IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 1, Ver. VI (Feb. 2014), PP 01-05
www.iosrjournals.org

Study of Web Crawler and its Different Types

Trupti V. Udapure[1], Ravindra D. Kale[2], Rajesh C. Dharmik[3]
[1]M.E. (Wireless Communication and Computing) student, CSE Department, G.H. Raisoni Institute of Engineering and Technology for Women, Nagpur, India
[2]Asst Prof., CSE Department, G.H. Raisoni Institute of Engineering and Technology for Women, Nagpur, India,
[3]Asso.Prof. & Head, IT Department, Yeshwantrao Chavan College of Engineering, Nagpur, India,

Abstract : *Due to the current size of the Web and its dynamic nature, building an efficient search mechanism is very important. A vast number of web pages are continually being added every day, and information is constantly changing. Search engines are used to extract valuable Information from the internet. Web crawlers are the principal part of search engine, is a computer program or software that browses the World Wide Web in a methodical, automated manner or in an orderly fashion. It is an essential method for collecting data on, and keeping in touch with the rapidly increasing Internet. This Paper briefly reviews the concepts of web crawler, its architecture and its various types.*

Keyword*: Crawling techniques, Web Crawler, Search engine, WWW*

I. Introduction

In modern life use of internet is growing in rapid way. The World Wide Web provides a vast source of information of almost all type. Now a day's people use search engines every now and then, large volumes of data can be explored easily through search engines, to extract valuable information from web. However, large size of the Web, searching all the Web Servers and the pages, is not realistic. Every day number of web pages is added and nature of information gets changed [1]. Due to the extremely large number of pages present on Web, the search engine depends upon crawlers for the collection of required pages [6].

Web crawling 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 [3]. 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 [1]. 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 [6]. Google crawlers run on a distributed network of thousands of low-cost computers and can therefore carry out fast parallel processing. This is why and how Google returns results within fraction of seconds [4].

Web crawlers-also known as robots, spiders, worms, walkers, and wanderers- are almost as old as the web itself. The first crawler, Matthew Gray's Wandered, was written in the spring of 1993, roughly coinciding with the first release of NCSA mosaic [5].

This paper analyses the concepts of web crawler. This work is organized as follows. Section 1 introduces the web crawler; section 2 is the literature review; section 3 is about the web crawler and its working; section 4 deliberates the different types of crawlers and Section 5 brings out the conclusion.

II. Literature Review

WWW 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 [18] , Matthew Gray wrote the first Crawler, the World Wide Web Wanderer, which was used from 1993 to 1996 [10]. J. Cho. in [18] describes various search techniques and how the search engines works by using crawler and in [26] 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. [19] Proposes how one can maintain local copies of remote data sources "fresh," when the source data is updated autonomously and independently.[21][24] gives an idea about different types crawler. Gautam Pant and Filippo Menczer examined the use of focused crawler in [16][17]. S.S. Dhenakaran1 and K. Thirugnana Sambanthan [3] 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 [4] 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 [5] has developed a Focused crawler using breadth-first search to extract only the relevant web pages of interested topic from the Internet. In [6] author Keerthi S. Shetty, Swaraj Bhat and Sanjay Singh, used symbolic model checking approach to model the basic operation of crawler and verify its properties by using The tool NuSMV. It helps to verify the constraints placed on the system by exploring the entire state space of the system. In [19] author Hiroshi Takeno, Makoto Muto, Noriyuki Fujimoto introduced a new Web crawler that collects Web content suitable for viewing on mobile terminals such as PDA or cell phones. They have described "Mobile Search Service" that provides content suitable for mobile terminals.

III. Web Crawler

A web crawler is a 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 [7].

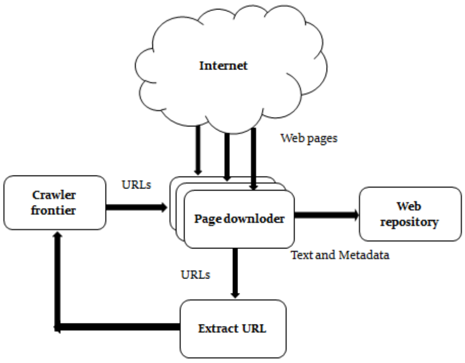

Fig1: architecture of a web crawler

Working of Web Crawler

Figure 1 shows the generalized architecture of web crawler. It has three main components: a ***frontier*** which stores the list of URL's to visit, ***Page Downloader*** which download pages from WWW and ***Web Repository*** receives web pages from a crawler and stores it in the database. Here the basic processes are briefly outline.

Crawler frontier: - It contains the list of unvisited URLs. The list is set with seed URLs which may be delivered by a user or another program [16]. Simply it's just the collection of URLs. The working of the crawler starts with the seed URL. The crawler retrieves a URL from the frontier which contains the list of unvisited URLs. The page corresponding to the URL is fetched from the Web, and the unvisited URLs from the page are added to the frontier [17]. The cycle of fetching and extracting the URL continues until the frontier is empty or some other condition causes it to stop. The extracting of URLs from the frontier based on some prioritization scheme [15].

