OPEN ACCESS

International
Journal
Of Modern Engineering Research (IJMER)

Review of Focused Crawler Algorithms

Rinki Tyagi Mrs. Deepali Dev
Student, A.B.E.S Engineering College
Asst. Professor, A.B.E.S Engineering College

ABSTRACT:- Focused web crawler is a type of crawler that only collects relevant pages and discards the irrelevant one. Because of searching according to the query given by user it is also known as topic specific crawler. In this paper we prepare a survey of focused web crawler, important terminologies used by a focused web crawler, architecture of focused web crawler, types of focused web crawler, various algorithms of a focused web crawler and the challenges faced by it.

Keywords:- focused crawler, blind search algorithms, heuristic search algorithms.

I. INTRODUCTION

As WWW is growing rapidly so extracting relevant information from the web is a big task to perform, Web crawler solves this problem of World Wide Web (WWW). Now there is a question in our mind that what is a web crawler??? Web crawler is basically a program that is used to extract relevant pages from web. There are millions of pages on web and extracting the relevant information from the web is very difficult. Focused web crawler is a type of web crawler which overcomes this problem [4]. Focused web crawler extracts the relevant information from the web and discards the irrelevant one. It traverses the web like a graph where nodes of a web graph represents web pages that contain relevant information and edges of graph represent hyperlinks.

This crawler is a tool i.e. used very much in our [1] society to extract the information according to the need of user. It extracts the information from web just like information from a database. It is also known as specific web crawler or vertical web crawler or topic specific crawler. In this paper we just prepare a survey of focused web crawler, some of the important terms of focused web crawler, its working, architecture, types of focused crawler and various algorithms used by it.

Nowadays because of some illegal activities some websites implement some rules on every visitor and they watch activities of every visitor and for this they have some tools and these tools are used to watch the activities of every visitor and if there is any violation of any activity then these tools restrict those websites and if this type of activity is done by the IP address [3] of any university or by IP address of any college server then these addresses gets blocked by these tools and then no one can access these websites ever and website restriction can be found in robot.txt file. Focused web crawler extract the relevant information from the web by navigating the web like a graph. Web crawler consists of root directory of all of the web sites[3].When robot.txt file is not present then we cannot conclude that there are no restrictions. The tools used to monitor the websites still can see the activities of every visitor and can check the date transfer rate and also can compare it with the allowed maximum data transfer rate for every visitor.

In the next section of this paper we are going to describe some important and very basic terms used by a crawler.

II. BASIC TERMS USED BY FOCUSED WEB CRAWLER

Some important terms used by focused web crawler are [1]:

HARVEST RATIO:

It is defined as the ratio of number of relevant pages to the total number of pages downloaded from the web. Harvest ratio also helps in predicting the relevancy of the page. If harvest ratio is high then relevancy of page will be high and vice-versa.

PARENT PAGE:

It is defined as the page through which link of relevant page is extracted.

TARGET PAGE:
The pages retrieved from parent page is known as target page. Another name of target page is child page.

HUBS:
Hubs are pages consisting of number of links to another page and another pages may consist of relevant information.

AUTHORITIES:
Authorities are those pages that have high relevancy.

RELATIONSHIP BETWEEN HUBS AND AUTHORITIES:
Hubs consist of authorities (highly relevant pages) and authorities are pages that have high relevancy.

CRAWLER FRONTIER:
It is a list of URL that has to be scanned. It is also known as URL frontier.

TUNNELING:
As web is very much complex and extracting relevant information is a big challenge. Sometimes it has been seen that irrelevant pages consist of some links to relevant pages, so a crawler have to navigate irrelevant pages also and the whole process is known as tunneling.

SEED SET:
It is a set of URL that consist of highly relevant information about a specific topic

III. FEATURES OF FOCUSED WEB CRAWLER

In this section we are going to describe some of the features of focused web crawler and they are described as follows [7]:

ROBUSTNESS:
As we know web contains number of pages known as web page and server consist of number of web pages. Sometimes, server gives number of faulty web pages in which crawler can be trapped. So, crawler does not respond to those web pages. The design of the crawler is like that it does not respond to that faulty page.

SCALABLE:
Crawler should provide scalability to provide freshness to the user and for this, high throughput and high bandwidth is required.

PERFORMANCE & EFFECIENCY:
Crawler should have high performance and high efficiency. In order to improve performance and efficiency harvest ration should be increased, time can be speed up and cost can be reduced, accuracy can be improved. So, there are number of terms on which performance and efficiency depends.

QUALITY:
When user input query into the crawler, crawler searches for the most relevant pages using some algorithms, if crawler provides most relevant pages first then quality is high otherwise quality is low. So, providing highly relevant pages to improve quality is a big challenge.

FRESHNESS:
Sometimes we saw that crawler searches for the web page continuously. So, providing fresh and updated information continuously is a big challenge but crawler always provides fresh web page to the user and crawler always eliminates redundant information.

IV. ARCHITECTURE OF FOCUSED WEB CRAWLER

In this section we are going to describe architecture of focused web crawler:

Firstly user input query to the crawler and seed URL is visited. Seed URL also known as set of URL's contained in URL frontier. So, crawler frontier consists of a set of URL's from the WWW. The crawler starts working with the seed URL i.e. extracted from WWW. So, crawler extracts URL from the set of URL's that are not yet visited. This process continues until the crawler frontier becomes empty. Page downloader is also used to download the web page from WWW. HTTP protocol [10] is required if client need to send a HTTP request.

