SURVEY OF WEB CRAWLING ALGORITHMS

Rahul kumar[1], Anurag Jain[2] and Chetan Agrawal[3]

[1]Department of CSE Radharaman Institute of Technology and Science, Bhopal, M.P, India
[2]Department of CSE Radharaman Institute of Technology and Science, Bhopal, M.P, India
[3]Assistant Prof. Department of CSE Radharaman Institute of Technology and Science,India

Abstract

The World Wide Web is the largest collection of data today and it continues increasing day by day. A web crawler is a program from the huge downloading of web pages from World Wide Web and this process is called Web crawling. To collect the web pages from www a search engine uses web crawler and the web crawler collects this by web crawling. Due to limitations of network bandwidth, time-consuming and hardware's a Web crawler cannot download all the pages, it is important to select the most important ones as early as possible during the crawling process and avoid downloading and visiting many irrelevant pages. This paper reviews help the researches on web crawling methods used for searching.

Keywords:

Web crawler, Web Crawling Algorithms, Search Engine

1. INTRODUCTION

A web crawler or spider is a computer program that browses the WWW in sequencing and automated manner. A crawler which is sometimes referred to spider, bot or agent is software whose purpose it is performed web crawling. The basic architecture of web crawler is given below (Figure1). More than 13% of the traffic to a web site is generated by web search [1]. Today the size of the web is thousands of millions of web pages that is too high and the growth rate of web pages are also too high i.e. increasing exponentially due to this the main problem for search engine is deal this amount of the size of the web. Due to this large size of web induces low coverage and search engine indexing not cover one third of the publicly available web [12]. By analyzing various log files of different web site they found that maximum web request is generated by web crawler and it is on an average 50% [15]. Crawling the web is not a programming task, but an algorithm design and system design challenge because of the web content is very large. At present, only Google claims to have indexed over 3 billion web pages. The web has doubled every 9-12 months and the changing rate is very high [1, 2, 3].About 40% web pages change weekly [5] when we consider lightly change, but when we consider changing by one third or more than the changing rate is about 7% weekly [7].

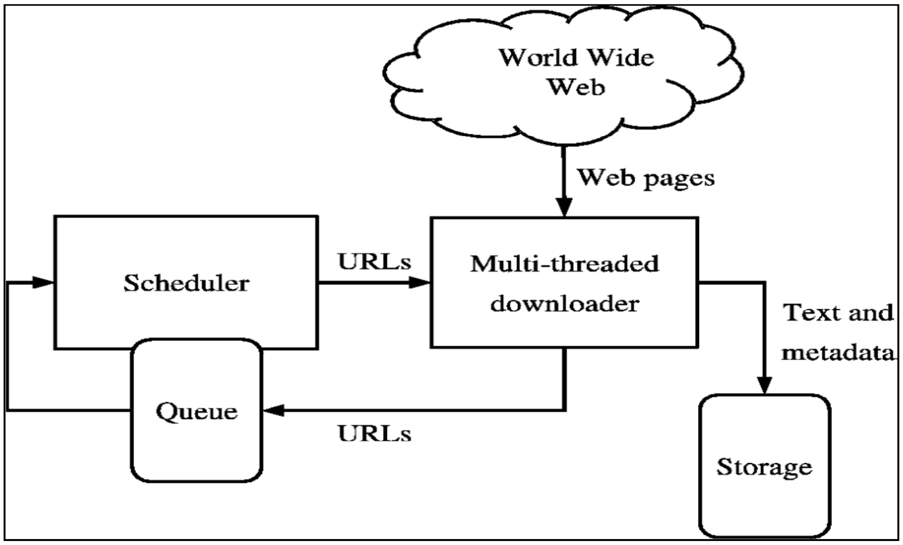

Figure1: Architecture of Crawler

Researchers are developing new scheduling policy for downloading pages for the world wide web which guarantees that, even if we want do not download all the web pages we still download the most important (by the user point of view) ones. As the size of Internet data grows, it will be very vital to download the important ones first, as it will be impossible to download all of them.

The rest of the paper is organized as follows. Section2 is explained about fundamentals of web crawling. Section3 gives detail about web crawler strategies with diagram. Section4 has critical analysis with tables. Research scope is in section 5. Conclusion and references are at last.

2. FUNDAMENTALS OF WEB CRAWLING

A crawler which is sometimes referred to spider, bot or agent is software whose purpose, it has performed web crawling [4]. This can be used for accessing the Web pages from the web server as per user pass queries commonly for search engine (see figure2). A web crawler also used sitemap protocol for crawling web pages. In the crawling process, generally starts with a set of Uniform Resource Locator (URLs) called the Seed URL.

