Finding Relevant Website Queries

Brian D. Davison, David G. Deschenes, and David B. Lewanda
Department of Computer Science & Engineering
Lehigh University
19 Memorial Drive West
Bethlehem, PA 18015 USA
{davison,dgd4,dbl2}@cse.lehigh.edu

ABSTRACT

Search engine traffic is central to the success of many websites. By analyzing the queries-to-results graph generated by a search engine, our tool can recommend relevant queries for website optimization. We demonstrate our technique on a data set of hundreds of thousands of queries.

Keywords

Relationship analysis, search engines, query suggestion, competitive intelligence

1. INTRODUCTION

Search engines provide a mapping from queries to documents. Additionally, that mapping is often considered to be one-to-many, as, in most cases, many documents are considered relevant by a search engine to a given query. However, that mapping is really many-to-many, as a particular document is often considered relevant to many queries. More specifically, the mapping may be viewed as a directed bipartite graph where a set of queries maps to a set of documents.

Mining that graph permits the discovery of meaningful relationships of entities in the graph. For example, related URLs can be found in a two-step graph exploration (as shown by the first two steps in Figure 1). Finding related queries can be performed analogously, starting with the given query.

The savvy content provider realizes the potential for a search engine to send traffic to a website. As a result, content providers will often attempt to optimize particular pages to rank highly in the results of a particular query. By analyzing Web server logs, the content provider is able to determine which queries are successfully sending visitors to the site. However, the content provider has only a narrow view of what queries might be utilized to find the site — the queries found in the web site log and intuition about other possible queries. The content provider does not have a global view of what queries are made, and in particular does not know what relevant queries exist that probably should rank his site highly, but do not.

Our system is able to suggest such queries. We start with a mechanism that finds related Web sites, but extend it one additional step (as illustrated in Figure 1). We take the related URLs and find the set of queries that generate them, and remove those queries that also include the starting site. We rank this set of queries by the number of URLs that the query has in common with the set of URLs related to the starting URL.

WWW2003, May 20–24, 2003, Budapest, Hungary.

2. BACKGROUND

Our approach has strong ties to what might generally be called relationship analysis. In bibliometrics, researchers analyze patterns and relationships of co-citation and bibliographic coupling. Sociologists study social networks among people. In textual data mining, term co-occurrence is often utilized. On the Web, the study of the relationships between pages is typically called link analysis. In each of the cases above, one entity is considered related (because of co-occurrence, co-citation, or explicit linkage) to another entity of the same type.

The use of query result vectors to measure query similarity was found to be better than calculating similarities using the query term vector [6]. This idea was subsequently applied to the Web [2, 3]. Others have used query click-through data to relate queries and documents [1, 7]. Recently, Jeh and Widom [5] proposed a generalized technique to incrementally calculate the similarities (based on a bipartite graph structure) of all pairs of objects (not necessarily of the same type).

3. EXPERIMENTS

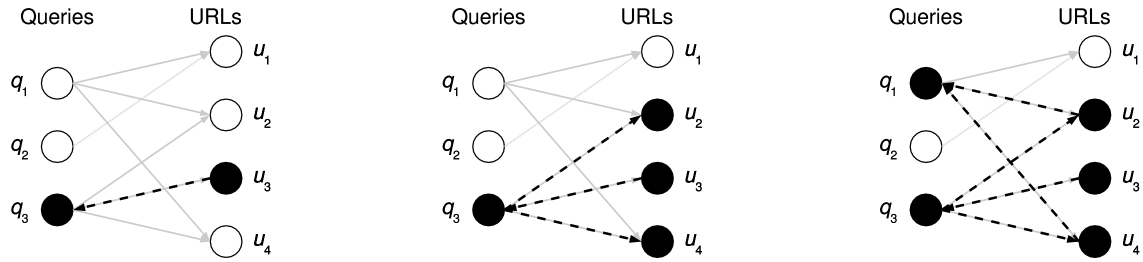
We used a trace of queries from the Excite[1] search engine that was collected on December 20, 1999. This trace[2] has been used by many others for query analysis (e.g., [4]). It contains almost 2.5 million requests recorded over an eight hour period. All queries in the trace were made lower-case, but not otherwise modified. Duplicate queries were then removed, after recording the frequency of each unique query.

