Chapter IX

An Intelligent Metasearch Engine with Link Prediction and Page Clipping Generation Capabilities

Cheng-Jye Luh, Yuan-Ze University, Taiwan
Lin-Chih Chen, National Taiwan University of Science and Technology, Taiwan

Abstract

This chapter presents an intelligent metasearch engine that can recommend a user’s next hyperlink access and relevant paragraphs extracted from metasearch results. The proposed design is based on the primacy effect of browsing behavior, that users prefer top ranking items in search results. Three search methods were implemented in this engine. First, the search engine vector voting (SVV) method rearranges search results gathered from six well-known search engines according to their weights obtained from user behavior function. The hyperlink prediction (HLP) method then arrange the most likely accessed hyperlinks from the URLs in SVV search results. Finally, the page clipping synthesis (PCS) method extracts relevant paragraphs from the HLP search results. A user study indicated that users are more satisfied with the proposed search methods than with general search engines. Moreover, performance measure results confirmed that the proposed search methods outperform other metasearch and search engines.

Copyright © 2007, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.
Introduction

Internet users generally hope that search engines can locate the exact required information. This task is recognized by search engine designers as difficult (Jansen, Spink, Baterman, & Saracevic, 1998) because the inputs supplied by the users are generally insufficient for collecting suitable data. Literature shows that the Web query averages 2.3 terms long (Spink, Wolfram, Jansen, & Saracevic, 2001). Thousands of results, typically returned for such a short query, generally are arranged in their descending order of relevance using ranking algorithms (Li, 1998). Often the users are concerned with the top ranked URLs, and ignore the rest (Sougné, 2000). Therefore, providing the most relevant Web pages directly to the users would significantly reduce their navigation time on a variety of hyperlinks.

Search engines generally measure the similarity between user query and document contents using information retrieval (IR) techniques including the vector space models, probability models, and fuzzy logic models (Harman, 1992; Yates & Neto, 1999). These models, primarily based on the frequencies of keyword occurrence in documents, are frequently attacked by “keyword spamming” technique. That is, a document’s rank can be manipulated by duplicating the same set of keywords in a document (Yates & Neto, 1999). However, search engines can now detect and penalize keyword spamming (Mall-Net, 2001).

Li (1998) developed the HVV method to solve the keyword spamming problem using a SCI (Science Citation Index) like ranking scheme, which ranks each Web page by the number of times other Web pages contain links to it. Thus, a Web page would have a high weight if several hyperlinks point to it. However, this method does not distinguish the ratings given by high-quality Web pages from those given by low-quality Web pages (Henzinger, 2001). Brin and Page (1998) created a PageRank algorithm, which is currently used by the Google search engine, to remedy this problem. The page rank of a page is computed by weighting each hyperlink pointing to the page proportional to the page rank of the referring page containing the hyperlink. However, the PageRank algorithm’s recursive nature has raised concern about considerable computing power it needs to analyze the hyperlinks and page ranks (Sobek, 2002).

Collecting search results from search engines is an effective way to collect and rate relevant Web pages. This method is called metasearch (Dreilinger & Howe, 1996; Selberg & Etzioni, 1997). The metasearch approach saves a lot of time by searching only in one place and eliminating the need to use and learn several separate search engines. The quality of metasearch results depends on the search engines used and how the results are organized.

A metasearch method, called search engine vector voting (SVV), was developed in this study to rearrange the search results obtained from several well-known search engines. A Web page wins a vote from a search engine if it is listed on the top 50 items returned for a given query. The weight of a particular Web page is determined by its actual rankings in the search results of search engines from which it wins votes. The SVV method can solve the two problems encountered in using HVV stated above, since well-known search engines are generally recognized as high-quality Web sites, and SVV only considers the
www.igi-global.com/article/business-models-b2b/1834?camid=4v1a