Query Suggestion based on User Landing Pages

Silviu Cucerzan and Ryen W. White
Microsoft Research
One Microsoft Way, Redmond, WA 98052, USA
{silviu, ryenw}@microsoft.com

ABSTRACT
This poster investigates a novel query suggestion technique that selects query refinements through a combination of many users' post-query navigation patterns and the query logs of a large search engine. We compare this technique, which uses the queries that retrieve in the top-ranked search results places where searchers end up after post-query browsing (i.e., the *landing pages*), with an approach based on query refinements from user search sessions extracted from query logs. Our findings demonstrate the effectiveness of using landing pages for the direct generation of query suggestions, as well as the complementary nature of the suggestions it generates with regard to traditional query log based refinement methodologies.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Query formulation, search process.*

General Terms
Algorithms, Experimentation, Human Factors.

Keywords
Search engine, query log analysis, landing pages, search sessions.

1. INTRODUCTION
Search engines facilitate access to the vast amount of information on the World Wide Web, and form a key enabling technology to help users address a broad range of information problems [3]. Given the diversity of their users' information needs, Web search engines have added many tools and features to enhance users' search effectiveness (e.g., spelling correction, machine translation, query federation, and question answering). Among these, all major commercial search engines provide *query suggestions* comprising semantically-related query refinements presented alongside search results to improve the effectiveness of subsequent query formulations.

In this poster, we investigate an approach to generate query suggestions based on the pages where many users who submit a particular query end up through post-query browsing. These so-called *landing pages* are frequently among the top-ranked search results, but can also lie on click trails far beyond the search results page, may not contain the queried terms, or may not even be indexed by the search engine. These popular landing pages can capture "the wisdom of the crowds" for information needs. We propose the suggestion of queries that return these pages as top-ranked search results, hence utilizing not only other users' querying decisions, but also the choices other users make following query submission. In a study conducted to evaluate this technique, we compare the suggestions it generates with a more traditional query suggestion approach similar to [6] that uses queries entered by many users following the original query in the same session.

SIGIR'07, July 23–27, 2007, Amsterdam, Netherland.
ACM 978-1-59593-597-7/07/0007.

2. RELATED WORK
There has been relevant related work on query refinement and the discovery of semantically similar queries. Beeferman and Berger [1] proposed an approach that exploits search engine click-through data, in which they represented user queries and visited search results as a bipartite graph, and applied an agglomerative clustering technique to identify related queries and Web pages. Cui *et al.* [4] showed that the lexical features of the query space and the Web document space are different, and investigated the mapping between query words and the words in visited search results in order to perform query expansion. Daumé and Brill [5] extracted suggestions based on document clusters that have common top-ranked documents. Various other sources of information have been investigated, such as pre-computed document abstracts [2], temporal query patterns [7], and query substitutions [6]. Our approach differs from all these techniques in that it utilizes user behavior information far beyond the search result list (and perhaps more like information that users eventually target) and leverages multiple sources of interaction data (i.e., queries and post-query browsing behavior).

3. STUDY
We conducted a study comparing landing pages and query sessions for query suggestions. The study did not compare relative effectiveness of these techniques but rather how the suggestions generated related to each other. In this section we describe the data from which suggestions are generated, the method used in both techniques to generate suggestions, the experiment conducted, and its findings.

3.1 Data
We used Live Search query logs, which contained the overall query frequency as well as anonymized user search session information. For landing page generation, we also used Web activity logs collected with permission from hundreds of thousands of users, which included details of all pages they visited. Both sets of logs were gathered during the same time period.

3.2 Query Suggestion Techniques
Two query suggestion techniques are compared in this study: one that uses landing pages (LP) and one using query sessions (QS).

3.2.1 Query Suggestion Using Landing Pages
From the Web activity logs we reconstructed temporally ordered sequences of viewed pages joined with hyperlink clicks and originating with a query submission to a commercial search engine such as Google, Yahoo!, and Live Search. We refer to these page sequences as *search trails*. These trails are terminated when one of the following events occurs: (1) a new query is submitted; (2) a user returns to their homepage, checks e-mail, logs in to an online service (e.g., MySpace or del.ico.us), types a URL or visits a bookmarked page; (3) a page is viewed for more than 30 minutes with no activity; (4) the user closes the active browser window. If a page meets any of these criteria, the trail terminates on the previous page (i.e., the *landing page*).

For each landing page of an original, unrefined query, we mine the search engine query logs to find queries sent by users that retrieved the landing page in the one of the top ten search results. These queries are collected and used as potential suggestions for the original query. The rationale behind this approach is that landing pages capture the underlying information intent of many users, so the queries that get users to these pages may be effective query suggestions. Given the potential for noise, we are particularly interested in using as suggestions those queries that return the landing page on the first position.

3.2.2 Query Suggestion Using Query Sessions

A more traditional method of generating query suggestions is to utilize the query refinements of many users as gathered from query logs (e.g., [6]). Using this approach suggestions are extracted for an original query if they immediately follow in a search *session* (i.e., a period of search activity terminated by 30 minutes or more of inactivity) conducted by at least three users.

Our goal is to explore the potential that these methods offer for query suggestion when used individually and in combination.

