IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Hershey PA 17033-1240, USA ITB8699
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

Chapter 10

World Wide Web Search Engines

Wen-Chen Hu
University of North Dakota

Jyh-Haw Yeh
Boise State University

ABSTRACT

The World Wide Web now holds more than 800 million pages covering almost all issues. The Web's fast growing size and lack of structural style present a new challenge for information retrieval. Numerous search technologies have been applied to Web search engines; however, the dominant search method has yet to be identified. This chapter provides an overview of the existing technologies for Web search engines and classifies them into six categories: 1) hyperlink exploration, 2) information retrieval, 3) metasearches, 4) SQL approaches, 5) content-based multimedia searches, and 6) others. At the end of this chapter, a comparative study of major commercial and experimental search engines is presented, and some future research directions for Web search engines are suggested.

INTRODUCTION

One of the most common tasks performed on the Web is to search Web pages, which is also one of the most frustrating and problematic. The situation is getting worse because of the Web's fast growing size and lack of structural style, as well as the inadequacy of existing Web search engine technologies (Lawrence & Giles, 1999a). Traditional search techniques are based on users typing in search keywords which the search services can then use to locate the desired Web pages. However, this approach normally retrieves too many documents, of which only a small fraction are relevant to the users' needs. Furthermore, the most relevant documents do not necessarily appear at the top of the query output list. A number of corporations and research organizations are taking a variety of approaches to try to solve these problems. These approaches are diverse, and none of them dominate the field. This chapter provides a survey and classification of the available World Wide Web search engine

techniques, with an emphasis on nontraditional approaches. Related Web search technology reviews can also be found in (Gudivada, Raghavan, Grosky, & Kasanagottu, 1997; Lawrence & Giles, 1998b; Lawrence & Giles, 1999b; Lu & Feng, 1998).

Requirements of Web Search Engines

It is first necessary to examine what kind of features a Web search engine is expected to have in order to conduct effective and efficient Web searches and what kind of challenges may be faced in the process of developing new Web search techniques. The requirements for a Web search engine are listed below, in order of importance:

1. effective and efficient location and ranking of Web documents;
2. thorough Web coverage;
3. up-to-date Web information;
4. unbiased access to Web pages;
5. an easy-to-use user interface which also allows users to compose any reasonable query;
6. expressive and useful search results; and
7. A system that adapts well to user queries.

Web Search Engine Technologies

Numerous Web search engine technologies have been proposed, and each technology employs a very different approach. This survey classifies the technologies into six categories: i) hyperlink exploration, ii) information retrieval, iii) metasearches, iv) SQL approaches, v) content-based multimedia searches, and vi) others. The chapter is organized as follows: Section 2 introduces the general structure of a search engine, and Sections 3 to 8 introduce each of the six Web search engine technologies in turn. A comparative study of major commercial and experimental search engines is shown in Section 9 and the final section gives a summary and suggests future research directions.

SEARCH ENGINE STRUCTURE

Two different approaches are applied to Web search services: genuine search engines and directories. The difference lies in how listings are compiled:

- Search engines, such as Google, create their listings automatically.
- A directory, such as Yahoo!, depends on humans for its listings.

Some search engines, known as hybrid search engines, maintain an associated directory. Search engines traditionally consist of three components: the crawler, the indexing software, and the search and ranking software (Greenberg & Garber, 1999; Yuwono & Lee, 1996). Figure 1 shows the system structure of a typical search engine.

Crawler

A crawler is a program that automatically scans various Web sites and collects Web documents from them. Crawlers follow the links on a site to find other relevant pages. Two search algorithms—breadth-first searches and depth-first searches—are widely used by crawlers to traverse the Web. The crawler views the Web as a graph, with the nodes being the objects located at Uniform Resource Locators (URLs). The objects could be (Hypertext Transfer Protocols (HTTPs), File Transfer Protocols (FTPs), mailto (e-mail), news, telnet, etc. They also return to sites periodically to look for changes. To speed up the collection of Web