# Web Crawler: A Review

Md. Abu Kausar
Dept. of Computer & System Sciences
Jaipur National University, Jaipur, India

V. S. Dhaka
Dept. of Computer & System Sciences
Jaipur National University, Jaipur, India

Sanjeev Kumar Singh
Dept. of Mathematics
Galgotias University, Gr. Noida, India

## ABSTRACT

Information Retrieval deals with searching and retrieving information within the documents and it also searches the online databases and internet. Web crawler is defined as a program or software which traverses the Web and downloads web documents in a methodical, automated manner. Based on the type of knowledge, web crawler is usually divided in three types of crawling techniques: General Purpose Crawling, Focused crawling and Distributed Crawling. In this paper, the applicability of Web Crawler in the field of web search and a review on Web Crawler to different problem domains in web search is discussed.

## Keywords

WWW, Web Crawler, Crawling techniques, Web Crawler Survey, Search engine, Parallel Crawler.

## 1. INTRODUCTION

The World Wide Web (WWW) is internet client server architecture. It is a powerful system based on complete autonomy to the server for serving information available on the internet. The information is arranged as a large, distributed, and non-linear text system known as Hypertext Document system. These systems define part of a document as being hypertext- pieces of text or images which are linked to other documents via anchor tags. HTTP and HTML present a standard way of retrieving and presenting the hyperlinked documents. Internet browsers, use search engines to explore the servers for required pages of information. The pages send by the servers are processed at the client side.

Now days it has become an important part of human life to use Internet to gain access the information from WWW. The current population of the world is about 7.049 billion out of which 2.40 billion people (34.3%) use Internet [3] (see Figure 1). From .36 billion in 2000, the amount of Internet users has increased to 2.40 billion in 2012 i.e., an increase of 566.4% from 2000 to 2012. In Asia out of 3.92 billion people, 1.076 billion (i.e.27.5%) use Internet, whereas in India out of 1.2 billion, .137 billion (11.4%) use Internet. Same growth rate is expected in future too and it is not far away when one will start thinking that life is incomplete without Internet. Figure 1: illustrates Internet Users in the World by Geographic Regions.

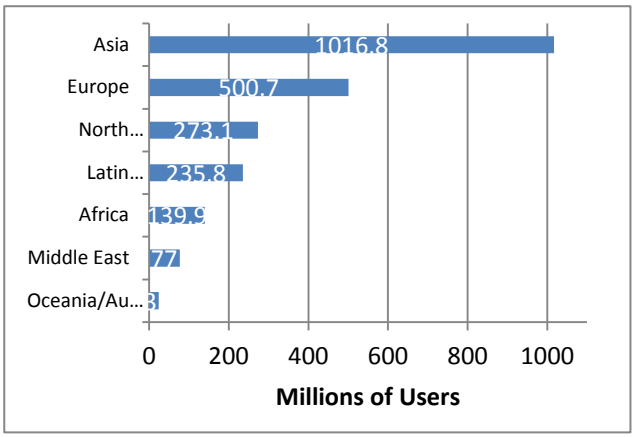


**Figure 1: Internet Users in the World by Geographic Regions (Source: http://www.internetworldstats.com accessed on May 7, 2012)**

Beginning in 1990, World Wide Web has grown exponentially in size. As of today, it is estimated that it contains about 55 billion publicly index able web documents [4] spread all over the world on thousands of servers. It is not easy to search information from such a vast collection of web documents available on WWW. It is not sure that users will be able to retrieve information even after knowing where to look for information by knowing its URLs as Web is continuously changing. Information retrieval tools are divided into three categories as follow:

a) Web directories

b) Meta search engines

**c)** Search engines

## 2. WEB CRAWLER

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. These pages are retrieved by a Web crawler that is an automated Web browser that follows each link it sees.

## 2.1 The History of Web Crawler

The first Internet "search engine", a tool called "Archie" — shortened from "Archives", was developed in 1990 and downloaded the directory listings from specified public anonymous FTP (File Transfer Protocol) sites into local files, around once a month [5], [6]. In 1991, "Gopher" was created, that indexed plain text documents. "Jughead" and "Veronica" programs are helpful to explore the said Gopher indexes [7], [8], [9], [10]. With the introduction of the World Wide Web in 1991 [11], [12] numerous of these Gopher sites changed to web sites that were properly linked by HTML links. In the year 1993, the "World WideWebWanderer" was formed the first crawler [13]. Although this crawler was initially used to measure the size of the Web, it was later used to retrieve URLs that were then stored in a database called Wandex, the first web search engine [14]. Another early search engine, "Aliweb" (Archie-Like Indexing for the Web) [15] allowed users to submit the URL of a manually constructed index of their site.