Page downloader: - The main work of the page downloader is to download the page from the internet corresponding to the URLs which is retrieved from the crawler frontier. For that, the page downloader requires a HTTP client for sending the HTTP request and to read the response. There should be timeout period needs to set by the client in order to ensure that it will not take unnecessary time to read large files or wait for response from slow server. In the actual implementation, the HTTP client is restricted to only download the first 10KB of a page. [8].

Web repository:* -** It use to stores and manages a large pool of data ***"objects," [12] in case of crawler the object is web pages. The repository stores only standard HTML pages. All other media and document types are ignored by the crawler [21]. It is theoretically not that different from other systems that store data objects, such as file systems, database management systems, or information retrieval systems. However, a web repository does not need to provide a lot of the functionality like other systems, such as transactions, or a general directory naming structure [12]. It stores the crawled pages as distinct files. And the storage manager stores the up-to-date version of every page retrieved by the crawler.

The working of a web crawler is as follows:

- Initializing the seed URL or URLs
- Adding it to the frontier
- Selecting the URL from the frontier
- Fetching the web-page corresponding to that URLs
- Parsing the retrieved page to extract the URLs[21]
- Adding all the unvisited links to the list of URL i.e. into the frontier
- Again start with step 2 and repeat till the frontier is empty.

The working of web crawler shows that it is recursively keep on adding newer URLs to the database repository of the search engine. This shows that the major function of a web crawler is to add new links into the frontier and to choose a recent URL from it for further processing after every recursive step [7]. Flow of basic crawler is shown in figure 2.

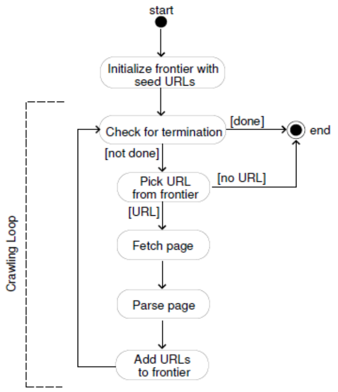

Fig2: flow of a basic crawler

The behaviour of a Web crawler is the outcome of a combination of policies:

- A ***selection policy:*** states which pages to download,
- A ***re-visit policy:*** states when to check for changes to the pages,
- A ***politeness policy:*** states how to avoid overloading Web sites, and
- A ***parallelization policy:*** states how to coordinate distributed Web crawlers [3][13].

Some Examples of web crawlers are ***Yahoo! Slurp*** was the name of the Yahoo! Search crawler, ***Bingbot*** is the name of Microsoft's Bing webcrawler, ***FAST Crawler*** is a distributed crawler, ***PolyBot*** is a distributed crawler, ***RBSE*** was the first published web crawler, ***WebCrawler*** was used to build the first publicly available full-text index of a subset of the Web, ***Googlebot*** is the name of the Google search crawler etc.

IV. Types Of Web Crawler

Different strategies are being employed in web crawling. These are as follows.

4.1 Focused Web Crawler

Focused Crawler is the Web crawler that tries to download pages that are related to each other [4][21]. It collects documents which are specific and relevant to the given topic [7][14]. It is also known as a Topic Crawler because of its way of working [4][17]. 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 [11]. The search exposure of focused web crawler is also huge [2][9].

4.2 Incremental Crawler

A traditional crawler, in order to refresh its collection, periodically replaces the old documents with the newly downloaded documents. On the contrary, an incremental crawler incrementally refreshes the existing collection of pages by visiting them frequently; based upon the estimate as to how often pages change [21]. It also exchanges less important pages by new and more important pages. It resolves the problem of the freshness of the pages. The benefit of incremental crawler is that only the valuable data is provided to the user, thus network bandwidth is saved and data enrichment is achieved [22][27].

4.3 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 [2]. 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 [23].

4.4 Parallel Crawler

Multiple crawlers are often run in parallel, which are referred as Parallel crawlers [24]. A parallel crawler consists of multiple crawling Processes [24] called as C-procs which can run on network of workstations [25]. The Parallel crawlers depend on Page freshness and Page Selection [20]. A Parallel crawler can be on local network or be distributed at geographically distant locations [2].Parallelization of crawling system is very vital from the point of view of downloading documents in a reasonable amount of time [25].

V. 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. The overview of different crawling technologies has been presented in this paper. When only information about a predefined topic set is required, "focused crawling" technology is being used. Compared to other crawling technology the Focused Crawling technology is designed for advanced web users focuses on particular topic and it does not waste resources on irrelevant material.

