Web Mining in Search Engines

Ricardo Baeza-Yates

Center for Web Research
Department of Computer Science
Universidad de Chile
Blanco Encalada 2120, Santiago, Chile
Email: rbaeza@dcc.uchile.cl

Abstract

Given the rate of growth of the Web, scalability of search engines is a key issue, as the amount of hardware and network resources needed is large, and expensive. In addition, search engines are popular tools, so they have heavy constraints on query answer time. So, the efficient use of resources can improve both scalability and answer time. One tool to achieve these goals is Web mining. Web mining has three branches: link mining, usage mining, and content mining. One important analysis in all these cases is the dynamic behavior. Here we give examples of link and usage mining related to search engines, as well as the related Web dynamics.

Keywords: Web usage mining, link analysis, Web dynamics, search engines.

Summary

Our main goal is to show how valuable is to perform log query mining, by presenting several different applications of this idea combined with standard link (Chakrabarti 2002) and Web dynamics (Levene et al. 2003). Although past research has focused in technical aspects of search engines, analyzing queries has a broader impact in Web search and design in two different aspects: *Web findability* and *information scent*. Web findability or ubiquity is a measure of how easy to find a Web site is, where search engines are the main access tools. To improve findability there are several techniques, and one of them is to use query log analysis of Web site search to include on the Web site text the most used query words. Information scent (Pirolli 1997) is how good it is a word with respect of words with the same semantics. For example, polysemic words (words with multiple meanings) may have less information scent.

We first present several Web characteristics, including the Web structure (Broder et al. 2000), and its relations (Baeza-Yates et al. 2001), and then the following applications:

- Queries to a search engine follow a power-law distribution, which is far from uniform. Using this query distribution, we present an inverted file organization that has three levels: precomputed answers, main, and secondary memory indexes. We show that by using half the index in main memory we can answer 80% of the queries, and that using a small number of precomputed answers we can improve the query answer time on at least 7% (Baeza-Yates et al. 2003).
- Second, we present an algorithm that uses queries and clicks to improve ranking (Zhang et al. 2002), which captures semantic relations of queries and Web pages. We include our own experiments and conclusions (Baeza-Yates et al. 2004). Queries to a search engine follow a power-law distribution, which is far from uniform.
- We end with the study of quantitative measures of the relation between the dynamics of the Web, its structure, and the quality of Web pages. Quality is studied using different link-based metrics considering their relationship with the structure of the Web and the last modification time of a page. We show that, as expected, Pagerank (Page et al. 1999) is biased against new pages, and we obtain information on how recency is related with Web structure (Baeza-Yates et al. 2002) and its evolution (Baeza-Yates et al. 2003), as well as other link based ranking measures such as authorities and hubs (Kleinberg 1998).

References

Ricardo Baeza-Yates and Carlos Castillo, Relating Web Characteristics with Link Based Web Page Raking, In *Proceedings of SPIRE 2001*, IEEE CS Press, Laguna San Rafael, Chile, pp. 21-32, 2001.

Baeza-Yates, Felipe Saint-Jean, and Carlos Castillo. Web Dynamics, Age and Page Quality, In *Proceedings of SPIRE 2002*, LNCS, Springer, Lisbon, Portugal, 2002.

Ricardo Baeza-Yates, and Felipe Saint-Jean. A Three Level Search Engine Index based in Query Log Distribution. SPIRE 2003, LNCS, Springer, Manaus, 2003.

Ricardo Baeza-Yates, and Barbara Poblete. Evolution of the Chilean Web Structure Composition. In *First Latin-American Web Congress*, Santiago, Chile, IEEE CS Press, 2003.

Ricardo Baeza-Yates. Query Usage Mining in Search Engines. In Web Mining: Applications and Techniques, Anthony Scime, editor. Idea Group, 2004.

Andrei Broder, Ravi, Kumar, Farzin Maghoul, Prabakhar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the Web: Experiments and models. In *9th World Wide Web Conference*, Amsterdam, 2000.

Copyright ©2004, Australian Computer Society, Inc. This paper appeared at Twenty-Seventh Australasian Computer Science Conference (ACSC2004), Dunedin, New Zealand. Conferences in Research and Practice in Information Technology, Vol. 26. Vladimir Estivill-Castro, Ed. Reproduction for academic, not-for profit purposes permitted provided this text is included.

Soumen Chakrabarti, Mining the Web: Discovering Knowledge from Hypertext Data. Morgan Kaufmann Publishers, San Francisco, CA, USA, 2002.

Jon Kleinberg. Authoritative sources in a hyperlinked environment. Proc. 9th Symposium on Discrete Algorithms, 1998.

Mark Levene and Alex Poulovassilis, editors. Web Dynamics, Springer, 2003.

Larry Page, Sergei Brin, Rajeev Motwani, and Terry Winograd. The Pagerank citation algorithm: bringing order to the Web. Tech. rep., Dept. of Computer Science, Stanford University, 1999.

Peter Pirolli. Computational Models of Information Scent-Following in a Very Large Browsable Text Collection, In Human Factors in Computing Systems: Proceedings of the CHI '97 Conference. ACM Press, New York, 3-10, 1997.

Dell Zhang, and Yisheng Dong. A Novel Web Usage Mining Approach For Search Engine. *Computer Networks* 39(3): 303-310, 2002.