International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Web Crawler: Essential Component of Search Engine

Akshada K. Dhakade[1], Deepak C. Dhanwani[2]

[1]M.E. Istyear (CSE), P. R. Pote COE&M, Amravati, India

[2]Assistant Professor, Department of Computer Science and Engineering, P. R. Pote COE&M, Amravati, India

Abstract: ***With the vast growth of the Internet, many web pages are available online. Search engines use a component called as web crawlers for collecting these web pages from the web for storage and indexing. Many web pages are autonomous and are updated independent of the users. As the web pages are updated autonomously; users do not come to know of how often the sources change. Web crawler is the central part of the search engine which browses through the hyperlinks and stores the visited links for the future use. This paper represents concepts of web crawlers, its architecture and its various types.***

Keywords: Search Engine, Web Crawler, Crawler Policies, Techniques

1. Introduction

A **web search engine** is a software system that is designed to search for information on the World Wide Web. The search results are generally presented in a line of results often referred to as search engine results pages (SERPs). The information may be a mix of web pages, images, and other types of files. Some search engines also mine data available in databases or open directories. Unlike web directories, which are maintained only by human editors, search engines also maintain real-time information by running an algorithm on a web crawler.

In Figure 1, the search engine accepts the query from the user. An interface is provided by the search engine to the user so that users submit the queries. And it contains the mechanism for serving these queries. This is the only part which is visible to the end-users.

The database stores the data crawled by the web crawlers. The search engine queries the database so as to answer any user's request. The database also feeds the downloader with the URLs to be downloaded .The processor processes the URLs it takes from the downloader and updates the database with the fresh information (URLs) [1].

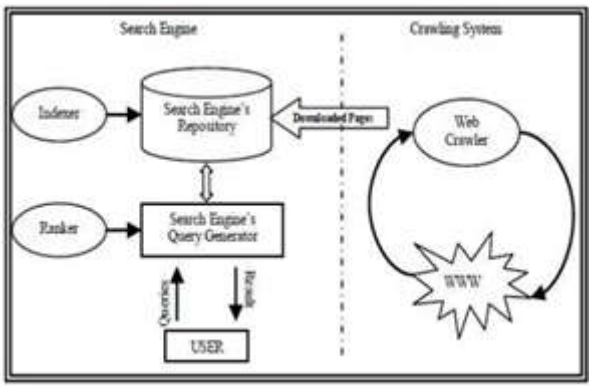

Figure1: Architecture of Search Engine

2. Web Crawler

Web crawler is an important method for collecting data and keeping up to date with the rapidly expanding Internet. A web crawler is a program, which automatically traverses the web by downloading documents and following links from page to page. It is a tool for the search engines and other information seekers to gather data for indexing and to enable them to keep their databases up to date. All search engines internally use web crawlers to keep the copies of data a fresh. Search engine is divided into different modules. Among those modules crawler module is the module on which search engine relies the most because it helps to provide the best possible results to the search engine. Crawlers are small programs that 'browse' the web on the search engine's behalf, similarly to how a human user would follow links to reach different pages [2].

A web crawler is a program/software or programmed script that browses the World Wide Web in a systematic, automated manner. The structure of the WWW is a graphical structure, i.e., the links presented in a web page may be used to open other web pages. Internet is a directed graph where webpage as a node and hyperlink as an edge, thus the search operation may be summarized as a process of traversing directed graph. By following the linked structure of the Web, web crawler may traverse several new web pages starting from a webpage. A web crawler move from page to page by the using of graphical structure of the web pages. Such programs are also known as robots, spiders, and worms. Web crawlers are designed to retrieve Web pages and insert them to local repository. Crawlers are basically used to create a replica of all the visited pages that are later processed by a search engine that will index the downloaded pages that help in quick searches. Search engines job is to storing information about several webs pages, which they retrieve from WWW [3].

3. Literature Review

Possibly the largest level study of Web page change was performed by Fetterly et al. They crawled 151 million pages

once a week for 11 weeks, and compared the modification across pages. Like Ntoulas et. al., they found a relatively small amount of change, with 65% of all page pairs remaining exactly the same. The study furthermore found that past change was a good judge of future change, this page length was correlated with change, and that the top level domain of a page was correlated with change. Cho and Garcia-Molina crawled around 720,000 pages once a day for a period of four months and seemed at how the pages changed. Ntoulas et. al. studied page change through weekly downloaded of 154 websites collected over a year. They found that a large number of pages did not modify according to a bags of words measure of similarity. Even for pages that did change, the changes were small. Frequency of change was not a big judge of the degree of change, but the degree of change was a good judge of the future degree of change. More recently, Olston and Panday crawled 10,000 random samples of URLs and 10,000 pages sampled from the Open Directory every second days for several months. Their analysis measured both change frequency and information longevity is the average lifetime of a shingle, and found only a moderate correlation between the two. They introduce new crawl policies that are aware to information longevity. In a study of changes examined via a proxy, Douglis et al. identified an association between re visitation rates and change. Hence, the study was limited to web content visited by a restricted population, and web pages were not aggressively crawled for changes among different visits.

