ISSN No:-2456-2165

Web Crawler: A Survey

S. S. Bhamare
School of Computer Sciences
Kavayitri Bahinabai Chaudhari North Maharashtra University
Jalgaon (M.S) India

Abstract:- In world wide web, Web Crawler is working as a software agent which helps in web indexing. Web crawler at times called as spider or internet bot explicitly operated by many search engines. In Information Retrieval to colleting an information Web crawler play an important role. For effective searching and web indexing is mostly depends on web crawlers. General Purpose, Distributed and Focused crawling are the types of crawling techniques of web crawling. This paper discussed overall survey on web crawlers, its types and working of web crawlers.

Keywords:- *Web Crawler, Crawling techniques, WWW, Search engine.*

I. INTRODUCTION

World Wide Web is "a wide-area hypermedia information retrieval initiative aiming to give universal access to a large universe of documents." In simpler terms, the Web is a computer network that allows users of one computer to access information stored on another through the world-wide network called the Internet [22]. The Web's application follows a standard client-server model. Currently, the World Wide Web (WWW) contains massive amounts of web pages which are accessible by users. This provides a very useful and helpful means of collecting information.

Web Search is the most important application on the World Wide Web. It is based on information retrieval (IR), which is an area of research that enables the user to discover essential information from a large collection of text documents. An IR system discovers a set of documents that is associated to the query from its underlying collection. In information retrieval through search engines, web crawler plays significant role for searching and indexing the web documents.

II. WEB CRAWLER AND ITS WORKING

In last decade World Wide Web expand very rapidly makes it the largest widely accessible data source in the world. The amount of data/information on the Web is huge and is still growing.

Most of the information on the Web is linked. There are Hyperlinks among various Web pages within and across different web sites. Within a site, hyperlinks manage information structures. Implicit transmission of authority to the focus pages is represented by hyperlinks in various sites

In Web, generally Web pages formatted in HTML (i.e. Hyper Text Markup Language) are found on this network of computers. Contains of web pages such as pictures, videos and other online content can be accessed through a different Web browser.

Web crawling supports information retrieval for automatic retrieval of appropriate documents, while at the same time it makes sure that retrieval of irrelevant documents is also avoided.

The main aim of web crawler is to make content index of web sites through the Internet thus these websites can be seen in results of search engine. Web crawler, or spider is typically functioned by search engines like Google and Bing etc.

Generally, crawling activity of web crawlers begin by downloading the robot.txt file of websites. This text file consists of sitemaps that list the URLs of web site the search engine can crawl. Once web crawlers start crawling a web page, new web pages are discovering through links. All these crawlers contain newly discovered URLs to the crawl queue so that they can be crawled later and index every individual page that is connected to others by these Web crawlers.

Web pages change frequently, and it is also necessary to identify exactly how search engines should crawl them. To determine factors various algorithms are used by Web Search engine crawlers like how frequently an existing page would be recrawled and how many number of web pages on a web site would be indexed.

III. WEB CRAWLING TECHNIQUES

Mostly Web Crawlers are used following web crawling techniques:

A. General Purpose Crawling

In this general-purpose crawling techniques Web Crawler gathers all the pages from a specific collection of URLs and their links. To fetch many pages from different locations this technique helps to the crawler. Due to fetching all the pages this technique can slow down the speed and network bandwidth.

B. Focused Crawling

In focused crawler is used to collect documents only on a particular or focus topic. This technique reduces amount of network traffic and downloads. Main objective of this focused crawler is for search only selective web pages that are appropriate to a particular set of problems. The related sections only crawls of the web and have an advantage to considerable savings of hardware and network resources.

C. Distributed Crawling

In distributed crawling, it allows users to extend their individual computing and bandwidth resources to crawling web pages by increasing the load of these tasks across many computers. Several methods are used to crawl and download pages from the Web [20].

