World Wide Web Search Technologies

Wen-Chen Hu
University of North Dakota, USA

Hung-Jen Yang
University of North Dakota, USA

Jyh-haw Yeh
Boise State University, USA

Chung-wei Lee
Auburn University, USA

INTRODUCTION

The World Wide Web now holds more than six billion pages covering almost all daily issues. The Web's fast growing size and lack of structural style present a new challenge for information retrieval (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. Numerous search technologies have been applied to Web search engines; however, the dominant search methods have yet to be identified. This article provides an overview of the existing technologies for Web search engines and classifies them into six categories: i) hyperlink exploration, ii) information retrieval, iii) metasearches, iv) SQL approaches, v) content-based multimedia searches, and vi) others. At the end of this article, a comparative study of major commercial and experimental search engines is presented, and some future research directions for Web search engines are suggested. Related Web search technology review can also be found in Arasu, Cho, Garcia-Molina, Paepcke, and Raghavan (2001) and Lawrence and Giles (1999b).

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. 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.

BACKGROUND

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

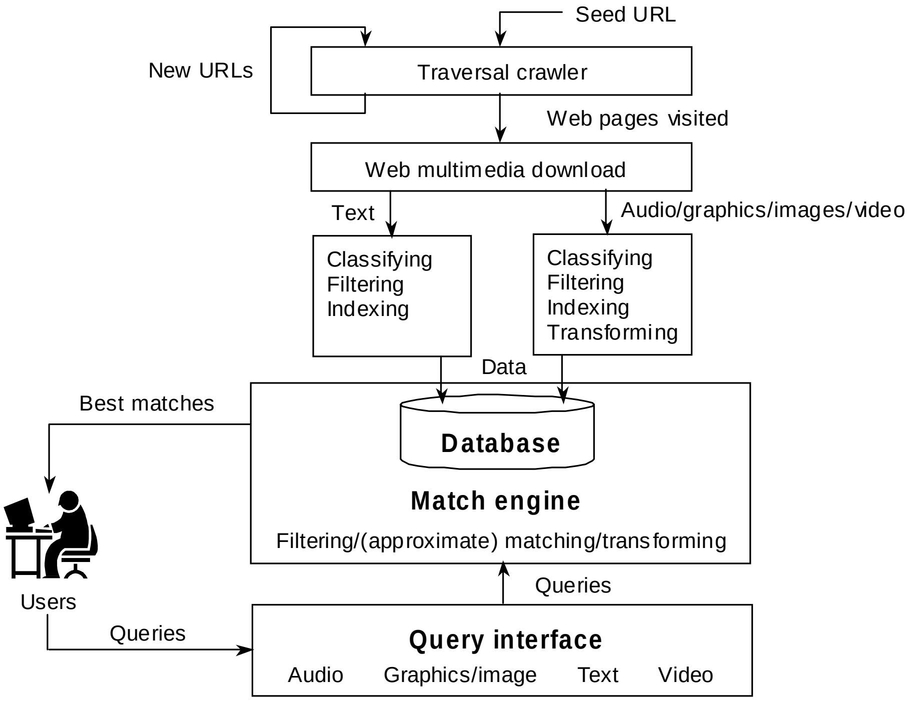
- Search engines, such as Google, create their listings automatically.
- A directory, such as Yahoo!, depends on humans for its listings.

Some search engines, known as hybrid search engines, maintain an associated directory. Figure 1 shows the system structure of a typical search engine. Search engines traditionally consist of three components: i) the crawler, ii) the indexing software, and iii) the search and ranking software:

- A crawler is a program that automatically scans various Web sites and collects Web documents from them. Two search algorithms, breadth-first searches and depth-first searches, are widely used by crawlers to traverse the Web.
- Automatic indexing is the process of algorithmically examining information items to build a data struc-

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. System structure of a Web search engine



ture that can be quickly searched. Traditional search engines utilize the following information, provided by HTML scripts, to locate the desired Web pages: i) content, ii) descriptions, iii) hyperlink, iv) hyperlink text, v) keywords, vi) page title, vii) text with a different font, and viii) the first sentence.

- Query processing is the activity of analyzing a query and comparing it to indexes to find relevant items. A user enters a keyword or keywords, along with Boolean modifiers such as "and", "or", or "not", into a search engine, which then scans indexed Web pages for the keywords. To determine in which order to display pages to the user, the engine uses an algorithm to rank pages that contain the keywords.

SEARCH ENGINE TECHNOLOGIES

This section examines the existing technologies for Web search engines and classifies them into six categories: i) hyperlink exploration, ii) information retrieval, iii) metasearches, iv) SQL approaches, v) content-based multimedia searches, and vi) others.

Hyperlink Exploration

Links can be tremendously important sources of information for indexers; the creation of a hyperlink by the author of a Web page represents an implicit endorsement of the page being pointed to. This approach is based on identifying two important types of Web pages for a given topic:

- Authorities, which provide the best source of information on the topic, and
- Hubs, which provide collections of links to authorities.

Authorities and hubs are either given top ranking in the search results or used to find related Web pages. A simple method to update a non-negative authority with a weight x_p and a non-negative hub with a weight y_p is given by Chakrabarti et al. (1999). If a page is pointed to by many good hubs, its authority weight is updated by using the following formula:

$$x_p = \sum_{q \text{ such that } q \to p} y_q ,$$

where the notation $q \to p$ indicates that q links to p. Similarly, if a page points to many good authorities, its hub weight is updated via

$$y_p = \sum_{q \text{ such that } p \to q} x_q .$$

5 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/world-wide-web-search-technologies/14753

Related Content

The Relationship Between Knowledge Automation and Employee Creativity
Shabina Shaikh (2022). *International Journal of Information Technology Project Management (pp. 1-25).*
www.irma-international.org/article/the-relationship-between-knowledge-automation-and-employee-creativity/311851

The Project Management Process of Planning and Budgeting in Public Construction Projects
Jesper Kranker Larsen, Lene Faber Ussing, Thomas Ditlev Brunoeand Søren Munch Lindhard (2015). *International Journal of Information Technology Project Management (pp. 20-33).*
www.irma-international.org/article/the-project-management-process-of-planning-and-budgeting-in-public-construction-projects/133221

Packet Inter-Arrival Distributions in Computer Network Workloads
Dennis Guster, David Robinsonand Paul Safonov (2005). *Encyclopedia of Information Science and Technology, First Edition (pp. 2260-2264).*
www.irma-international.org/chapter/packet-inter-arrival-distributions-computer/14595

IT Supporting Strategy Formulation
Jan Achterbergh (2005). *Encyclopedia of Information Science and Technology, First Edition (pp. 1728-1734).*
www.irma-international.org/chapter/supporting-strategy-formulation/14503

Project Management for IT Projects
Len Asprey (2005). *Encyclopedia of Information Science and Technology, First Edition (pp. 2341-2347).*
www.irma-international.org/chapter/project-management-projects/14610