IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 15, Issue 6 (Nov. - Dec. 2013), PP 34-37
www.iosrjournals.org

Design Issues for Search Engines and Web Crawlers: A Review

Deepak Kumar, Aditya Kumar
WCTM, Gurgaon, Haryana,

Abstract: *The World Wide Web is a huge source of hyperlinked information contained in hypertext documents. Search engines use web crawlers to collect these web documents from web for storage and indexing. The prompt growth of the World Wide Web has posed incomparable challenges for the designers of search engines and web crawlers; that help users to retrieve web pages in a reasonable amount of time. In this paper, a review on need and working of a search engine, and role of a web crawler is being presented.*
Key words: *Internet, www, search engine, types, design issues, web crawlers.*

I. Introduction

Today Internet [1,5] has become the universal source of information. Within a span of few years, it has changed the way we business and communicate. It has given a worldwide platform to us. The World Wide Web (WWW or web) [1, 2, 4, 12, 13] is a hyperlinked repository of hypertext documents lying in different websites distributed over far end distant geographical locations.

There is unbelievable growth in the number of web documents and it is estimated that currently there are millions of web servers around the world hosting trillions of web pages. With such a huge web repository, finding the right information at the right time is a very challenging task. Hence, there is a need for a more effective way of retrieving information from the web.

Search engines [3,7,8,9,11,12,13,14] operate as a link between web users and web documents. Without search engines, this vast source of information in web pages remain veiled for us. A search engine is a searchable database which collects information from web pages on the Internet, indexes the information and then stores the result in a huge database where from it can be searched quickly.

II. Search Engines

A general web search engine (Fig. 1) has three parts [3,10,11,12,13,14] i.e. Crawler, Indexer and Query engine. The web crawler (also called robot, spider, worm, walker or wanderer) is a module that searches the web pages from the web world. These are small programs that peruse the web on the search engine's behalf, and follow links to reach different pages. Starting with a set of seed URLs, crawlers extract URLs appearing in the retrieved pages, and store pages in a repository database.

The indexer extracts all the words from each page and records the URL where each word has occurred. The result is stored in a large table containing URLs; pointing to pages in the repository where a given word occurs.

Indexing methods used in web database creation are full text indexing, keyword indexing and human indexing. Full text indexing is where every word on the page is put into a database for searching. It helps user to find every example in response to a specific name or terminology. A general topic search will not be very useful in the database and one has to dig through a lot of false drops.

In keyword indexing only important words or phrases are put into a database. In human indexing a person examines the page and determines a very few key phrases that describes it. This allows the user to find a good start of works on a topic, assuming that the topic was picked by the human as something that describes the page.

The query engine is responsible for receiving and filling search requests from users. It relies on the indexes and on the repository. Because of the web's size, and the fact that users typically only enter one or two keywords, result sets are usually very large.

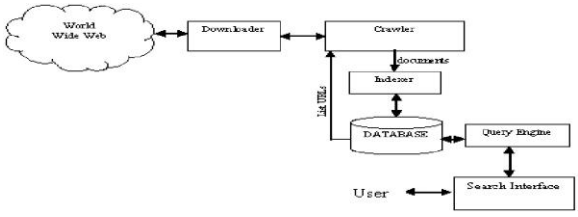

Fig. 1 : Architecture of a typical web search engine

2.1 Types of Search Engines

Search engines are good at finding unique keywords, phrases, quotes, and information contained in the full text of web pages. Search engines allow user to enter keywords and then search for them in its table followed by database. Various types [9] of search engines available are Crawler based search engines, Human powered directories, Meta search engines and Hybrid search engines.

i. **Crawler based search engines :**

Crawler based search engines create their listings (that make up the search engine's index or catalog) automatically with the help of web crawlers. It uses a computer algorithm to rank all pages retrieved. Such search engines are huge and often retrieve a lot of information. For complex searches, it allows to search within the results of previous search and enables us to refine search results. Such types of search engines contain full text of web pages they link to. Here, one can find pages by matching words in the pages user wants.

ii. Human powered directories :

Human powered directories are built by human selection i.e. they depend on humans to create repository. They are organized into subject categories and classification of pages is done by subjects. Such directories do not contain full text of the web page they link to, and are smaller than most search engines.

iii. Hybrid search engine :

Hybrid search engines differ from traditional text oriented and directory based search engines. These search engines typically favor one type of listing over the other. Many search engines today combine a crawler based search engine with a directory service.

iv. Meta search engines :

Meta search engines accumulate search and screen the results of multiple primary search engines. Unlike search engines, meta crawlers don't crawl the web themselves to build listings. Instead, they allow searches to be sent to several search engines all at once. The results are then blended together onto one page.