Researchers have also peeped at how search results modify over time. The main focus in this study was on recognizing the dynamics of the consequences change and search engines has for searchers who want to return to previously visited pages. Junghoo Cho and Hector GarciaMolina proposed the design of an effective parallel crawler. The size of the Web grows at very fast speed, it becomes essential to parallelize a crawling process, to complete downloading pages in a reasonable amount of time. Author first proposes multiple architectures for a parallel crawler and then identifies basic issues related to parallel crawling. Based on this understanding, author then propose metrics to evaluate a parallel web crawler, and compare the proposed architectures using millions of pages collected from the Web. Rajashree Shettar, Dr. Shobha G presented a new model and architecture of the Web Crawler using multiple HTTP connections to WWW. The multiple HTTP connection is applied using multiple threads and asynchronous downloader part so that the overall downloading process is optimum. The user gives the initial URL from the GUI provided. It begins with a URL to visit. As the crawler visits the URL, it identifies all the hyperlinks available in the web page and appends them to the list of URLs to visit, known as the crawl frontier. URLs from the frontier is iteratively visited and it ends when it reaches more than five levels from every home pages of the websites visited and it is accomplished that it is not required to go deeper than five levels from the home page to capture most of the pages visited by the people while trying to retrieve information from the internet [3].

World Wide Web contains millions of information beneficial for the users, many information seekers usage search engine to initiate their Web activity. Every search engine rely on a crawler module to provide the grist for its operation, Matthew Gray wrote the first Crawler, the World Wide Web Wanderer, which was used from 1993 to 1996 . J. Cho. in describes various search techniques and how the search engines works by using crawler and in he has described how the search engines should cope with the evolving Web, in an attempt to provide users with up-to-date results. He has made the various studies on crawler policies. Proposes how one can maintain local copies of remote data sources “fresh," when the source data is updated autonomously and independently. Gautam Pant and Filippo Menczer examined the use of focused crawler. S.S. Dhenakaran1 and K. Thirugnana Sambanthan give an overview about Different types of Web crawler and the policies being used in the web crawlers and their evolution. Ms. Swati Mali and Dr. B.B. Meshram in implements effective multiuser personal web crawler where one user can manage multiple topics of interest. This type of web crawler can be configured to target precisely what user needs. It offers a high degree of control over the information that is returned for a particular search, vastly increasing the likelihood that it will be relevant. A crawler is a program that downloads and stores web pages often for a web search engine. The rapid growth of World Wide Web poses challenges to search for the most appropriate link. Author Pooja gupta and Mrs. Kalpana Johari has developed a Focused crawler using breadth-first search to extract only the relevant web pages of interested topic from the Internet [2].

4. Architecture of Web Crawler

A web crawler is one of the main components of the web search engines. The growth of web crawler is increasing in the same way as the web is growing. A list of URLs is available with the web crawler and each URL is called a seed. Each URL is visited by the web crawler. It identifies the different hyperlinks in the page and adds them to the list of URLs to visit. This list is termed as crawl frontier. Using a set of rules and policies the URLs in the frontier are visited individually. Different pages from the internet are downloaded by the parser and the generator and stored in the database system of the search engine. The URLs are then placed in the queue and later scheduled by the scheduler and can be accessed one by one by the search engine one by one whenever required. The links and related files which are being searched can be made available whenever required at later time according to the requirements. With the help of suitable algorithms web crawlers find the relevant links for the search engines and use them further. Databases are very big machines like DB2, used to store large amount of data [4].

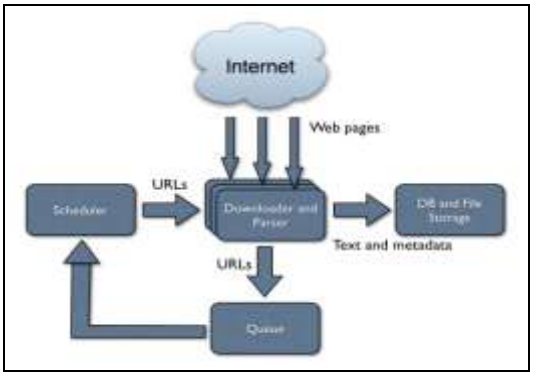

Figure 2: Architecture of Web Crawler

5. Crawler Policies

A Web crawler has various tasks and goals that must be handled carefully in spite of various contradictions amongst them. Also the various resources that are available must be used by web crawlers efficiently, also including network bandwidth which must exhibit a high degree of parallelism without affecting the web server by overloading. The behaviour of Web Crawler is the outcomes of combination of policies.

a) A selection policy that states which pages to download.
b) A re-visit policy that states when to check for changes to the pages.
c) A politeness policy that states how to avoid overloading Web sites.
d) A parallelization policy that states how to coordinate distributed Web crawlers [5].

