Mining Generalized Query Patterns from Web Logs

Charles X. Ling*
Dept. of Computer Science, Univ. of Western Ontario, Canada
ling@csd.uwo.ca

Jianfeng Gao
Microsoft Research China
jfgao@microsoft.com

Huajie Zhang*
Dept. of Computer Sciences, Univ. of Western Ontario, Canada
hzhang@csd.uwo.ca

Weining Qian*
Dept. of Computer Science, Fudan University, China
wnqian@fudan.edu.cn

Hongjiang Zhang
Microsoft Research China
hjzhang@microsoft.com

Abstract
User logs of a popular search engine keep track of user activities including user queries, user click-through from the returned list, and user browsing behaviors. Knowledge about user queries discovered from user logs can improve the performance of the search engine. We propose a data-mining approach that produces generalized query patterns or templates from the raw user logs of a popular commercial knowledge-based search engine that is currently in use. Our simulation shows that such templates can improve search engine’s speed and precision, and can cover queries not asked previously. The templates are also comprehensible so web editors can easily discover topics in which most users are interested.

1. Introduction
Data mining is a process of discovering implicit and useful knowledge from large datasets [Fayyad, et al, 1996]. Knowledge about datasets can be as simple as statistical information such as means and deviations, but data mining aims at finding higher level or more powerful knowledge, such as classifiers, predictive or descriptive rules, Bayesian networks, cluster descriptions, and so on. Two most important vertical applications of data mining nowadays are in the financial industry and the Internet (World Wide Web).

We study data mining application on the Internet in this paper. In particular, we apply data mining to discover useful and implicit knowledge from web logs, which keep traces of information when users visit web pages on web servers. The purpose of web-log mining is to improve web performance (which is defined precisely later in the paper) by utilizing the mined knowledge. Indeed, data mining is very promising since popular web sites get millions of hits each single day, traditional methods or human would be infeasible to analyze such logs.

The log we study in this paper is from a popular commercial knowledge-based search engine that is currently in use. The web of the search engine allows users to submit keyword queries to a search engine to find articles about the queries, and to browse articles organized under a hierarchical structure of several levels. The log keeps users’ queries and their clicks, as well as their browsing activities. The file is quite large: one-day log would be over several hundreds megabytes in size.

It is well known that queries for web search engines are often too short to contain sufficient information to discriminate ambiguous documents. Therefore, additional knowledge, such as informative search terms

*This work was done while those authors were visiting Microsoft Research China.