The index contained a list of URLs and a list of user wrote keywords and descriptions. The network overhead of crawlers initially caused much controversy, but this issue was resolved in 1994 with the introduction of the Robots Exclusion Standard [16] which allowed web site administrators to block crawlers from retrieving part or all of their sites. Also, in the year 1994, "WebCrawler" was launched [17] the first "full text" crawler and search engine. The "WebCrawler" permitted the users to explore the web content of documents rather than the keywords and descriptors written by the web administrators, reducing the possibility of confusing results and allowing better search capabilities. Around this time, commercial search engines began to appear with [18], [19], [20], [21], [22], [23], [24] and [25] being launched from 1994 to 1997 [26]. Also introduced in 1994 was Yahoo! , a directory of web sites that was manually maintained, though later incorporating a search engine. During these early years Yahoo! and Altavista maintained the largest market share [26]. In 1998 Google was launched, quickly capturing the market [26]. Unlike many of the search engines at the time, Google had a simple uncluttered interface, unbiased search results that were reasonably relevant, and a lower number of spam results [27]. These last two qualities were due to Google's use of the PageRank [28] algorithm and the use of anchor term weighting [29].

While early crawlers dealt with relatively small amounts of data, modern crawlers, such as the one used by Google, need to handle a substantially larger volume of data due to the dramatic enhance in the amount of the Web.

## 2.2 Working of Web Crawler

The working of Web crawler is beginning with initial set of URLs known as seed URLs. They download web pages for the seed URLs and extract new links present in the downloaded pages. The retrieved web pages are stored and well indexed on the storage area so that by the help of these indexes they can later be retrieved as and when required. The extracted URLs from the downloaded page are confirmed to know whether their related documents have already been downloaded or not. If they are not downloaded, the URLs are again assigned to web crawlers for further downloading. This process is repeated till no more URLs are missing for downloading. Millions of pages are downloaded per day by a crawler to complete the target. Figure 2 illustrates the crawling processes.

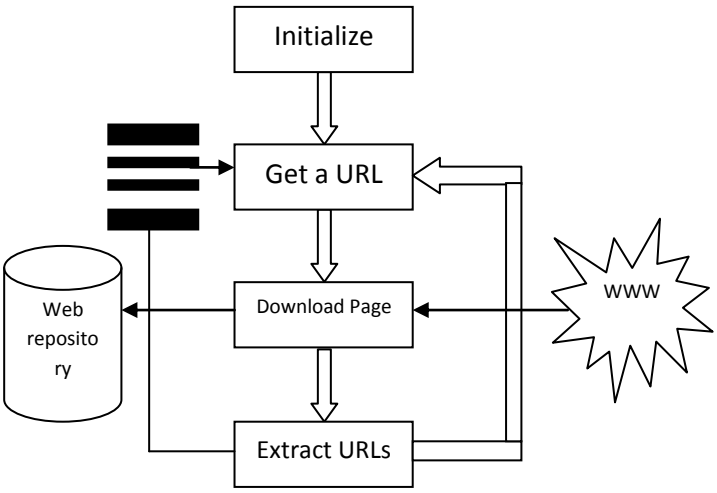


**Figure 2: Flow of a crawling process**

The working of a web crawler may be discussed as follows:

- Selecting a starting seed URL or URLs
- Adding it to the frontier
- Now picking the URL from the frontier
- Fetching the web-page corresponding to that URL
- Parsing that web-page to find new URL links
- Adding all the newly found URLs into the frontier
- Go to step 2 and reiterate till the frontier is empty

Thus a web crawler will recursively keep on inserting newer URLs to the database repository of the search engine. So we can see that the major function of a web crawler is to insert new links into the frontier and to choose a fresh URL from the frontier for further processing after every recursive step.

# 3. CRAWLING TECHNIQUES

There are a few crawling techniques used by Web Crawlers, mainly used are:

*A. General Purpose Crawling*

A general purpose Web Crawler collects as many pages as it can from a particular set of URL's and their links. In this, the crawler is able to fetch a large number of pages from different locations. General purpose crawling can slow down the speed and network bandwidth because it is fetching all the pages.

*B. Focused Crawling*

A focused crawler is designed to collect documents only on a specific topic which can reduce the amount of network traffic and downloads. The purpose of the focused crawler is to selectively look for pages that are appropriate to a pre-defined set of matters. It crawl only the relevant regions of the web and leads to significant savings in hardware and network resources.

