Web Usage Mining in Search Engines

Ricardo Baeza-Yates
Universidad de Chile, Chile

ABSTRACT

Search engine logs not only keep navigation information, but also the queries made by their users. In particular, queries to a search engine follow a power-law distribution, which is far from uniform. Queries and related clicks can be used to improve the search engine itself in different aspects: user interface, index performance, and answer ranking. In this chapter we present some of the main ideas proposed in query mining and we show a few examples based on real data from a search engine focused on the Chilean Web.

INTRODUCTION

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. In this chapter we focus on Web usage mining of logs of queries and user clicks to improve search engines and Websites. We do not consider other kinds of Web mining such as link analysis (Chakrabarti, 2002), content mining, or Web dynamics (Levene & Poulouvassilis, 2003).

There are few papers that deal with the use of query logs to improve search engines, because this information is usually not disclosed. The exceptions deal with strategies for caching the index and/or the answers (Markatos, 2000; Saraiva et al., 2001; Xie &
O’Hallaron, 2002), and query clustering using click-through data associated with queries (obtaining a bipartite graph) for ranking or related goals (Beeferman & Berger, 2000; DirectHit, 1997; Wen, Nie & Zhang, 2001; Xue et al., 2002; Zhang & Dong, 2003). Other papers are focused on user behavior while searching, for example detecting the differences among new and expert users or correlating user clicks with Web structure (Baeza-Yates & Castillo, 2001; Holscher & Strube, 2000; Pradumonio et al., 2002). Recently, there has been some work on finding queries related to a Website (Davison et al., 2003) and weighting different words in the query to improve ranking (Schaale et al., 2003).

The main goal of this chapter is to show how valuable it is to perform log query mining, by presenting several different applications of this idea combined with standard usage mining. Although past research has focused on the 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\(^1\) or ubiquity is a measure of how easy it is to find a Website when search engines are the main access tools. To improve findability, there are several techniques. One is to use query log analysis of Website search and include in the Website text the most used query words.

Information scent (Pirolli, 1996) is how good a word is with respect to words with the same semantics. For example, polysemic words (words with multiple meanings) may have less information scent. The most common queries are usually the ones with more information scent. When analyzing Web search queries we find words that are found (or not found) on a site but have more or a similar information scent to words in the home page, and words that are not found that imply new information that needs to be added.

This chapter is organized as follows. We start with some basic concepts followed by some primary statistics of search engine usage. Next we present two applications of the search engine log. First, using the query distribution, we present an inverted file organization with three levels: precomputed answers, and main and secondary memory indexes. We show that by using half the index in main memory we can answer 80% of all queries, and that using a small number of precomputed answers we can improve the query answer time on at least 7% (Baeza-Yates & Saint-Jean, 2003b). Second, we present an algorithm that uses queries and clicks to improve ranking (Zhang & Dong, 2003) by capturing semantic relations of queries and Web pages. We conclude with some open problems.

### PRELIMINARIES

#### Zipf’s Law

Zipf’s law was introduced in the late 1940s to describe several empirical observations such as the distribution of the population of cities or the frequency of words in English written text (Zipf, 1932). If \( F_i \) is the frequency of the \( i \)-th most frequent event, we have that \( F_i \sim 1/i^\alpha \) where \( \alpha \) is a constant, and the parameter of the distribution. In a log-log graph, \( \alpha \) is the slope (without the sign) of the line. In the case of words, it means that there are few very frequent words (usually called stop words) and many unusual words. In Figure 1, the first part of the curve is the frequent words, while the last part of the curve is the unusual words. Perhaps, due to this distribution, the number of distinct words in a text (vocabulary) does not grow linearly, but follows a sublinear curve of the form

---

Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.
A Subspace Clustering Framework for Research Group Collaboration
[www.igi-global.com/article/subspace-clustering-framework-research-group/2602?camid=4v1a](www.igi-global.com/article/subspace-clustering-framework-research-group/2602?camid=4v1a)