International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 2 Issue: 3 488 – 492

Analysis of Web Crawling Algorithms

Rashmi Janbandhu[#1], Prashant Dahiwale[*2], M.M.Raghuwanshi[#3]

[#]*Comp.Sc. & Engg. Department -Comp.Sc. & Engg Department,*
Rashtrasant Tukadoji Maharaj Nagpur University - Rashtrasant Tukadoji Maharaj Nagpur University
Rajiv Gandhi College of Engineering and Research, Nagpur,India

[1]rashmi.janbandhu@gmail.com
[3]m_raghuwanshi@rediffmail.com

[*]*Comp.Sc. & Engg Department*
Rajiv Gandhi College of Engineering and Research, Nagpur,India
[2]prashant.dahiwale@gmail.com

Abstract— The web today is huge and enormous collection of data today and it goes on increasing day by day. Thus, searching for some particular data in this collection has a significant impact. Researches taking place give prominence to the relevancy and relatedness of the data that is found. Inspite of their relevance pages for any search topic, the results are still huge to be explored. Another important issue to be kept in mind is the users' standpoint differs from time to time from topic to topic. Effective relevance prediction can help avoid downloading and visiting many irrelevant pages. The performance of a crawler depends mostly on the opulence of links in the specific topic being searched. This paper reviews the researches on web crawling algorithms used for searching.

Keywords— *Web Crawling Algorithms, Crawling Algorithm Survey, Search Algorithms, Lexical Database, Metadata, Semantic.*

*****

I. INTRODUCTION

Web search is currently generating more than 13% of the traffic to Web sites [1].The main problem search engines have to deal with is the size of the Web, which currently is in the order of thousands of millions of pages that is too enormous and is increasing exponentially. This large size induces a low coverage, with no search engine indexing more than one third of the publically available Web[3].

Researchers are developing scheduling policy for downloading pages from the Web which guarantees that ,even if we do not download all the pages, we still download the most important ones. As the size of Internet data grows, it will be very vital to download the superior ones first, as it will be impossible to download all of them.

Typing "Java" as keywords into Google search engine would lead to around 25 million results with quotation marks and 237 million results without quotation marks. With the same keywords, Yahoo search engine leads to around 8 million results with quotation marks and 139 million results without quotation marks, while MSN search engine leads to around 8 million results with quotation marks and around 137 million results without quotation marks[2]. These gigantic numbers of results are brought to the user, of which only few are relevant and rest are uninteresting to the users.

This complete set of circumstances fetches attention to a prime issue which is the relevance of a webpage to a specific topic. The process used by search engines to index their databases is very clandestine, they use varied number of web crawlers for collecting and arranging information.

The section II discusses the fundamentals of web crawling process. Section III gives a detailed description of the various crawling policies and the working of a web crawler and various prioritizing algorithms. Section IV deals with our conclusion.

II. FUNDAMENTALS OF WEB CRAWLING

Web crawlers are programs which traverse through the web searching for the relevant information [4] using algorithms that narrow down the search by finding out the most closer and relevant information. This process of crawling is iterative, as long the results are in close proximity of user's interest. The algorithm determines the relevancy based on the factors such as frequency and location of keywords in the web pages.

Crawlers have web robots that fetches new and recently changed websites, and indexes them. A extremely large number of websites i.e. billions of websites are crawled and indexed using algorithms (When algorithms are published, there is often an important lack of details that prevents other from reproduce the work. There are also emerging concerns about "search engine spamming" which prevent major search engines from publishing their

ranking algorithms [5].) depending on a number of factors. The search engines innovates the use of the considerations and features very often to improve the search engines process.

The process generally starts with a set of Uniform Resource Locator (URLs) called the Seed URLs. When an user initiates a search, the key words are extracted and the index for the websites is searched to categorise which websites are most relevant. Relevancy is determined by a number of factors and also it differs for the different search engines in accordance to the methodology and strategies it follows.

We have to start from any URL (Seed), but we should keep in mind that the starting URL will not reach all the web pages and will not refer to any page which in turn would refer back to the seed URL again, because if this happens then eventually it makes us to restart the crawl. It is always better to take good
seed URL. For example Google or Yahoo can be used to get seed URL by simply entering the keywords into them and considering their resulting links as our seed URLs. This is because these are amongst the popular search engines whose results are prominent and accepted by majority of users around the world.