*C. Distributed Crawling*

In distributed crawling, multiple processes are used to crawl and download pages from the Web.

# 4. PARALLEL CRAWLER

Now search engines do not depend on a single but on multiple web crawlers that run in parallel to complete the target. While functioning in parallel, crawlers still face many challenging difficulties such as overlapping, quality and network. Given below Figure illustrates the flow of multiple crawling processes.

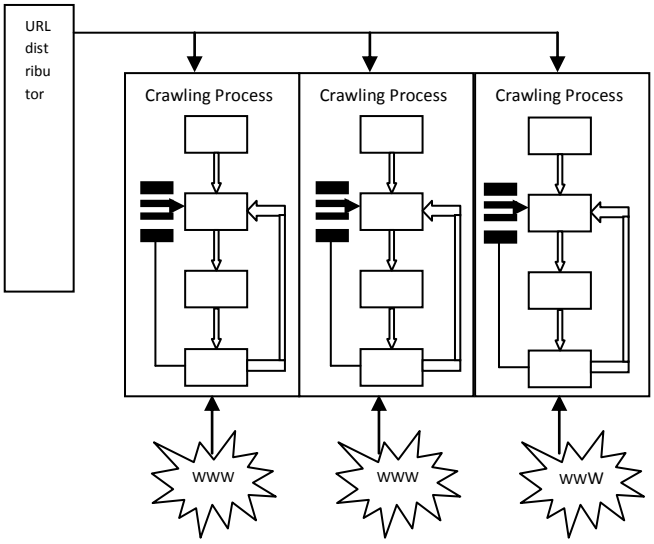


**Figure 3: Flow of multiple crawling processes.**

# 5. LITERATURE SURVEY

Possibly the largest level study of Web page change was performed by Fetterly et al. [46]. They crawled 151 million pages once a week for 11 weeks, and compared the modification across pages. Like Ntoulas et. al. [50], 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. Describing the amount of change on the Web has been of significant interest to researchers [44], [46], [47], [48], [49], [50], [52]. Cho and Garcia-Molina [44] crawled around 720,000 pages once a day for a period of four months and seemed at how the pages changed. Ntoulas et. al. [50] 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 [51] crawled 10,000 random samples of URLs and 10,000 pages sampled from the OpenDirectory 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. [45] 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 [53], [54]. 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 Garcia-Molina [30] 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 [31] 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. Eytan Adar et. al [32] described algorithms, analyze, and models for characterizing the evolution of Web content. Proposed analysis gives insight into how Web content changes on a finer grain than previous study, both in terms of the time intervals studied and the detail of change analyzed. A. K. Sharma et. al. [33] Parallelization of crawling system is necessary for downloading documents in a reasonable amount of time. The work has done reported here to focuses on providing parallelization at three levels: the document, the mapper, and the crawl worker level. The bottleneck at the document level has been removed. The efficacy of DF (Document Fingerprint) algorithm and the efficiency of volatile information has been tested and verified. This paper specifies the major components of the crawler and their algorithmic detail. Ashutosh Dixit et. al. [34] developed a mathematical model for crawler revisit frequency. This model ensures that frequency of revisit will increase with the change frequency of page up to the middle threshold value after that up to the upper threshold value remains same i.e., unaffected by the change frequency of page but after the upper threshold value it starts reducing automatically and settles itself to lower threshold. Shruti Sharma et. al. [35] present architecture for a parallel crawler which includes multiple crawling processes; called C-procs. Each C-proc performs the vital tasks that a single process crawler performs. It downloads pages from the WWW, stores the pages locally, extracts URLs from them and follows their links. The C-proc's executing these tasks may be spread either on the same local network or at geographically remote locations. Alex Goh Kwang Leng et. al. [36] Developed algorithm which uses the standard Breadth-First Search strategy to design and develop a Web Crawler called PyBot. Initially it takes a URL and from

