

Predicting when Browsing Context Is Relevant to Search

Mandar Rahurkar
University of Illinois-Urbana Champaign
405 North Mathews Avenue
Urbana, Illinois 61801
rahurkar@uiuc.edu

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

ABSTRACT

We investigate a representative case of sudden information need change of Web users. By analyzing search engine query logs, we show that the majority of queries submitted by users after browsing documents in the news domain are related to the most recently browsed document. We investigate ways of identifying whether a query is a good candidate for contextualization conditioned on the most recently browsed document by a user. We build a successful classifier for this task, which achieves 96% precision at 90% recall.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms

Algorithms, Experimentation, Measurement

Keywords

Contextualized Web search, query log analysis, query expansion

1. INTRODUCTION & PREVIOUS WORK

The importance of search personalization and contextualization has been well understood by the information retrieval community, as shown in [2] and in [3]. Several previous studies have explored the usage of query context to improve the relevance of the search results, by modeling the context as the history of executed searches (e.g., [5]), the search results clicked by a user (e.g., [4]), the documents on the user's computer (e.g., [1]), the documents browsed by the user in the past (e.g., [6]), or even richer models, which include some or most of these resources (e.g., [7]). However, solutions that employ data aggregated over a long period of time cannot adapt to the immediate needs of the user as these may change over time or even abruptly. Thus, using such aggregated data to bias the search may actually be detrimental to the quality of the search results if it does not capture the current user's intent. One scenario that we investigate in this work, in which the user's information needs are highly dynamic, is that of the users querying a search engine while browsing the news. For this scenario, we examine the use for contextualization of only the most recent document browsed by a user before issuing a query, hypothesized as the most relevant context for the query. Experimental results based on log data from the Microsoft Live Search engine confirm our hypothesis.

We first establish that the last document visited by the user plays an important role in identifying the information needs of the user,

especially when this document is from the news domain. We then show that the relatedness of a query to the most recently browsed document can be predicted accurately.

2. DATA COLLECTION

To collect data about the queries submitted by users after browsing a Web page in the news domain, we analyzed the query logs of a major commercial 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 (URL, query) pairs were extracted from search engine logs over a period of several months. 6,149 of the pairs contained URLs that were indexed by the search engine employed at the time of running the experiments. In order to facilitate further testing of ranking/re-ranking methods using this search engine (omitted in this poster version because of space limitations), only this restricted set of pairs were used for experimentation.

Ground truth decisions on whether a query q is semantically related to a Web page d in a logged pair (d, q) were obtained by asking human subjects to annotate the pairs with one of six category labels: Relevant, Irrelevant, Cannot Say, Navigational Query For News, Generic Page, and Technical Error. Because news pages have a lot of dynamic content such as current headlines and advertisements in addition to the main story, the annotators were asked to identify first the news story that constitutes the focus of the page and then judge whether the Web page is relevant to the given query.

For the annotation effort, we designed a user interface in which a document, query pair from the extracted log data is presented at a time to an annotator. The annotator has the option to search the Web for documents related to the query before making his/her decision. The annotator also has the option to see the query words that appear in the document highlighted.

Out of the 1,026 obtained tuples with non-conflicting labels, 53.9% belong to the class Relevant, 33% are Irrelevant, 5.8% fall under the Cannot say category, and 7.3% were classified as Navigational.

3. QUERY CONTEXTUALIZATION

Our annotation effort shows that the most recently browsed document by a user who sends a query to a search engine is related to the query almost 54% of the time when the document is in the news domain. In these cases, the last document visited can play an important role in understanding the information need of a user that issues a query to a search engine. However, the use of the browsed document could be detrimental for search relevance when the query is not clearly related to this document. Therefore, we investigate the task of accurately identifying these two types of situations.

As a baseline, we employ a classifier that labels a document as relevant to the corresponding query if and only if all the query terms are found in the document. The precision obtained by this classifier is very high (95%), but the recall is very low (17.5%).

We attempt to determine a better solution to this problem by employing a logistic regression classifier. We investigate three types of features designed to capture the document/query similarity on three important axes: lexical overlap between the two, similarity of the search results retrieved for the query and the document, and the time elapsed between the browsing and the querying action.

3.1 Query/Document Matching Ratio (QDMR)

The absolute number of query terms that appear in the corresponding browsed document may not necessarily be a good measure of the match. Thus, we employ instead the ratio between the number of query terms matching the document and the total number of terms present in a query. Formally, we compute the query/document matching ratio as:

$$QDMR(d, q) = \frac{|d \cap q|}{|q|}$$

where $|\cdot|$ denotes the cardinality of a set.