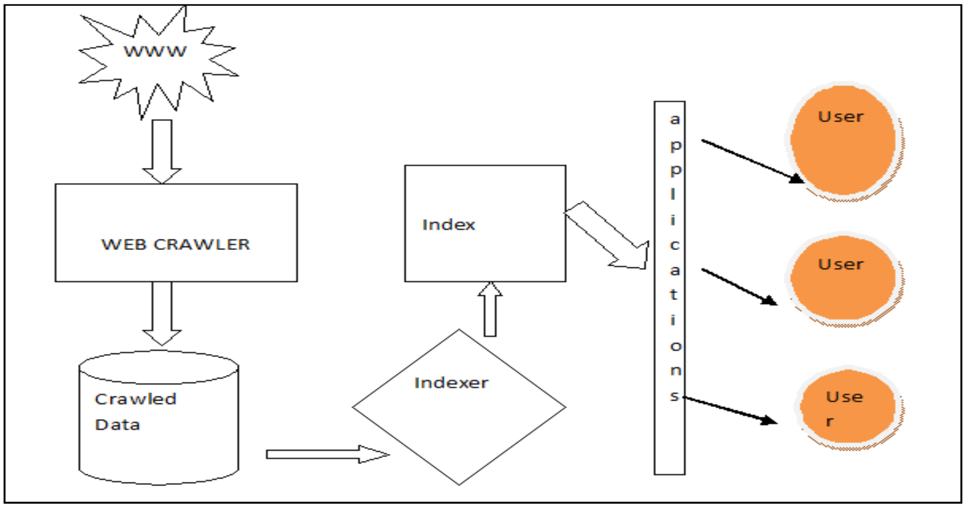

Figure2: Accessing the web pages through WWW by web crawlers

In web crawling processes start from a URL set (Seed URL), but we should keep in mind that the starting URL will not reach all the web pages. The basic web crawling algorithm fetches (i) a web page (II) Parse it to extract all linked URLs (III) For all the web URLs not seen before, repeat (I) -(III). The size of the web is large, so this web search engine can’t cover all the websites in www. There should be high chances of the relevant pages to be in the first few download, as the web crawler always downloads web pages infractions. By the analyzed various log files of different web site. Web crawler request for a web is equivalent to 50%. So we can find most valuable web pages so crawler can download these pages for search engine [16].

3. WEB CRAWLER STRATEGIES

3.1 Breadth First Search Algorithm

Breadth first algorithm work on a level by level, i.e. algorithm starts at the root URL and searches the all the neighbors URL at the same level. If the desired URL is found, then the search terminates. If it is not, then search proceeds down to the next level and repeat the processes until the goal is reached. When all the URLs are scanned, but the objective is not found, then the failure reported is generated. Breadth first Search algorithm is generally used where the objective lies in the depthless parts in a deeper tree. [6][13].

3.2 Depth First Crawling Algorithm

Depth first search algorithm is a more useful search which starts at the root URL and traverse depth through the child URL. First, we move to the left most child if one or more than one child exist and traverse deep until no more is available (Figure3). Here backtracking is used to the next unvisited node and processes is repaid in similar manner [9]. By the use of this algorithms authors makes sure that all the edges, i.e. all URL is visited once breath [10]. It is very efficient for search problems, but when the child is large then this algorithm goes into an infinite loop [8].

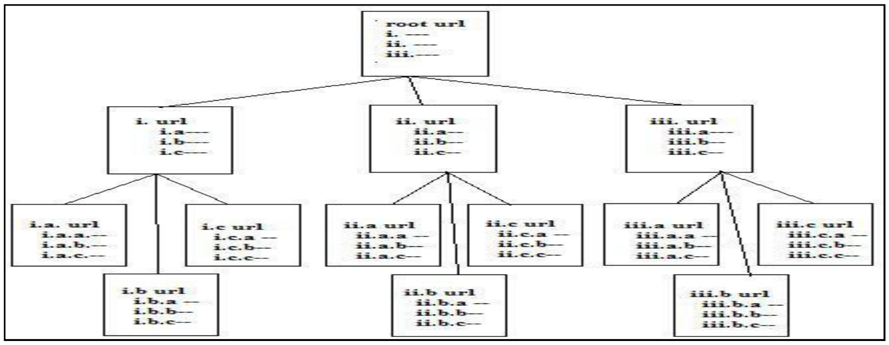

Figure3: Depth First Search

3.3 Page Rank Algorithm

By Page rank algorithm web crawler determines the importance of the web pages in any web site by the total number of back links or citations in providing page [10]. The page rank of a provided web page is calculated as Relatedness between the web pages are taken into account by the Page Rank algorithm. The web page whose number of input link is high is considered of more importance relative to other web page, i.e. interest degree of the page to another. When the number of input link is increased, then interest degree of a page obviously also increases. Therefore, the total weighted sum of input links defines the page rank of a web page [11]