that URL, it gets all the hyperlinks. From the hyperlinks, it crawls again until a point that no new hyperlinks are found. It downloads all the Web Pages while it is crawling. PyBot will output a Web structure in Excel CSV format on the website it crawls. Both downloaded pages and Web structure in Excel CSV format are stored in storage and are used for the ranking. The ranking systems take the Web structure in Excel CSV format and apply the PageRank algorithm and produces ranking order of the pages by displaying the page list with most popular pages at the top. Song Zheng [37] Proposed a new focused crawler analysis model based on the genetic and ant algorithms method. The combination of the Genetic Algorithm and Ant Algorithm is called the Genetic Algorithm-Ant Algorithm whose basic idea is to take advantages of the two algorithms to overcome their shortcomings. The improved algorithm can gets higher recall rate. Lili Yan et. al. [38] Proposed Genetic Pagerank Algorithms. A genetic algorithm (GA) is a search and optimization technique which is used in computing to find optimum solutions. Genetic algorithms are categorized as global search heuristics. Andoena Balla et. al. [39] presents a method for detecting web crawlers in real time. Author use decision trees to categorize requests in real time, as beginning from a crawler or human, while their session is ongoing. For this purpose author used machine learning techniques to recognize the most vital features that distinguish humans from crawlers. The technique was tested in real time with the help of an emulator, using only a small number of requests. Results show the effectiveness and applicability of planned approach. Bahador Saket and Farnaz Behrang [40] presented a technique to determine correctly the quality of links that have not been retrieved so far but a link is accessible to them. For this reason author apply an algorithm like an AntNet routing algorithm. To avoid local search difficulty, author recommended a method which is based on genetic algorithms (GA). In this technique, address of some pages is given to crawler and their associated pages are retrieved and the first generation is created. In selection task, the degree of relationship among the pages and the specific topic is studied and each page is given a special score. Pages whose scores exceed a definite amount is selected and saved and other pages are discarded. In cross-over task, the links of current generation pages are extracted. Each link is given a unique score depending on the pages in which link is placed. After that a previously determined number of links will be selected randomly and the related pages will retrieve and new generation is created. Anbukodi.S and Muthu Manickam.K [41] proposed approach which employs mobile agents to crawl the pages. Mobile agent is created, sent, finally received and evaluated in its owner's home context. These mobile crawlers are transferred to the site of the source where the data reside to filter out any unnecessary data locally before transported it back to the search engine. These mobile crawlers can decrease the network load by reducing the quantity of data transmitted over the network. Using this approach filter those web pages that are not modified using mobile crawlers but retrieves only those web pages from the remote servers that are actually modified and perform the filtering of non-modified pages without downloading the pages. Their migrating crawlers shift to the web servers, and carry out the downloading of web documents, processing, and extraction of keywords. After compressing, transfer the results back to the central search engine. K. S. Kim et. al. [42] proposed a dynamic web-data crawling techniques, which contain sensitive inspection of web site changes, and dynamic retrieving of pages from target sites. Authors develop an optimal collection cycle model according the update characteristics of the web contents. The model dynamically predicts collection cycle of the web contents by calculating web collection cycle score.

# 6. CONCLUSION

The Internet and Intranets have brings a lots of information. People usually have the option to search engines to find necessary information. Web Crawler is thus vital information retrieval which traverses the Web and downloads web documents that suit the user's need. 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 which are later processed by a search engine that will index the downloaded pages that help in quick searches. The major objective of the review paper is to throw some light on the web crawling previous work. This article also discussed the various researches related to web crawler.

# 7. REFERENCES

[1] 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.

[2] Berners-Lee, Tim, and Cailliau, CN, R., "Worldwide Web: Proposal for a Hypertext Project" CERN October 1990, available at: http://www.w3.org/Proposal.html.

[3] "Internet World Stats. Worldwide internet users", available at: http://www.internetworldstats.com (accessed on May 5, 2012).

[4] Maurice de Kunder, "Size of the World Wide Web", Available at: http://www.worldwidewebsize.com (accessed on May 5, 2012).

[5] P. J. Deutsch. Original Archie Announcement, 1990. URL http://groups.google.com/group/comp.archives/msg/a77343f9175b24c3?output=gplain.

[6] A. Emtage and P. Deutsch. Archie: An Electronic Directory Service for the Internet. In roceedings of the Winter 1992 USENIX Conference, pp. 93–110, San Francisco, California, USA, 1991.

[7] G. S. Machovec. Veronica: A Gopher Navigational Tool on the Internet. Information Intelligence, Online Libraries, and Microcomputers, 11(10): pp. 1–4, Oct. 1 1993. ISSN 0737-7770.

[8] R. Jones. Jughead: Jonzy's Universal Gopher Hierarchy Excavation And Display. unpublished, Apr. 1993.

[9] J. Harris. Mining the Internet: Networked Information Location Tools: Gophers, Veronica, Archie, and Jughead. Computing Teacher, 21(1):pp. 16–19, Aug. 1 1993. ISSN 0278-9175.

[10] H. Hahn and R. Stout. The Gopher, Veronica, and Jughead. In The Internet Complete Reference, pp. 429–457. Osborne McGraw-Hill, 1994.