IV. LITERATURE SURVEY

In literature survey, there exist few research work on web crawler carried out by many researchers to improve the effectiveness of web crawler,

[2] The key objective of the paper is to extract data from websites using provided hyperlinks. This extracted data are mainly unstructured data. In the end, the authors show a difference between TF-IDF algorithm and the BFS algorithm to show the accuracy rate suggesting TF-IDF algorithm offers more precise results.

[3] New web crawler was implemented that is also called a web spider, that browses the web in a methodical manner to gather information. This system was designed, developed, and implemented using python. To develop this system an algorithmic program is designed for implemented the developed module on the required system.

[6] are suggested the design of an efficient parallel crawler. Due to size of web increases rapidly every day. It is necessary to make a crawling process parallelize. In a sufficient amount of time, it supports to complete downloading web pages. Researchers propose system of measurement to evaluate a parallel web crawler and compare the proposed architectures using lots of pages collected from the Web.

[7] proposed a novel model and its design of the Web Crawler using multiple HTTP connections to World Wide Web. It begins with a URL, Crawler visits the URL and it detects all the hyperlinks available in the web page and adds them to the list of URLs to visit, known as the crawl frontier. Up to the level five from every home page of web sites URLs is repeatedly visited and then stops while trying to retrieve information from the internet.

[9] proposed design architecture for execute multiple crawling processes (C-procs) as a parallel crawler. Each process executes the dynamic tasks as a single process crawler executes. By downloading pages from the World Wide Web, it stores the web pages locally and extracts URLs from web pages by following their hyperlinks. The crawling process executing these tasks may be over on the same local network or at geographically remote locations.

[10] proposed and developed new web Crawler called PyBot which is based on algorithm standard Breadth-First Search strategy. Primarily this crawler takes a URL and it's all hyperlinks. From that page hyperlinks, crawler crawls again till to the point that have a no other hyperlinks found. This crawler crawls and downloads all the Web Pages. It uses download web pages and web structure in Excel CSV format for the ranking pages, page rank algorithm used for produces ranking order of pages with page list.

[11] present a new Genetic PageRank Algorithms, is a search and optimization technique which is used in computing to find optimal solutions.

[12] proposed a method for detecting web crawlers in real time. Researchers use decision trees to categorize requests in real time, as start from a crawler or human, while their session is ongoing.

[13] offered a method to correctly determine the quality of hyperlinks that is not retrieved but this hyperlink is accessible to them. So that researcher applies an algorithm such as AntNet routing algorithm.

[14] present a method which employs mobile agents to crawl the web pages. These mobile crawlers filter out any unnecessary data locally before moved it back to the search engine. These mobile crawlers decrease the network load by reducing the amount of data transmitted over the network.

[20] discussed different existing research work on Web Crawlers and its techniques has been carried out by various researchers.

V. CHALLENGES AND ISSUES

The main challenge is increasing size of web. Every day large number of web pages emerging on web and crawling of this large size of web is new challenge for the web crawlers. The major challenge is also associated with crawling multimedia data, Web crawler execution time and scaling of the web size.

Many methods and options have been designed and developed according to above challenges. Due to rate of increase of data on web makes web crawling a more challenging and difficult task. Here is need that continuously updating in web crawling algorithms to match up with the requiring data from users and increasing data on web.

VI. CONCLUSION

Web crawlers plays an important role for all search engines which helps to search web data. To develop an efficient web crawler that match with todays need is not difficult task. Developing such web crawler with proper approach and architecture leads to implement a smart web crawler for smart searching. Many researchers developed various web crawler algorithms for searching but smart web crawler algorithms need to implement for better results and high performance.