3.4 Online Page Importance Calculation Algorithm

On-line Page Importance Computation (OPIC) in this method, to find that importance of any page in web site, i.e. each page has a unique cash value that is equally distributed to all output links, initially all pages in any website have the same cash and it is equal to 1/n. The crawler will start downloading web pages with higher cashes in each and every stage and cash will be distributed among all the pages it points when a web page is downloaded. Unfortunately, by the use of in this method, each web page will be downloaded many times so that the web crawling time also increase [14]

3.5 Crawling the large sites first

In 2005 Ricardo BaezaYates et al “Crawling a Country: Better Strategies than Breadth First for Web Page Ordering” perform experiments in approx 100 million web pages and find that crawling the large site first scheme has practically most useful then on-line page importance computation. The web crawler fined first of all un –crawled web pages to find high priority web page for picking a web site, and starts with the sites with the large number of pending pages [3].

3.6 Crawling through URL Ordering

Junghoo Cho et al “Efficient Crawling Through URL Ordering” find that a crawler is to select URLs & to scan from the queue of known URLs so as to find more important pages first when it visits earlier URLs that have anchor text which is similar to the driving query or link distance is also short to a page and that type of web pages to be known important [20].

3.7 Batch-page rank

Batch-page rank In This strategy first calculation of an estimation of Page rank by the help of pages seen so far, every N page downloaded. After those again next N pages are selected to download are those web pages with the highest estimated Page rank [20].

3.8 Partial-page rank

Partial-page rank This is also similar to batch-page rank, but indifference between Page rank re-calculations of N pages in batch page rank and partial page rank; a temporary page rank is assigned to all new pages by using the total sum of the Page rank of the web pages pointing to it and it is divided by the total number of out-links of those pages [21].

3.9 By HTTP Get Request and Dynamic Web Page

It is a Query based Approach to minimize the Web Crawler or spider Traffic by using HTTP Get Request and also Dynamic Web Page. According to the author it is a query based approach to inform all updates on the web site by web crawler using by Dynamic web page and also HTTP GET Request [17]. And crawler download only updated web pages after the last visit.

3.10 Using Customized Sitemap

Increasing the Efficiency of Crawler by the Using Customized Sitemap.When a crawler revisiting the websites and find that which web pages have been updated or newly added since last visit, then there is no need to download the complete website every time. With this scheme, it will be less time consuming for web crawlers to maintain the freshness of downloaded websites used by search engines. [18].

3.11 By the use of filter

“A Novel Web Crawler Algorithm on Query based Approach with Increases Efficiency” The authors proposed a modify approach for crawling by the use of a filter and this is a query based approach. Filter always redirects the updated web pages and crawler downloads all updated web pages after LAST_VISIT [19].

4. CRITICAL ANALYSIS

Ref. No.	Method	Concept	Advantage	Limitations
3	Crawling the large sites first	Crawling starts with the sites with the large number of pending pages, i.e. web pages for crawling.	Large web site crawled first	When important pages exist in short web site, then this is crawled latter.
6	Breadth First Search Algorithm	Starts at the root URL and searches the all the neighbors URL at the same level	Well suited for situations where the objective is found on the shallower parts in a deeper tree	It will not perform so well when the branches are so many in a game tree
9	Depth First Search Algorithm	Starts at the root URL and traverse depth through the child URL.	Well suited for such problems	When the branches are large then this algorithm takes might end up in an infinite loop
10	Page Rank Algorithm	Download the web pages on the basis of page rank.	In the very limited time important pages are downloaded.	In high Page Rank pages Are always good in quality and we just download it
14	Online Page Importance Calculation Algorithm	The crawler will download web pages with higher cashes in each stage and cash will be distributed between the pages it points when a page is downloaded	The cash value is calculated in one step and very short duration of time.	Each page will be downloaded many times that will increase crawling time
17	By HTTP Get Request and Dynamic Web Page	It is a query based approach and crawler just download updated web pages after the last visit.	Web crawler download only downloads latest updated web pages.	We do not see before last visit updated web pages.
19	By the use of filter	Crawling is done by the use of filter and this is also a query based approach	Reduces web crawling or network traffic.	An extra filter is used so crawler done extra work.
20	Crawling through URL Ordering	It visits earlier URLs that have anchor text which is similar to the driving query or link distance is also short to a page	Extremely useful when we are trying to crawl a fraction of the Web, and we need to revisit pages often to detect changes	When many clusters have existed on the web site then performance is decreased.