Based on the application for which search engines are used, they can be categorized as follows:-

i. Primary search engines scan entire sections of the www and produce their results from databases of web page content, automatically created by computers.
ii. Business and Services search engines essentially National yellow page directories.
iii. Employment and Job search engines either provide potential employers access to resumes of people interested in working for them or provide prospective employees with information on job availability.
iv. Finance-oriented search engines facilitate searches for specific information about companies (officers, annual reports etc.).
v. Image search engines help us search the www for images of all kinds.
vi. News search engines search newspaper and news web site archives for the selected information.
vii. People search engines search for names, addresses, telephone numbers and e-mail addresses.
viii. Subject guides are like indexes in the back of a book. They involve human intervention in selecting and organizing resources, so they cover fewer resources and topics but provide more focus and guidance.
ix. Specialized search engines search specialized databases, allow users to enter search terms in a particularly easy way, look for low prices on items they are interested in purchasing, and even give users access to real, live human beings to answer questions.

2.2 Design issues

Designing a search engine [7,8,9,10,11] is a tricky task and there are differences in the ways they work. In common they all perform three basic tasks; search the Internet based on important words, keep an index of the words they find and where they find them, and allow users to look for words (or combinations of words) found in that index.

There are different ways to improve performance of search engines but three main characteristics are improving algorithms to search the web, using filters towards the user's results; and improving the user interface for query input. Factors that determine the quality of a search engine are freshness of contents, index quality, search features, retrieval system and user behaviour. Few more issues are :-

i. Primary goals:

The primary goal of a search engine is to provide high quality search results over a rapidly growing world wide web. The most important appraise for a search engine is its search performance, quality of the results, ability to crawl and indexing the web efficiently.

ii. Diversity of documents:

On the web, documents are written in several different languages. Documents need to be indexed in a way which allows it to search for documents written in diverse languages with just one query. Search engines should be able to index documents written in multiple formats, as each file format provides certain difficulties for the search engines.

iii. Behaviour of web users:

Web users are very heterogeneous and search engines are used by professionals as well as by laymen. The search engine needs to be enough smart to cater for both types of users. Moreover, there is a tendency that users often only look at the results set that can be seen without scrolling, and the results which are not on first page are nearly invisible for the general user.

iv. Freshness of database :

Search engines find problems in keeping its database up-to-date with the entire web because of its enormous size and the different update cycles of individual websites.

Other characteristics [10] that a large search engine is expected to have are scalability, high performance, politeness, continuity, extensibility and portability.

III. Web Crawlers

Web crawler (Fig. 2) is a program that fetches information from www in an automated manner. The objective of a web crawler is to maintain freshness [6] of pages in its collection as high as possible. To download a document, a crawler starts by placing an initial set of seed URLs in a queue, where all URLs to be retrieved are kept and prioritized. From this queue the crawler extracts a URL, downloads the page, extracts URLs from the downloaded page, and places the new URLs in the queue. This process is repeated and the collected pages are used by a search engine. The browser parses the document and makes it available to the user.

The general algorithm of a web crawler is given below :
Begin

Read a URL from the set of seed URLs;
Determine the IP address for the host name;
Download the Robot.txt file which carries downloading permissions and also specifies the files to be excluded by the crawler;
Determine the protocol of underlying host like http, ftp, gopher etc.;

Based on the protocol of the host, download the document;
Identify the document format like doc, html, or pdf etc.;
Check whether the document has already been downloaded or not;
If the document is fresh one then
Read it and extract the links or references to the other cites from that documents;
else

Continue;
Convert the URL links into their absolute URL equivalents;
Add the URLs to set of seed URLs;
End.