3.3 Experiment and Findings

We ran our experiments on a set of 5,000 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). We were able to compute at least one landing page for 2,073 of these queries. In total, we obtained 347,193 landing pages, giving us an average of 167 landing pages per covered query. We then mined the query logs of the search engine to locate queries distinct from the original queries and that retrieved the landing pages in the top 10 search results. We found a total of 290,117 such queries for 31,456 landing pages of 1,612 of the original queries. Figure 1 shows the average numbers of queries in the query log that returned the target landing pages as the top n ranked result (where n = 1..10). Note the large number of queries found to return the landing page in the top position. These could be viewed as "ideal" LP suggestions in a stratified variant of the technique. Overall, we were able to derive a total of 88,554 suggestions that would retrieve 14,452 landing pages on the top position for 1,366 of the original queries. QS suggestions were also derived from user search sessions in query logs (as described in Section 3.2.2). Using this method we generated 338,003 QS suggestions for 4,076 of the original queries.

The average Jaccard coefficient (defined as the size of the intersection divided by the size of the union) for the sets of original queries covered by the LP and QS was 0.38, suggesting that there was a fair amount of overlap between the queries for which suggestions were offered, with many more original queries having QS suggestions. Figure 2 illustrates differences in the coverage and average number of suggestions obtained by the two methods. While LP provides a much lower coverage than the QS, it has the merit of hypothesizing quite different suggestions. The average Jaccard coefficient for the suggestion sets was 0.13, and the LP suggestions appeared to be complementary to the QS suggestions. For example, for the query "academy awards", LP suggests queries focused on winners and the ceremony; while QS generated a scattered space of suggestions, most of them related to nominated movies and movie stars.

Each method offers distinct advantages and disadvantages. LP guarantees to return in the top search results of at least one of the user landing pages, but is highly biased toward pages retrieved in the past as top search results and pages which can be reached through hyperlinks from these. QS suggestions are guaranteed to match previously seen query refinements, but have the disadvantage of being correlated with search results retrieved by the search engine in those previous sessions and may be outdated rapidly with changing search indices and ranking.

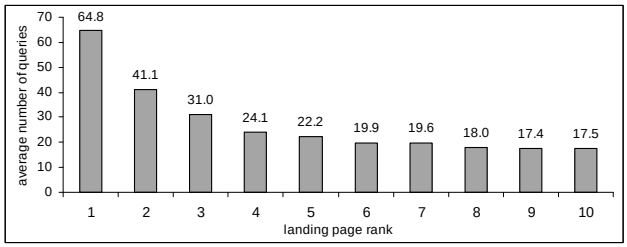

Figure 1. The average number of suggestions that retrieve the landing pages as a search result with a certain rank.

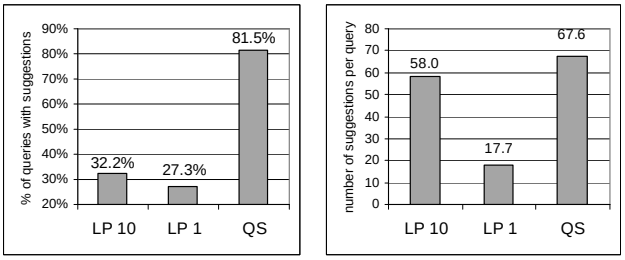

Figure 2. Coverage and number of suggestions obtained for the landing page and the query session -based methods (LP *n* denote suggestions that return the landing page in top *n* results).

4. CONCLUSION

Our findings show that computing landing pages for the user queries and extracting other queries that retrieve these pages as top-ranked search results obtains a reasonably high coverage and can provide quite different query suggestions from the method of mining query refinements from user sessions. This strongly suggests employing a combination of the two methods. The use of landing pages captures mostly the cases in which the users' needs are satisfied by a search engine (or pages a few clicks away from a result page), whereas the use of query sessions captures mostly the other cases, in which users refine queries to direct the search engine into a new result space because they were not completely satisfied with the results for the original query. Landing page suggestions generate less substantial modifications of the search result space, whereas query session suggestions could more significantly alter the search direction.

5. REFERENCES

[1] Beeferman, D. and A.L. Berger (2000). Agglomerative clustering of a search engine query log. *Proc. KDD*, 407-416.

[2] Billerbeck, B. and J. Zobel (2004). Techniques for efficient query expansion. *Proc. of SPIRE*, 30-42.

[3] Broder, A. (2002). Taxonomy of web search. *SIGIR Forum*

[4] Cui, H., J.R. Wen, J.Y. Nie, and W. Ma (2002). Probabilistic query expansion using query logs. *Proc.of WWW*, 325-332.

[5] Daumé, H. and E. Brill (2004). Web search intent induction via automatic query reformulation. *Proc. HLTNAACL*, 49-52.

[6] Jones, R., B. Rey, O. Madani, G. Greiner (2006). Generating query substitutions. *Proc.of WWW*, 387-396

[7] Vlachos, M., C. Meek, Z. Vagena, and D. Gunopulos (2004). Identification of similarities, periodicities and bursts for online search queries. *Proc. of SIGMOD*, 131-142.