The size of the web is huge, search engines practically can't be able to cover all the websites. There should be high chances of the relevant pages to be in the first few downloads, as the web crawler always download web pages in fractions. This scenario calls for prioritizing Web pages. The relevancy or cost of any web page is function of its eminent quality, its standing in terms of out links it contain or the number of visits to it by user. There are many strategies for selecting the websites to be downloaded, some of them are breadth first, depth first, page rank, etc.

Researchers of Web crawling have focused on parallelism, discovery and control of crawlers for Web site administrators, accessing content behind forms which is often known as the hidden web, detecting mirrors, keeping the freshness of the search engine copy high, long-term scheduling and focused crawling. There have been also studies on characteristics of the Web, which are relevant to the crawler performance, such as detecting communities, characterizing server response time, studying the distribution of web page changes, studying the macroscopic web structure, and proposing protocols for web servers to cooperate with crawlers [5].

The freshness, newness and revisiting of a page also has a significant importance while crawling the web so that user is benefited by updated and latest information. Two types of visiting policy has been proposed –Uniform change frequency - the revisiting is done at the uniform regardless of its change and non uniform change frequency – the revisiting is not uniform and the revisiting is done more frequently and the visiting frequency is directly proportional to the change frequency[6].

III. WEB CRAWLER STRATEGIES

A. Breadth First Search Algorithm

This algorithm starts at the root URL and searches the all the neighbour URL at the same level. If the goal is reached, then it is reports success and the search terminates. If it is not, search proceeds down to the next level sweeping the search across the neighbour URL at that level and so on until the goal is reached. When all the URLs are searched, but the objective is not met then it is reported as failure.

Breadth first is 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, especially like chess game and also when all the path leads to the same objective with the same length of the path [7][8].

In our example fig.1 search stars from the root URL and will collect URL i, ii and iii. The search then proceeds by downloading these three URL, next step will be to download i.a, i.b, i.c also ii.a, ii.b, ii.c and iii.a, iii.b, iii.c which are at the same level. Further URL i.a.a, i.a.b, i.a.c along with ii.a.a, ii.a.b, ii.a.c and iii.a.a, iii.a.b, iii.a.c are downloaded.

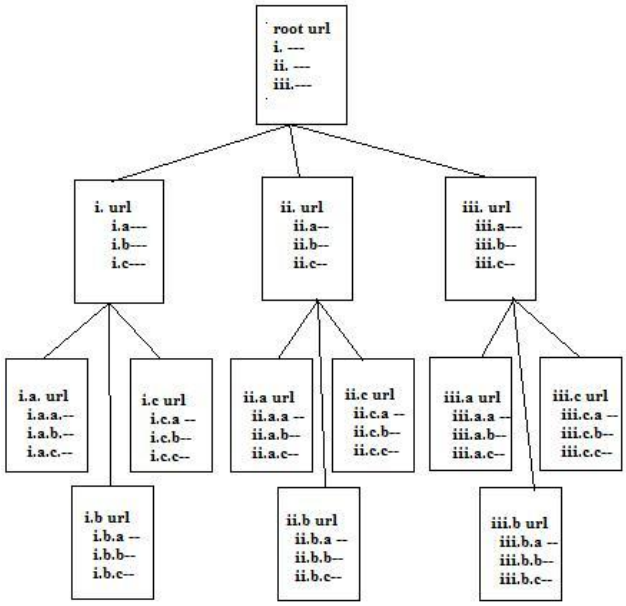

Fig. 1. Best First Search and Depth First Search

B. Depth First Search Algorithm

This powerful search which starts at the root URL and traverse deeper through the child URL. If there are more than one child, then priority is given to the left most child and traverse deep until no more child is available. It is backtracked to the next unvisited node and then continues in a similar manner [9].

This algorithm makes sure that all the edges are visited once breadth [10]. It is well suited for search problems, but when the branches are large then this algorithm takes might end up in an infinite loop [8].

In our example fig.1 search starts with root URL and download the first i URL. The search proceeds by downloading i.a then i.a.a , i.a.b , i.a.c then i.b and its child and then i.c and its child. After this step ii url is downloaded and the process continues.

C. Page Rank Algorithm

Relatedness between the web pages are taken into account by the Page Rank algorithm. For example, if page P1 has a link to page P2, then, P2's content is probably appealing for P1's creator. Therefore, the number of input links to a web page shows the interest degree of the page to others. Obviously, the interest degree of a page increases with the growing number of input links.