The Google API[3] was used to collect result sets of the top ten document URLs for the most frequently occurring queries. In total, we have recorded results for 430,351 unique queries, generating 3,177,721 unique document URLs. While this is only about one third of all of the unique queries, it accounts for all queries with frequency greater than one, and many queries that were requested only once. Since this Excite dataset is fairly old, and will soon be exhausted, we have begun collection of our own query data, and will shortly augment our system with it.

In order to efficiently access our data we make use of established search engine data structures. The queries to documents mapping is broken down into two indices — one that maps a query to its resulting document URLs and another that maps the URL of a document to those queries for which it is a result.

We tested our system[4] with a number of requests for related pages and queries, and suggested website queries. Due to space

[1] `http://www.excite.com/`
[2] We thank Amanda Spink for providing access to this data.
[3] `http://www.google.com/apis/`
[4] Available from `http://wume.cse.lehigh.edu/`



(a) Find associated queries. (b) Find related URLs. (c) Find queries returning related URLs.

Figure 1: The process of discovering queries that should, but do not, rank a given site highly.

constraints, we show suggested queries for just two websites, and what might be learned from those suggestions. In Table 1, we discover that it might behoove mp3.com to somehow incorporate the terms “free”, “songs”, and “files” in their site. Similarly, Hallmark might want to include something incorporating the terms “greetings” and “electronic” into their site, so that it ranks higher on these related queries (Table 2).

4. DISCUSSION

We note that in time, the cached query results may become out-of-date. While presently we are more concerned with populating our data set than with freshness, a larger-scale implementation will require tracking and renewing result sets that are likely stale.

In the future, we also hope to add other search engine results to determine the effect that the search engine (with different data sets and ranking algorithms) has on the quality of results from our system. Google tends to rank home pages and popular pages highly, while a system focusing on textual analysis might generate a different flavor. Similarly, expanding the search engine results (to 100, for example) from the current 10 would allow for significantly larger related object calculations, and might prompt the use of a URL weighting scheme based on the rank of the URL in the result set.

We are currently working to incorporate query frequency into our rankings, as we expect that users of our system will want to focus on site optimization for popular queries.

5. SUMMARY

In recent years the use of link analysis has made significant improvements in the quality of Web search results. By turning to similar analyses of relationships between queries and the set of Web page results, we have demonstrated that these techniques can be used to recommend relevant queries for website optimization.

6. REFERENCES

[1] D. Beeferman and A. Berger. Agglomerative clustering of a search engine query log. In *Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 407–415, 2000.

[2] L. Fitzpatrick and M. Dent. Automatic feedback using past queries: Social searching? In *Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 306–313, Philadelphia, PA, July 1997.

[3] N. S. Glance. Community search assistant. In *Artificial Intelligence for Web Search*, pages 29–34. AAAI Press, July 2000. Presented at the AAAI-2000 workshop on Artificial Intelligence for Web Search, Technical Report WS-00-01.

[4] B. J. Jansen, A. Spink, and T. Saracevic. Real life, real users, and real needs: A study and analysis of user queries on the web. *Information Processing and Management*, 36(2):207–227, 2000.

[5] G. Jeh and J. Widom. SimRank: A measure of structural-context similarity. In *Proc. of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Alberta, Canada, July 2002.

[6] V. V. Raghavan and H. Sever. On the reuse of past optimal queries. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 344–350, Seattle, WA, July 1995.

[7] J.-R. Wen, J.-Y. Nie, and H.-J. Zhang. Clustering user queries of a search engine. In *Proceedings of the Tenth International World Wide Web Conference*, Hong Kong, May 2001.

where can i listen to free music
”mp3 songs”
where can i get mp3’s?
listen to music
where can i listen to music
where can i find free mp3s
where can i find mp3 songs?
mp3 files
where can i find mp3 files
download mp3 files

Table 1: Suggested queries found when starting with http://www.mp3.com/.

free animated post cards
online greetings cards
birthday ”electronic greeting”
christmas e greeting cards
free greetings cards
how can i send a greeting card?
’thank you greetings cards’
electronic cards
birthday e-cards
how can i find web cards to send

Table 2: Suggested queries found when starting with http://www.hallmark.com/.