs.

We construct the altered queries q' by appending one named entity from the corresponding browsed document at a time to original query q . By employing expanded queries, we may obtain cosine similarity improvements simply because the results for the expanded queries must contain more terms from the target document. To account for this, we look at the improvement scale for different ways of choosing the named entities for expansion, and also, for different sets of (document, query) pairs.

Each of the documents from the collected (document, query) set is parsed to extract a set of named entities by employing the

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, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CIKM'08, October 26–30, 2008, Napa Valley, California, USA.
Copyright 2008 ACM 978-1-59593-991-3/08/10 ...\$5.00.

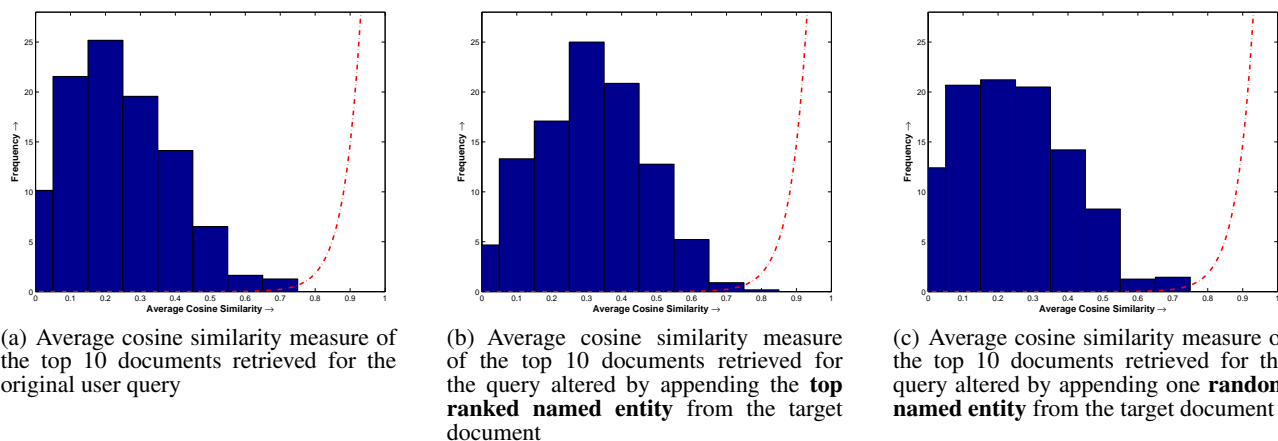


Figure 1: Comparison of the histograms of average cosine similarity between the browsed document and the top 10 search results retrieved for the original queries and for the queries altered by appending one named entity from the browsed document when the (document, query) pair is in the Relevant class.

Wikipedia-based system described in [1].¹ The named entities from each document are then ranked using a TF*IDF scheme. Term frequencies are obtained by calculating number of times a named entity occurs in the given document, including the identified co-references. Document frequencies are set as the estimated number of documents in the index of the Live Search engine that contain the Wikipedia canonical forms of the name entities, as obtained by querying for those forms.

We focus first on the pairs in which the query was related to the browsed document. Histograms of the cosine similarity between the target document and the top 10 search results for the original query as well as the two altered versions of the query are shown in Figures 1(a), 1(b), and 1(c). Note how the similarity mass shifts to larger values when the top-ranked entity is appended to the user query. However, the average cosine similarity increases very little when adding a random entity to the query, from **0.24** for the top 10 results for the original query to 0.25; in contrast, when adding the top ranked entity to the query, the average cosine similarity for the retrieved top 10 results increases substantially to **0.31**. By repeating this experiment for the case when the query was *not* related to the currently browsed document, we obtain a similarity score of 0.11 using the top ranked entity. To quantify these improvements for the two cases, we employed *skewness* as our performance metric. Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable. Ideally we want the left tail to be longer as shown by the dotted curve in Figure 1(b), which would imply higher similarity between the target document and the retrieved results. When the query was not related to the target document, skewness decreased only by **65%**, compared to the **242%** decrease for the queries relevant to the browsed document.

To better understand the scale of the improvements obtained by performing query alteration, we re-ranked the top 100 results returned for the original query by using the browsed document for feedback in a BM-25 scheme. When the query is relevant to the currently browsed document, the improvements in average similarity obtained by using this re-ranking strategy are nearly identical to those obtained by query expansion. Moreover, the document most similar to the browsed document in the top 10 re-ranked sets was on average exactly as relevant as the the most similar document in the

top 10 documents returned for the altered queries (cosine similarity to the browsed document of 0.46). When the query is not related to the browsed document, the change in similarity score for the re-ranked sets is again minimal, as in the case of query expansion.

As an additional evaluation of our query expansion strategy, we employ an approach that relies on the observation that the most similar document to the currently browsed document d is the document d itself. We compared the rank of the browsed document d in the list of top 10,000 search results retrieved by the original query and the expanded query. If document d is not found in the list, it is assigned a rank of 10,000. We then compute the mean reciprocal rank (MRR) averaged over all the Relevant document/query pairs. The MRR score for the original queries is **0.06**, while the MRR score for the queries expanded using the top ranked named entity in the corresponding document is **0.11**. Repeating the experiment for the non-related query-document pairs results in an observed change in MRR from 0.001 to 0.012. As expected, the MRR numbers for the pairs in the Irrelevant class are orders of magnitude lower than those for the pairs in the Relevant class.

By performing an exhaustive search over all named entities in the browsed document, we found that if we always knew the best possible named entity term to append, we could achieve an MRR of 0.57 for the expanded queries. This calls for further investigation on strategies to pick a named entity for expansion, including methods that use Web-based and query-log-based correlation statistics between terms.

4. CONCLUSION

We explored query expansion based on contextual information as a technique for leveraging the context information. We evaluated our algorithm on (document,query) pairs for which the query was determined to be related to the browsed document, as well as for which it was not relevant.

5. REFERENCES

- [1] S. Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In *EMNLP-CoNLL'07*, pages 708–716, 2007.
- [2] M. Rahrkar and S. Cucerzan. Predicting when browsing context is relevant to search. In *SIGIR'08*, pages 842–842, 2008.
- [3] M. Speretta and S. Gauch. Personalized search based on user search histories. In *IEEE - Web Intelligence*, pages 622–628, 2005.
- [4] J. Teevan, S. T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In *SIGIR'05*, pages 449–456, 2005.

¹This system solves not only the extraction of named entity references, but also performs their disambiguation to the canonical Wikipedia entities, and accounts for in-document co-references.