PR(P1) = PR(A1)/L(A1) + ... + PR(An)/L(An)

In order to find the Page Rank for a page, called PR(P1), we need to find all the pages that linked to page P1 and Out Link from P1. We found a page A1, which has link from P1 then page L(A1) will give no. of Outbound links to page P1. We do the same for A2, A3 and all other pages linking to Main page P – and Sum of the values will provide Rank of the web page. Moreover, when a web page receives links from an important page then certainly it should have a high rank. Therefore, Page Rank of a web page corresponds to the weighted sum of input links[11].

D. Path-Ascending Crawling Algorithm

The path ascending crawling algorithm would crawl each path from the home to the last file of that URL.This nature of the crawler helps get more information from that site. In the above way a crawler would ascend to every path in each URL (Uniform Resource Locator) that it intends to crawl. For example, when given a seed URL of http://rashmi.org/rash/profit.html, it will attempt to crawl /rashmi.org/, /rash/ and /profit.html.

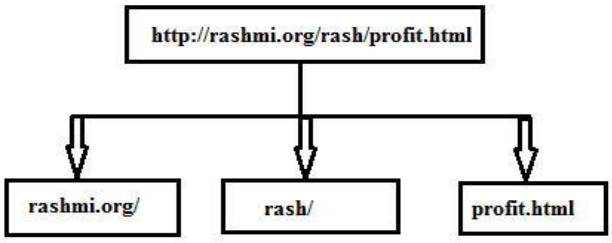

Fig. 1. Path - Ascending Crawler

The advantage with Path-ascending crawler is that they are very effective in finding isolated resources, or resources for which no inbound link which would have been found in regular crawling[5].

E. Focused Crawling Algorithm

The significance of a page for a crawler can also be expressed as a function of the similarity of a page to a given query. In this approach we can intend web crawler to download pages that are similar to each other, thus it would be called focused crawler or topical crawler[12].

The main problem in focused web crawling is that, we would like to envisage the similarity of the text of a given page to the query before actually downloading the page. This crawler would use that content of the pages which is already visited and infer the similarity between the driving query and the pages that have not been visited yet. The features such as URL, anchor text which are available without downloading that particular page are used to predict the similarity of unvisited page. Focused crawling usually relies on a general Web search engine for providing starting points i.e. its seed URLs. This type of crawler can be used to have specific type of search engines based on their file types[13].

F. Online Page Importance Calculation Algorithm

On-line Page Importance Computation (OPIC) in this method, each page has a cash value that is distributed equally to all output links, initially all pages have the same cash equal to 1/n. This is similar to Page Rank while it is done in one step.

If http://rashmi.org has “m” no. of pages in it,

Then each page obtains 1/m cash.

The crawler will download web pages with higher cashes in each stage and cash will be distributed among the pages it points when a page is downloaded. There is no comparison between OPIC and other crawling strategies. Unfortunately, in this method, each page will be downloaded many times that will increase crawling time[14].

G. Naïve Bayes Classification Algorithm

Naïve Bayes algorithm is based on Probabilistic learning and classification. It assumes that one feature is independent of another. This algorithm proved to be efficient over many other approaches although its simple assumption is not much applicable in realistic world cases.

Mejdl S. Safran, Abdullah Althagafi and Dunren Che proposed an efficient crawler based on Naïve Bayes to gather many relevant pages[2]. Four relevance attributes have been taken into consideration, i.e., the URL words, the anchor text, the parent pages text and the surrounding text of an URL. WordNet (a free online lexical database) is used to find and add new related keywords to further improve prediction accuracy.

H. Sematic Web Crawler Algorithm

Web sites are indexed through special process known as crawling and it helps the search engines to provide the results based on user query request very quickly. Query may be refined through a special query processor which removes stop word, does stemming of the query words to refine it. It is very likely that the search results would be more precise and just-the-thing if we add semantic techniques to search engine.

The crawler algorithm we analyzed uses the metadata of the web page to determine the sense of that web page. The special query processor introduced in the previous paragraph has some significant components which are describes as follows:

1) *Stop Word Identification Module*: Stop Words are the words such as the prepositions which should be filtered out because they make poor index terms for a search result. Basically stop words are removed from the search query if it happens to appear. There are over 421 Stop Words [15] it should have maximum efficient and effective in filtering the most frequently occurring and semantically neutral words in general literature in English language. Removing stop word is the initial process of our proposed query processing technique [16].

Sample S = {Photograph is very beautiful}

On applying Stop Words Removal process to srting S, we get,
S’ = {Photograph beautiful}

2) *Stemming Module:* Stemming is the process of refining the words to make them more effective, expressive, precise and perfect. In our proposed methodology[17].