5. RESEARCH SCOPE

As, the defined concepts for web crawling and improving its performance by the various crawling algorithms have been explained here. It has not end of the work for improving performance of crawling. There are many more techniques and algorithms may be considered for crawler to improve its performance. We can also improve its performance to modify the sitemap of any web site, i.e. in sitemap protocol all URL has a static priority and we can change it by dynamic priority and this priority is calculated through user interest i.e. number of hits has high priority.

6. CONCLUSION

The paper surveys several crawling methods or algorithms that are used for downloading the web pages from the World Wide Web. We believe that all of the algorithms discuss in this paper are well effective and high performance for web search, reduce the network traffic and crawling costs, but overall advantages and disadvantage favor more for By using HTTP Get Request and also Dynamic Web Page and download updated web pages By the using of filter is produce relevant results.

REFERENCES

[1] K. Bharat and A. Z. Broder. A technique for measuring the relative size and overlap of public web search engines. In Proceedings of the 7th World Wide Web Conference, pages 379-388, 1998.

[2] S. Lawrence and C. L. Giles. Searching the World Wide Web. Science, 280(5360):98-100, 1998

[3] Carlos Castillo, Mauricio Marin, Andrea Rodriguez, and Ricardo Baeza-Yates. Scheduling algorithms for Web crawling. In Latin American Web Conference (WebMedia/LA-WEB), Riberao Preto, Brazil, 2004. IEEE Cs. Press.

[4] S. Lawrence and C. L. Giles. Accessibility of information on the web. Nature, 400:107-109, 1999

[5] J. Cho and H. Garcia-Molina. The evolution of the web and implications for an incremental crawler. In Proceedings of the 26th International Conference on Very Large Databases, 2000.

[6] Junghoo Cho and Hector Garcia-Molina ―Effective Page Refresh Policies for Web Crawlers‖ ACM Transactions on Database Systems, 2003.

[7] D. Fetterly, M. Manasse, M. Najork, and J. L. Wiener. A large-scale study of the evolution of web pages. In Proceedings of the 12th International World Wide Web Conference, 2003.

[8] Carlos Castillo , Mauricio Marin , Andrea Rodriguez, ―Scheduling Algorithms for Web Crawling‖in the proceedings of WebMedia and LA-Web, 2004.

[9] Ben Coppin ―Artificial Intelligence illuminated‖ Jones and Barlett Publishers, 2004, Pg 77.

[10] Narasingh Deo ―Graph theory with applications to engineering and computer science‖ PHI, 2004 Pg 301

[11] Sergey Brin and Lawrence Page "Anatomy of a Large scale Hypertextual Web Search Engine" Proc. WWW conference 2004

[12] Ricardo BaezaYates Carlos Castillo Mauricio Marin Andrea Rodriguez," Crawling a Country: Better Strategies than BreadthFirst for Web Page Ordering" International World Wide Web Conference Committee (IW3C2). WWW, Chiba, Japan 2005

[13] Steven S. Skiena ―The Algorithm design Manual‖ Second Edition, Springer Verlag London Limited, 2008, Pg 162

[14] Mehdi Ravakhah, M. K. "Semantic Similarity BasedFocused Crawling" 'First International Conference on Computational Intelligence, Communication Systems and Networks', 2009.

[15] Yang Sun, Isaac G. Councill, C. Lee Giles," The Ethicality of Web Crawlers" IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology2010.

[16] Yang Sun, Isaac G. Councill, C. Lee Giles," The Ethicality of Web Crawlers" 2010

[17] Shekhar Mishra, Anurag Jain, Dr. A.K. Sachan," A Query based Approach to Reduce the Web Crawler Traffic using HTTP Get Request and Dynamic Web Page" International Journal of Computer Applications (0975 – 8887) Volume 14– No.3, January 2011

[18] Dr. Bharat Bhushan, Meenakshi Gupta, Garima Gupta" Increasing The Efficiency Of Crawler Using Customized Sitemap" International Journal of Computing and Business Research (IJCBR) Volume 3 Issue 2 May 2012

[19] S S Vishwakarma, A Jain ,A K Sachan," A Novel Web Crawler Algorithm on Query based Approach with Increases Efficiency" International Journal of Computer Applications (0975 – 8887) Volume 46– No.1, May 2012.

[20] Junghoo Cho, Hector Garc´ıa-Molina, and Lawrence Page. Efficient crawling through URL ordering. In Proceedings of the seventh conference on World Wide Web, Brisbane, Australia, April 1998.

[21] Paolo Boldi, Massimo Santini, and Sebastiano Vigna. Do your worst to make the best: Paradoxical effects in pagerank incremental computations. In Proceedings of the third Workshop on Web Graphs (WAW), volume 3243 of Lecture Notes in Computer Science, pages 168–180, Rome, Italy, October 2004. Springer.

Authors

Rahul Kumar has received his Bachelor's Degree in Computer science Engineering from RGPV UniversityBhopal India.