Although $QDMR$ was found in our empirical experiments as a good discriminative feature, about 2% of relevant document, query pairs in the training set do not have word tokens in common (e.g., the query “xom”, which is the stock symbol for Exxon Mobil, sent by a user after browsing a story on “the rise in crude oil prices”).

3.2 Query Results Similarity (QRS)

Starting from the empirical evidence that the degree of lexical overlap between a query and a document is a good indicator of the relatedness between the two, we investigate the use of the top results returned by a search engine for the query as an expanded query representation. Intuitively, we distinguish three cases: most search results are similar to the previously browsed document, only a few are similar, and none of the returned documents are similar to it. In the first case (high lexical overlap), contextualizing the query should not hurt the relevance, as most search results are already similar with the document to be used for contextualization. The last case (very low lexical overlap) can be seen as bearing not enough evidence to warrant contextualization. The case that needs more attention is the middle one, when the query returns some lexically similar search results to the browsed document. We hypothesize that in such a case, it is still desirable to contextualize the query. A feature that appears to account for all these cases is the maximum lexical similarity between the browsed document and any of the top search results.

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 of their corresponding vectorial representations. We then compute QRS as the maximum of these similarities:

$$QRS_m(d, q) = \max_{i=1..m} sim(d, S_i)$$

Because retrieving the entire documents in the top search results to compare them with the target document is prohibitively expensive for a real-time search engine (unless the vector forms of the retrieved documents are available), we approximate the lexical content of interest of the retrieved documents with the snippet of the document as generated by the search engine for the target query.

3.3 Time Elapsed Before Querying

Intuitively, the time elapsed since opening the most recently browsed document and issuing a query seems like a promising feature for

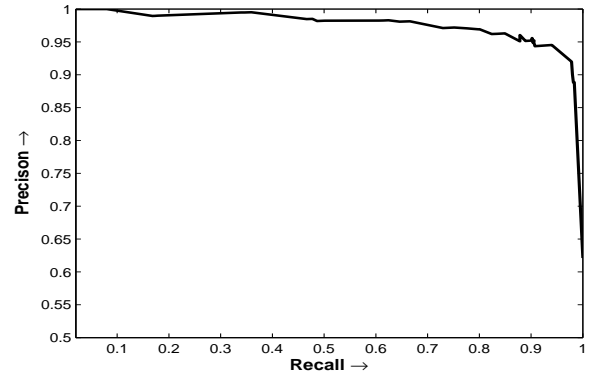


Figure 1: Precision-recall curve obtained for the task of identifying whether a browsed document is relevant to the subsequent query sent by a user to a search engine

determining the relationship between the two. However, this feature was found experimentally to be non-discriminative between the Relevant and Irrelevant document/query classes.

4. CLASSIFICATION RESULTS

Based on the empirical evidence obtained on the development set for each feature, we trained a logistic regression classifier using as features $QDMR$ and QRS_{10} .

Figure 1 shows the obtained precision-recall curve. We achieved **96%** precision at **90%** recall, which corresponds to an F-measure of 0.93. This is a considerable improvement compared to the 95% precision at 17.5% recall (corresponding F-measure of 0.3) obtained by the baseline system. While precision can be viewed as the most important of the two performance numbers because we want to avoid mistakenly contextualizing queries (and thus, generating user dissatisfaction), the recall number is also extremely important for a commercial search engine, as every percentage point earned in recall could translate to millions of additional queries to be contextualized.

5. CONCLUSION

The high performance achieved in identifying instances of browsed document and related query pairs substantiates the hypothesis that query contextualization using the most recently browsed document is a very promising area of investigation in relevance improvement for Web search engines.

6. REFERENCES

- [1] P. A. Chirita, C. S. Firan, and W. Nejdl. Personalized query expansion for the web. In *Proceedings of SIGIR'07*, pages 7–14, 2007.
- [2] C. Cool and A. Spink. Issues of context in information retrieval: an introduction to the special issue. *Information Processing and Management*, 38(5):605–611, 2002.
- [3] P. Ingwersen and K. Järvelina. *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer, 2005.
- [4] F. Qiu and J. Cho. Automatic identification of user interest for personalized search. In *Proceedings WWW'06*, pages 727–736, 2006.
- [5] X. Shen and C. Zhai. Exploiting query history for document ranking in interactive information retrieval. In *Proc. of SIGIR'03*, pages 377–378, 2003.
- [6] K. Sugiyama, K. Hatano, and M. Yoshikawa. Adaptive web search based on user profile constructed without any effort from users. In *Proceedings of WWW'04*, pages 675–684, 2004.
- [7] J. Teevan, S. T. Dumais, and E. Horvitz. Characterizing the value of personalizing search. In *Proceedings of SIGIR'07*, pages 757–758, 2007.