6. Crawling Techniques

The crawling method used by various search engines in order to download pages that have already been downloaded and those that are yet to be downloaded relies greatly on various techniques such as follows:

6.1 Focused Crawler

Focused Crawler is the Web crawler that tries to download pages that are related to each other. It collects documents which are specific and relevant to the given topic. It is also known as a Topic Crawler because of its way of working .The focused crawler determines the following – Relevancy, Way forward. It determines how far the given page is relevant to the particular topic and how to proceed forward. The benefits of focused web crawler is that it is economically feasible in terms of hardware and network resources, it can reduce the amount of network traffic and downloads. The search exposure of focused web crawler is also huge [2].

6.2 Incremental Web Crawler

An incremental crawler is one, which updates an existing set of downloaded pages instead of restarting the crawl from scratch each time. This involves some way of determining whether a page has changed since the last time it was crawled. A crawler, which will continually crawl the entire web, based on some set of crawling cycles. An adaptive model is used, which uses data from previous cycles to decide which pages should be checked for updates, thus high freshness and results in low peak load is achieved [6].

6.3 Parallel Crawler

As the size of the Web grows, it becomes more difficult to retrieve the whole or a significant portion of the Web using a single process. Therefore, many search engines often run multiple processes in parallel to perform the above task, so that download rate is maximized. This type of crawler is known as a parallel crawler [6].

6.4 Distributed Crawler

Distributed web crawling is a distributed computing technique. Many crawlers are working to distribute in the process of web crawling, in order to have the most coverage of the web. A central server manages the communication and synchronization of the nodes, as it is geographically distributed. It basically uses Page rank algorithm for its increased efficiency and quality search. The benefit of distributed web crawler is that it is robust against system crashes and other events, and can be adapted to various crawling applications [2].

7. Conclusion

Web Crawler is the vital source of information retrieval which traverses the Web and downloads web documents that suit the user's need. Web crawler is used by the search engine and other users to regularly ensure that their database is up-to-date. Web crawlers are an important aspect of all the search engines. They are the basic component of all the web services so they need to provide high performance. Data manipulation by the web crawlers covers a wide area. Building an effective web crawler to solve different purposes is not a difficult task, but choosing the right strategies and building an effective architecture will lead to implementation of highly intelligent web crawler application.

8. Future Scope

Already a lot of research is going on in the field of web data extraction techniques. In future work can be done to improve the efficiency of algorithms. Also, the accuracy and timeliness of the search engines can also be improved. The work of the different crawling algorithms can be extended further in order to increase the speed and accuracy of web crawling. A major open issue for future work about the scalability of the system and the behavior of its components. This could probably be best done by setting up a simulation test bed, consisting of several workstations, that simulates the web using either artificially generated pages or a stored partial snapshot of the web.

References

[1] Komal Sachdeva and Ashutosh Dixit, Ph.D, “Estimating Page Importance based on Page Accessing Frequency”, International Journal of Computer Applications (0975 – 8887) Volume 86 – No 5, January 2014
[2] Trupti V. Udapure, Ravindra D. Kale, Rajesh C. Dharmik, “Study of Web Crawler and its Different Types”, IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 1, Ver. VI (Feb. 2014)
[3] Md. Abu Kausar, V. S. Dhaka, Sanjeev Kumar Singh, “Web Crawler: A Review”, International Journal of Computer Applications (0975 – 8887), Volume 63– No.2, February 2013
[4] Mini Singh Ahuja, Dr Jatinder Singh Bal, Varnica, “Web Crawler: Extracting the Web Data”, International Journal of Computer Trends and Technology (IJCTT) – volume 13 number 3 – Jul 2014

[5] Raja Iswary and Keshab Nath, "WEB CRAWLER", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 10, October 2013
[6] Dhiraj Khurana, Satish Kumar, ―"Web Crawler: A Review", IJCSMS International Journal of Computer Science & Management Studies, Vol. 12, Issue 01, January 2012 ISSN (Online): 2231 –5268.

Author Profile

Akshada K. Dhakade. Currently she is pursuing M.E. in Computer Science and Engineering from P. R. Pote College of Engineering And Management Amravati, Maharashtra, India.

Prof. Deepak C. Dhanwani. Completed his M.E. Currently he is working as assistant professor in P. R. Pote College of Engineering and Management Amravati, Maharashtra, India.