REFERENCES

[1.] D. Debraj and P. Das, "Study of deep web and a new form-based crawling technique," International Journal of Computer Engineering and Technology (IJCET), Vol. 7, No. 1, pp. 36-44, 2016.
[2.] Ahmed, Tanvir & Chung, Mokdong ―Design and application of intelligent dynamic crawler for web data mining, Korea Multimedia Society, Spring Conference 2019.
[3.] F. M. Javed Mehedi Shamrat, Zarrin Tasnim, A.K.M Sazzadur Rahman, Naimul Islam Nobel, Syed Akhter

Hossain An Effective Implementation Of Web Crawling Technology To Retrieve Data From The World Wide Web (Www) International Journal Of Scientific & Technology Research Volume 9, Issue 01, January 2020 Issn 2277-8616 1252 Ijstr©2020 www.ijstr.org

[4.] Z. Guojun, J. Wenchao, S. Jihui, S. Fan, Z. Hao, L. Jiang, et al., "Design and application of intelligent dynamic crawler for web data mining," Proceeding of 2017 32nd Youth Aca demic Annual Conference of Chinese Associ ation of Automation (YAC) IEEE , pp. 1098-1105, 2017.

[5.] Berners-Lee, Tim, "The World Wide Web: Past, Present and Future", MIT USA, Aug 1996, available at: http://www.w3.org/People/Berners-Lee/1996/ppf.html.

[6.] Junghoo Cho and Hector Garcia-Molina "Parallel Crawlers". Proceedings of the 11th international conference on World Wide Web WWW '02", May 7–11, 2002, Honolulu, Hawaii, USA. ACM 1-58113-449-5/02/0005.

[7.] Rajashree Shettar, Dr. Shobha G, "Web Crawler On Client Machine", Proceedings of the International MultiConference of Engineers and Computer Scientists 2008 Vol II IMECS 2008, 19-21 March, 2008, Hong Kong

[8.] K. Sharma, J.P. Gupta and D. P. Agarwal "PARCAHYD: An Architecture of a Parallel Crawler based on Augmented Hypertext Documents", International Journal of Advancements in Technology, pp. 270-283, October 2010.

[9.] Shruti Sharma, A.K.Sharma and J.P.Gupta "A Novel Architecture of a Parallel Web Crawler", International Journal of Computer Applications (0975 – 8887) Volume 14– No.4, pp. 38-42, January 2011

[10.] Alex Goh Kwang Leng, Ravi Kumar P, Ashutosh Kumar Singh and Rajendra Kumar Dash "PyBot: An Algorithm for Web Crawling", IEEE 2011

[11.] Lili Yana, Zhanji Guia, Wencai Dub and Qingju Guoa "An Improved PageRank Method based on Genetic Algorithm for Web Search", Procedia Engineering, pp. 2983-2987, Elsevier 2011

[12.] Andoena Balla, Athena Stassopoulou and Marios D. Dikaiakos (2011), "Real-time Web Crawler Detection", 18th International Conference on Telecommunications, pp. 428-432, 2011

[13.] Bahador Saket and Farnaz Behrang "A New Crawling Method Based on AntNet Genetic and Routing Algorithms", International Symposium on Computing, Communication, and Control, pp. 350-355, IACSIT Press, Singapore, 2011

[14.] Anbukodi.S and Muthu Manickam.K "Reducing Web Crawler Overhead using Mobile Crawler", PROCEEDINGS OF ICETECT, pp. 926-932, 2011

[15.] K. S. Kim, K. Y. Kim, K. H. Lee, T. K. Kim, and W. S. Cho "Design and Implementation of Web Crawler Based on Dynamic Web Collection Cycle", pp. 562-566, IEEE 2012

[16.] MetaCrawler Search Engine, available at: http://www.metacrawler.com.

[17.] Cho, J. and H. Garcia-Molina. The evolution of the Web and implications for an incremental crawler. VLDB '00, 200-209, 2000.

[18.] Douglis, F., A. Feldmann, B. Krishnamurthy, and J. Mogul. Rate of change and other metrics: A live study of the World Wide Web. USENIX Symposium on Internet Technologies and Systems, 1997.

[19.] Fetterly, D., M. Manasse, M. Najork, and J. Wiener. A large-scale study of the evolution of Web pages. WWW '03, 669-678, 2003.

[20.] 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.

[21.] Kim, J. K., and S. H. Lee. An empirical study of the change of Web pages. APWeb '05, 632-642, 2005.

[22.] The World-Wide Web, Henrik Frystyk, July 1994 https://www.w3.org/People/Frystyk/thesis/WWW.html