[11] T. Berners-Lee, R. Cailliau, J. Groff, and B. Pollermann. World-Wide Web: The Information Universe. Electronic Networking: Research, Applications and Policy, 1(2): pp. 74–82, 1992. URL http://citeseer.ist.psu.edu/berners-lee92worldwide.html.

[12] T. Berners-Lee. W3C, Mar. 2008. URL http://www.w3.org/.

[13] M. K. Gray. World Wide Web Wanderer, 1996b. URL http://www.mit.edu/people/mkgray/net/.

[14] W. Sonnenreich and T. Macinta. Web Developer.com, Guide to Search Engines. John Wiley & Sons, New York, New York, USA, 1998.

[15] M. Koster. ALIWEB - Archie-Like Indexing in the WEB. Computer Networks and ISDN Systems, 27(2): pp. 175–182, 1994a. ISSN 0169-7552. doi: http://dx.doi.org/10.1016/0169-7552(94)90131-7.

[16] M. Koster. A Standard for Robot Exclusion, 1994b. URL http://www.robotstxt.org/wc/norobots.html. http://www.robotstxt.org/wc/exclusion.html.

[17] B. Pinkerton. Finding What People Want: Experiences with the WebCrawler. In Proceedings of the Second International World Wide Web Conference, Chicago, Illinois, USA, Oct. 1994.

[18] Infoseek, Mar. 2008. URL www.infoseek.co.jp

[19] Lycos, Mar. 2008. URL http://www.lycos.com

[20] Altavista, Mar. 2008. URL www.altavista.com

[21] Excite, Mar. 2008. URL www.excite.com

[22] Dogpile, Mar. 2008. URL www.dogpile.com

[23] Inktomi, Mar. 2008. URL www.inktomi.com

[24] Ask.com, Mar. 2008. URL http://ask.com/.

[25] Northern Light, Mar. 2008. URL http://www.northernlight.com

[26] D. Sullivan. Search Engine Watch: Where are they now? Search Engines we've Known & Loved, Mar. 4 2003b. URL http://searchenginewatch.com/sereport/article.php/2175241.

[27] Google. Google's New GoogleScout Feature Expands Scope of Search on the Internet, Sept. 1999. URL http://www.google.com/press/pressrel/pressrelease4.html.

[28] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998. URL http://citeseer.ist.psu.edu/page98pagerank.html

[29] S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In P. H. Enslow Jr. and A. Ellis, editors, WWW7: Proceedings of the Seventh International Conference on World Wide Web, pp. 107–117, Brisbane, Australia, Apr. 14–18 1998. Elsevier Science Publishers B. V., Amsterdam, The Netherlands. doi: http://dx.doi.org/10.1016/S0169-7552(98)00110-X.

[30] 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.

[31] 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

[32] Eytan Adar, Jaime Teevan, Susan T. Dumais and Jonathan L. Elsas "The Web Changes Everything: Understanding the Dynamics of Web Content", ACM 2009.

[33] A. 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.

[34] Ashutosh Dixit and Dr. A. K. Sharma, "A Mathematical Model for Crawler Revisit Frequency", IEEE 2nd International Advance Computing Conference, pp. 316-319, 2010.

[35] 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

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

[37] Song Zheng, "Genetic and Ant Algorithms Based Focused Crawler Design", Second International Conference on Innovations in Bio-inspired Computing and Applications pp. 374-378, 2011

[38] 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

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

[40] 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

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

[42] 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

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

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

[45] 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.

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

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

[48] Koehler, W. Web page change and persistence: A four-year longitudinal study. JASIST, 53(2), 162-171, 2002.

[49] Kwon, S. H., S. H. Lee, and S. J. Kim. Effective criteria for Web page changes. In Proceedings of APWeb ’06, 837-842, 2006.

[50] Ntoulas, A., Cho, J., and Olston, C. What’s new on the Web? The evolution of the Web from a search engine perspective. WWW ’04 , 1-12, 2004.

[51] Olston, C. and Pandey, S. Recrawl scheduling based on information longevity. WWW ’08, 437-446, 2008.

[52] Pitkow, J. and Pirolli, P. Life, death, and lawfulness on the electronic frontier. CHI ’97, 383-390, 1997.

[53] Selberg, E. and Etzioni, O. On the instability of Web search engines. In Proceedings of RIAO ’00, 2000.

[54] Teevan, J., E. Adar, R. Jones, and M. A. Potts. Information reretrieval: repeat queries in Yahoo's logs. SIGIR ‘07, 151-158, 2007.