Volume 8, No. 5, May-June 2017
ISSN No. 0976-5697
International Journal of Advanced Research in Computer Science
REVIEW ARTICLE
Available Online at www.ijarcs.info

Survey of Web Crawler Algorithms

Abhinav Garg, Kratika Gupta* and Abhijeet Singh
Department of Computer Science,
Galgotias College of Engineering and Technology,
Greater Noida, India

Abstract: In today's scenario, World Wide Web (WWW) is flooded with huge amount of information. Due to growing popularity of the internet, finding the meaningful information among billions of information resources on the WWW is a challenging task. The information retrieval (IR) provides documents to the end users which satisfy their need of information. Search engine is used to extract valuable information from the internet. Web crawler is the principal part of search engine; it is an automatic script or program which can browse the WWW in automatic manner. This process is known as web crawling. Crawling algorithms are crucial in selecting the pages that satisfies the users' needs. This paper reviews the researches on web crawling algorithms used on searching.

Keywords: WWW, Search Engine,WebCrawler,Web Crawling, Web Crawling Algorithms.

1. INTRODUCTION

With the amount of data increasing on the World Wide Web, it becomes extremely important to extract the most relevant information in the shortest span of time. A lot of research is being done to improve the efficiency of search engines by providing crawling algorithms which could traverse through large chunks of data in a short span of time and return the results sorted based on their relevance.These are days of competitive world, where each and every second is considered valuable backed up by information. Timely Information retrieval is a solution for survival. Due to the abundance of data on the web and different user perspective, information retrieval becomes a challenge .

When a data is searched, hundreds and thousands of results appear. The user's don't have persistence and stretch to go through each and every page listed. So the search engines have a bigger job of sorting out the results, in the order of interestingness of the user within the first page of appearance and a quick summary of the information provided on a page.
Web crawlers are programs which traverse through the web searching for the relevant information[1]using algorithms that narrow down the search by finding out the most closer and relevant information.Web pages needs not only relevance but also authoritativeness – from a trusted source of strong, precise information[3].Search engines uses algorithms which sorts, ranks the result in the order of authority, that is closer to the user's query. Many algorithms are is in use - Breadth first search, Best first search, Page Rank algorithm, Genetic algorithm, Naïve Bayes clssification algorithm to mention a few.
Not all information represented are useful. The search engine techniques may become useless or junky if the information it draws are not attracting users, especially if the malicious user who are trying to attract more traffic in to their site by embedding the most used keywords invisibly in to their site. The challenges are relevancy, robustness and the ability to download large number of pages.

2. LITERATURE SURVEY

When a data is searched, hundreds of thousands of results appear. Users do not have the persistence and stretch to go through each and every page listed. So search engines have a big job of sorting out the results, in the order of interest to the user within the first page of appearance and a quick summary of the information provided on a page [3]
Retrieving effective content from the Web is a crucial task because it heavily influences the perceived effectiveness of a search engine. Users often look at only a few top hits, making the precision achieved by the ranking algorithm of paramount importance. Early search engines ranked pages principally based on their lexical similarity to the query. The key strategy was to devise the best weighting algorithm to represent Web pages and query in a vector space, so that closeness in such a space would be correlated with semantic relevance. Web Crawler is a program/software or automated script which browses the World Wide Web in a methodical, automated manner. Crawlers have bots that fetch new and recently changed websites, and then indexes them. By this process billions of websites are crawled and indexed using algorithms (which are usually well-guarded secrets) depending on a number of factors. Several commercial search engines change the factors often to improve the search engines process[4].
The basic procedure executed by any web crawling algorithm takes a list of seed URLs as its input and repeatedly executes the following steps[6]:

- Remove a URL from the URL lit.
- Download the corresponding page.
- Check the Relevancy of the page.
- Extract any links contained in it.
- Add these links back to the URL list.
- After all URLs are processed, return the most relevant page.

3. WEB CRAWLING 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[5].

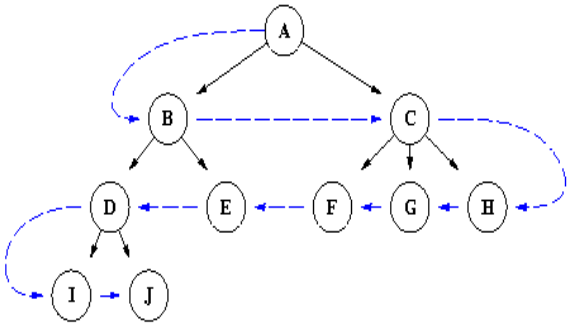

Fig 1.breadth first search

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. Here backtracking is used to the next unvisited node and processes is repaid in similar manner[8]. By the use of this algorithms authors makes sure that all the edges, i.e. all URL is visited once breath. It is very efficient for search problems, but when the child is large then this algorithm goes into an infinite loop.

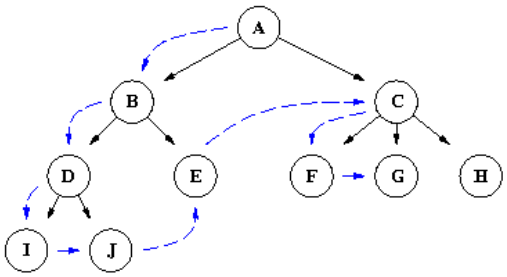

Fig 3.2. Depth first search

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. 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[7].

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[9].

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[10].

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.[11]

3.7 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[2]. And crawler download only updated web pages after the last visit.

METHOD	CONCEPT	ADVANTAGE	LIMITATION
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.
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
Depth First Search Algorithm	Starts at the root URL and traverse depth through the	Well suited for such problems	When the branches are large then this algorithm takes might end up in

	child URL.		an infinite loop
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
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
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
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.

4. 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.

5. CONCLUSION

The main objective of the review paper was to throw some light on the web crawling algorithms. We also discussed the various search algorithms and the researches related to respective algorithms and their strengths and weaknesses associated. 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] Pavalam S M, Jawahar M, Felix K Akorli, S V Kashmir Raja " Web Crawler in Mobile Systems" International Conference on Machine Learning (ICMLC 2011), Vol. , pp

[2] 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

[3] AlessioSignorini, "A Survey of Ranking Algorithms"retrievedfrom http://www.divms.uiowa.edu/~asignori/phd/report/asurvey-of-ranking-algorithms.pdf 29/9/2011

[4] Pavalam, S. M., SV Kashmir Raja, Felix K. Akorli, and M. Jawahar, "A Survey of Web Crawler Algorithms," International Journal of Computer Science, vol. 8, iss. 6, no 1, Nov. 2011.

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

[6] Menczer, Filippo, Gautam Pant, and Padmini Srinivasan, "Topical webcrawlers: Evaluating adaptive algorithms," ACM Transactions on Internet Technology (TOIT), vol. 4, no. 4, pp. 378-419, 2004

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

[8] Ben Coppin "Artificial Intelligence illuminated"Jones and Barlett Publishers, 2004, Pg 77.

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

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

[11] unghoo 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.