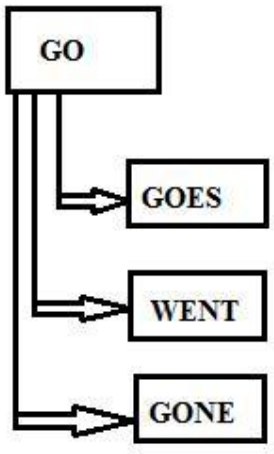

Fig.2. Stemming words for ―GO

Word ―GO can be expressed as going, gone, goes, went, gone, etc can be reduced to GO using Stemming Algorithm[18].
Both the above module, i.e., stop word identification module and stemming module helps to determine the accurate keywords and hence the accurate sense could be determined using lexical database described below. Thus, stop word identification and stemming has significant impact in semantic web crawler algorithm.

3) *Lexical Database:* Lexical database i.e. related terms database is a large collection of Synonyms, Holonyms, Meronymy, Antonyms of English words and are essential component of semantic crawler. Lexical database also provides the synonym or sense of a given word which is called as Synset[19]. Holonyms is the relationship between a term denoting a part or a member of, the

synonym, Meronymy is just opposite for Holonyms, Antonyms will provide alternate opposite meaning for synonym of a word. WordNet[19], ConceptNet[20] and YAGO[21] are best and widely used Lexical Database.

4)Crawler: Crawler fetches the metadata and determines the sense of related page with the help of any lexical database. This determined sense is stored for future reference. The query is then refined and its sense is determined using the same lexical database. The senses of the query and the web page are compared. This process is called sense binding. After this process the result clarifies if they both match or not.

Meta data is used in the abobe process because it contains important features of the web page such as Author name, Description, Keywords and Geographical position of the web page most probably keywords and description.

IV. CONCLUSIONS

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 surveyed in this paper are effective for web search, but the advantages favors more for Focused Crawling Algorithm due to its smallest response time.

REFERENCES

[1] StatMarket. Search engine referrals nearly double worldwide.

[2] http://websidestory.com/pressroom/-pressreleases.html?id=181, 2003.

[3] Mejdl S. Safran, Abdullah Althagafi and Dunren Che “Improving Relevance Prediction for Focused Web Crawlers”, in the proceding of 2012 IEEE/ACIS 11th International Conference on Computer and Information Science

[4] S. Lawrence and C. L. Giles. Searching the World Wide Web. Science, 280(5360):98.100, 1998.

[5] Pavalam S M, Jawahar M, Felix K Akorli, S V Kashmir Raja “ Web Crawler in Mobile Systems”in the proceedings of International Conference on Machine Learning (ICMLC 2011), Vol. , pp.

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

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

[8] Steven S. Skiena “The Algorithm design Manual” Second Edition, Springer Verlag London Limited, 2008, Pg 162.

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

[10] Alexander Shen “Algorithms and Programming: Problems and solutions” Second edition Springer 2010, Pg 135

[11] Narasingh Deo “Graph theory with applications to engineering and computer science” PHI, 2004 Pg 301

[12] Kim, S. J. and Lee, S. H. "An improved computation of the PageRank algorithm” in Proc. of the European Conference on Information Retrieval (ECIR', 2002, pp. 73—85).

[13] Debashis Hati, Biswajit Sahoo, A. K. "Adaptive Focused Crawling Based on Link Analysis," 2nd International Conference on Education Technology and Computer (ICETC), 2010.

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

[15] Serge Abiteboul, Mihai Preda, G. C. "Adaptive On-Line Page Importance Computation," Second International Conference on Emerging Trends in Engineering and Technology. ICETET-09,2009. Fox, C. "A stop list for general text," SIGIR Forum (24:1-2), r 90, pp. 19--21.

[16] Ho, T. K. "Stop Word Location and Identification for Adaptive Text Recognition, Int'l," J. of Document Analysis and Recognition (:3), 2000, pp. 16--26.

[17] Asuncion Honradot, Ruben Leon, R. O. D. S. "A Word Stemming Algorithm for the Spanish Language".

[18] Hull, D. A. and Grefenstette, G. "A Detailed Analysis of English Stemming Algorithms", Technical report, Xerox Research and Technology, 1996.

[19] Miller, G. A. "WordNet: A Lexical Database for English," Communications of the ACM (Vol. 38, No. 11), 1995, pp. 39-41.

[20] Liu, H. & Singh, P. (2004) ConceptNet: A Practical Commonsense Reasoning Toolkit. BT Technology Journal, To Appear. Volume 22, forthcoming issue. Kluwer Academic. Publishers.

[21] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In Proceedings of the 16th international conference on World Wide Web (WWW '07).