Chapter III

Using Context Information to Build a Topic-Specific Crawling System

Fan Wu
National Chung-Cheng University, Taiwan

Ching-Chi Hsu
Institute for Information Industry, Taiwan

ABSTRACT

One of the major problems for automatically constructed portals and information discovery systems is how to assign proper order to unvisited Web pages. Topic-specific crawlers and information seeking agents should try not to traverse the off-topic areas and concentrate on links that lead to documents of interest. In this chapter, we propose an effective approach based on the relevancy context graph to solve this problem. The graph can estimate the distance and the relevancy degree between the retrieved document and the given topic. By calculating the word distributions of the general and topic-specific feature words, our method will preserve the property of the relevancy context graph and reflect it on the word distributions. With the help of topic-specific and general word distribution, our crawler can measure a page’s expected relevancy to a given topic and determine the order in which pages should be visited. Simulations are also performed, and the results show that our method outperforms the breath-first and the method using only the context graph.

INTRODUCTION

The Internet has now become the largest knowledge base in the human history. The Web encourages decentralized authoring in which users can create or modify documents
locally, which makes information publishing more convenient and faster than ever. Because of these characteristics, the Internet has grown rapidly, which creates a new and huge media for information sharing and exchange. There are more than two billion unique and publicly accessible pages on the Internet, and the Internet is estimated to continue to grow at an accelerating rate, 7.3 million pages per day (Cyveillance, 2003). As a result of the tremendous size of information on the Web and its substantial growth rate, it is increasingly difficult to search for useful information on the Web.

Traditional information retrieval methods can be utilized to help users search for their needed information in a database. But they appear to be ineffective when facing this mammoth Web. Researchers have proposed many techniques to facilitate the information seeking process on the Web. Search engines are the most important and commonly used tools (Brown et al., 1998), which usually have large indexes of the Web and attempt to use a centralized architecture to solve the problem of information seeking in a decentralized environment. Web search engines are usually equipped with multiple powerful spiders that traverse the Web information space. But the engines have difficulties in dealing with such huge amounts of data. For example, Google (http://www.google.com/) claimed to be the largest search engine in the world, can only index about 60% of the Web. The other problem is that search engines usually return hundreds or thousands of results for a given query, which makes users bewildered in finding relevant answers for their information need (Lawrence 1999, 2000).

The general-purpose search engines, such as Altavista (http://www.altavista.com/), offer high coverage of the possible information on the Web, but they often provide results with low precision. That is, the information does not match what the user wants. Directory search engines, such as Yahoo! (http://yahoo.com.tw/), can limit the scope of its search upon such manually compiled collections of the Web contents that are related to some specified categories. The directory search engines return the results with higher precision, exhibiting the beauty of labor-intensive efforts. However, compiling a well-organized directory search engine for each directory would be tedious and impossible, and automatic construction of such an engine seems to have a long way to go.

The topic-specific search engine is another type of search engines that is constructed and optimized in accordance with domain knowledge. When users are aware of the topic or category of their need, a topic-specific search engine can provide the information with higher precision than a general, or directory search engine does. For example, ResearchIndex (http://citeseer.nj.nec.com/cs/) is a full-text index of scientific literature that aims at improving the dissemination and feedback of scientific literature. Travel-Finder (http://www.travel-finder.com/) is a topic-specific Web service designed to help individuals find travel professionals and travel information. LinuxStart (http://www.linuxstart.com/) is another topic-specific Web service that provides a deliberate hierarchy of Linux topics and a search engine for queries focused on Linux-related topics. Topic-specific search engines usually incorporate in their system domain knowledge and use focused crawlers to construct their data collection. The crawlers try not to traverse the off-topic areas, concentrating themselves on the links that lead to documents of interest. One of the major problems of the topic-specific crawler is how to assign a proper order to the unvisited pages that the crawler may visit later. The method to measure the importance of a document on the Web through linkage information has been adopted by the general and topic-specific search engines. But the topic-specific crawlers can further incorporate domain-specific knowledge to facilitate subjective search.

Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.
Ranking Web Services using Web Service Popularity Score
www.igi-global.com/article/ranking-web-services-using-web-service-popularity-score/115936?camid=4v1a