References

[1] Bharat Bhushan1, Narender Kumar2," *Intelligent Crawling On Open Web for Business Prospects*", IJCSNS International Journal of Computer Science and Network Security, VOL.12 No.6, June 2012

[2] Pavalam S. M., S. V. Kasmir Raja, Jawahar M., and Felix K. Akorli, *"Web Crawler in Mobile Systems",* International Journal of Machine Learning and Computing, Vol. 2, No. 4, August 2012

[3] S.S. Dhenakaran1 and K. Thirugnana Sambanthan2, "*WEB CRAWLER - AN OVERVIEW",* International Journal of Computer Science and Communication Vol. 2, No. 1, January-June 2011, pp. 265-267

[4] Ms. Swati Mali, Dr. B.B. Meshram, "*Implementation of Multiuser Personal Web Crawler",*CSI Sixth International Conference on Software Engineering (CONSEG), IEEE Conference Publications, 2012

[5] Pooja Gupta and Mrs. Kalpana Johari, *"Implementation of Web Crawler",* Second International Conference On Emerging Trends In Engineering and Technology, ICETET-09, IEEE Conference Publications,2009

[6] Keerthi S. Shetty, Swaraj Bhat and Sanjay Singh, *"Symbolic Verification of Web Crawler Functionality and Its Properties",* International Conference on Computer Communication and Informatics (ICCCI -2012), Coimbatore, INDIA, IEEE Conference Publications,2012

[7] 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

[8] Wenxian Wang, Xingshu Chen, Yongbin Zou, Haizhou Wang, Zongkun Dai, *"A Focused Crawler Based on Naive Bayes Classifier",* Third International Symposium on Intelligent Information Technology and Security Informatics, IEEE Conference Publications,2010

[9] Manas Kanti Dey, Debakar Shamanta, Hasan Md Suhag Chowdhury, Khandakar Entenam Unayes Ahmed, *"Focused Web Crawling: A Framework for Crawling of Country Based Financial Data*", Information and Financial Engineering (ICIFE), IEEE Conference Publications, 2010

[10] Dr Rajender Nath, Khyati Chopra, *"Web Crawlers: Taxonomy, Issues & Challenges",* International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 4, April 2013

[11] Debashis Hati, Biswajit Sahoo, Amritesh Kumar, *"Adaptive Focused Crawling Based on Link Analysis",* 2nd International Conference on Education Technology and Computer (ICETC),2010

[12] Jun Hirai Sriram Raghavan Hector Garcia-Molina Andreas Paepcke, *"WebBase : A repository of web pages"* , available: http://ilpubs.stanford.edu:8090/380/1/1999-26.pdf

[13] Frank McCown, Michael L. Nelson, "*Evaluation of Crawling Policies for a Web Repository Crawler"* , Copyright 2006 ACM 1595934170/06/0008, HT'06, August 22–25, 2006

[14] Shashi Shekhar, Rohit Agrawal and Karm Veer Arya, *"An Architectural Framework of a Crawler for Retrieving Highly Relevant Web Documents by Filtering Replicated We Collections",* 2010 International Conference on Advances in Computer Engineering, IEEE Conference Publications 2010.

[15] Ioannis Avraam, Ioannis Anagnostopoulos*, "A Comparison over Focused Web Crawling Strategies"* 2011 Panhellenic Conference on Informatics, IEEE Conference Publications, 2011

[16] Pant Gautam, Srinivasan Padmini, Menczer Filippo, *"Crawling the Web"* In Levene, Mark; Poulovassilis, Alexandra. Web Dynamics: Adapting to Change in Content, Size, Topology and Use. Springer. pp. 153-178. 2004

[17] Gautam Pant, Padmini Srinivasan, *"Learning to Crawl: Comparing Classification Schemes",* ACM Transactions on Information Systems, Vol. 23, No. 4, October 2005, Pages 430–462.

[18] A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke, and S. Raghavan. "*Searching the Web"*. ACM Transactions on Internet Technology, 1(1), 2001.

[19] J. Cho, Hector Garcia-Molina *"Effective Page Refresh Policies for Web Crawlers"*, ACM Transactions on Database Systems, Vol. 28, No. 4, December 2003,

[20] AH Chung Tsol, Daniele Forsali, Marco Gori, Markus Hagenbuchner, Franco Scarselli, *"A Simple Focused Crawler"* Proceeding 12th International WWW Conference 2003(poster), pp. 1.

[21] Junghoo Cho and Hector Garcia-Molina. 2000a. *"The evolution of the web and implications for an incremental crawler"*, In Proceedings of the 26th International Conference on Very Large Databases.

[22] A. K. Sharma and Ashutosh Dixit*, "Self Adjusting Refresh Time Based Architecture for Incremental Web Crawler"* International Journal of Computer Science and Network Security, vol.8 no.12, 2008, pp. 349-354

[23] Vladislav Shkapenyuk Torsten Suel *"Design and Implementation of a High-Performance Distributed Web Crawler",* CIS Department Polytechnic University Brooklyn, NY 11201

[24] Junghoo Cho, Hector Garcia-Molina, *"Parallel Crawlers", WWW2002, May 7–11, 2002, Honolulu, Hawaii, USA, ACM 1-58113-449-5/02/0005.*

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

[26] 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.

[27] Niraj Singhal, Ashutosh Dixit, and Dr. A. K. Sharma, *"Design of a Priority Based Frequency Regulated Incremental Crawler",* International Journal of Computer Applications (0975 – 8887) vol.1, no. 1, 2010, pp. 42-47.