A web crawler deals with two main issues, a good crawling strategy for deciding which pages to download next and, to have a highly optimized system architecture that can download a large number of pages per second. On other side it has to be robust against crashes, manageable, and considerate of resources and web servers.

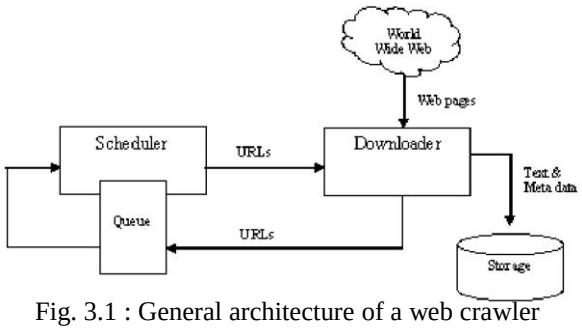

Fig. 3.1 : General architecture of a web crawler

Several available web crawling techniques are Parallel crawler, Distributed crawler, Focussed crawler, Hidden crawler, Incremental crawler, that differs in their mechanism, implementation and objective.

IV. Conclusion and future research

In this paper we conclude that designing a search engine to crawl the complete web is a nontrivial task. They have to be enough committed to perform the crawling process efficiently and reliably. Besides some design issues of a search engine, we also discussed that a web crawler supports the search engine to update its database to improve quality of contents retrieved by the user. From several available web crawling techniques, appropriate one may be used to crawl the web while designing a search engine.

References

[1]. A.K. Sharma, J. P. Gupta, D. P. Agarwal, "Augment Hypertext Documents suitable for parallel crawlers" , accepted for presentation and inclusion in the proceedings of WITSA-2003, a National workshop on Information Technology Services and Applications, Feb'2003, New Delhi.
[2]. Alexandros Ntoulas, Junghoo Cho, Christopher Olston, "What's New on the Web ? The Evolution of the Web from a Search Engine Perspective", In Proceedings of the World-Wide Web Conference (WWW), May 2004.
[3]. Arvind Arasu, Junghoo Cho, Hector Garcia-Molina, Andreas Paepcke, Sriram Raghavan, "Searching the Web", ACM Transactions on Internet Technology, 1(1): August 2001.
[4]. Baldi, Pierre, "Modeling the Internet and the Web: Probabilistic Methods and Algorithms", 2003.
[5]. Barry M. Leiner, Vinton G. Cerf, David D. Clark, Robert E. Kahn, Leonard, Kleinrock, Daniel C. Lynch, Jon
[6]. Postel, Larry G. Roberts, Stephen Wolff, "A Brief History of the Internet", www.isoc.org/internet/his tory,
[7]. Brian E. Brewington, George Cybenko, "How dynamic i s the web.", In Proceedings of the Ninth Internationa l World-Wide Web Conference, Amsterdam, Netherlands, May 2000.
[8]. Dirk Lewandowski, "Web searching, search engines an d Information Retrieval, Information Services & Use", 25 (2005) 137-147, IOS Press, 2005.
[9]. Franklin, Curt, "How Internet Search Engines Work", 2002, www.howstuffworks.com.
[10]. Grossan. B, "Search Engines : What they are, how th ey work, and practical suggestions for getting the most out of them", Feb97, webreference.com.
[11]. Heydon A., Najork M., "Mercator: A scalable, extens ible Web crawler.", World Wide Web, vol. 2, no. 4, pp. 2 19-229, 1999.
[12]. Mark Najork, Allan Heydon, "High- Performance Web Crawling", September 2001.
[13]. Mike, Burner, "Crawling towards Eternity : Building an archive of the World Wide Web", Web Techniques Magazine, 2(5), May 1997.
[14]. Niraj Singhal, Ashutosh Dixit, "Retrieving Informat ion from the Web and Search Engine Application", in proceedings of National Conference on "Emerging Trends in Software and Network Techniques– ETSNT'09", Amity University, Noida, India, Apr 2009 .
[15]. Sergey Brin, Lawrence Page, "The anatomy of a large - scale hyper textual Web search engine", Proceedings of the Seventh International World Wide Web Conference, pages 107-117, April 1998.