The use of web repository is that it is used for saving and managing the large amount of data and it mostly stores HTML pages. In crawling strategy relevancy is calculated and URL priority is assigned using some algorithms [5-8] such as Naïve Best First Search algorithm, Page Rank algorithm, shark search algorithm [10], fish search algorithm [10], T-Page Rank algorithm [8] and many more defined in section 6 of this paper and this section is very important in this paper.

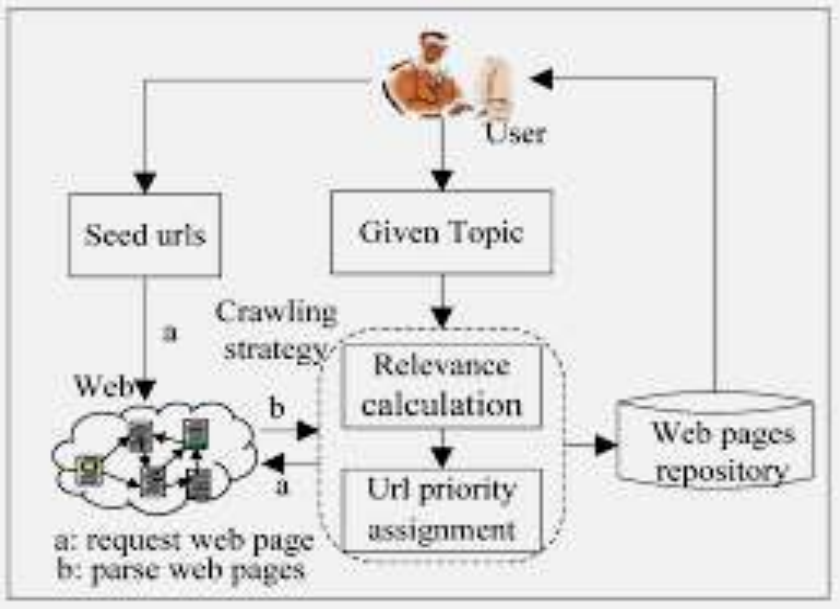

Figure (1): Focused Crawler Architecture

V. TYPES OF FOCUSED WEB CRAWLER

In this section of paper we will describe several types of focused web crawler and concept that how they work, their advantages and disadvantages and various challenges faced by them. Some types of focused web crawlers are as follows:

VSM CRAWLER:

It is a type of classic focused web crawler. Here number of pages having higher priority is downloaded first and the queue is maintained as priority queue, pages are maintained in priority queue and the pages having high relevancy is shown to the user. This crawler works on the concept of vector space model, vector space model only checks the similarity between the terms and vector space model takes the advantage of cosine similarity matrix. Terms are calculated based on TF*IDF values and weights are also calculated to compute the similarity between the term and topic [2].

SEMANTIC WEB CRAWLER:

This model is just a variation of classic focused web crawler [5]. Here the terms are considered similar if they consist of same meaning. So, it overcomes the disadvantages of vector space model but it does not checks the similarity between the terms instead it checks the similarity in terms of meaning. So, it is a disadvantage that it does not check the similarity in between terms. When user input any query then relevancy of the page is computed in terms of meaning and the pages having higher priority is downloaded first.

HAWK:

This crawler is used to find the most relevant pages. This crawler uses the concept of content and link analysis [6]. It avoids the concept of searching the pages which are not relevant. Here priority of URL is computed is placed in queue so that page relevancy is computed and boundaries to search topic is also increased. It uses the concept of relevancy, shark search and page rank algorithm. Here queue consist of three types of queue: first one saves the URL that are not yet visited, second one queue is used to save visited URL, and that one is queue is used to save the URL's that are useless.

SEMANTIC SIMILARITY VECTOR SPACE MODEL CRAWLER:

We have seen that both type of crawler have some disadvantage i.e. VSM crawler cannot search in terms of meaning and semantic web crawler cannot search only similarity between terms instead it checks the similarity between terms in terms of meaning. So a new crawler is prepared by combining both the crawlers (VSM crawler and semantic crawler) [2] to overcome the disadvantages of both type of crawlers. Harvest ratio is increased and average error rate gets decreases in this crawler.

VI. ALGORITHMS OF FOCUSED WEB CRAWLER

Various types of algorithms used by focused web crawler [4-7] are defined as:

BREADTH FIRST SEARCH:

This algorithm comes under blind search approach. Here, crawler searches for the pages level by level. In this algorithm, if a new URL is found then it is added at the tail of the queue and the URL's are maintained in queue by using the concept of first in first out or first come first serve basis. If any URL is searched by the crawler which consist of useless pages and if those useless pages also consist of some useless links then crawler will travel these useless links also. So, it blindly searches for the web pages without checking it's relevancy that's why it is known as blind search algorithm.

NAIVE BEST FIRST SEARCH:

It is a basic heuristic search algorithm. In this algorithm a queue is maintained as a priority queue. It will only navigate through the web pages according to priority assigned to the URL's in the queue. It overcomes the disadvantage of BFS algorithm because it does not traverse the pages blindly, it firstly checks [11] the relevancy of web page by maintaining priority queue. So harvest ratio gets improved.

PAGE RANK ALGORITHM:

It is the first technique to check the relevancy of web page by using the concept of link analysis. This technique is assigned to every node in a web graph. It consist a numerical value between 0 and 1 that shows the rank of the page. It depends on link structure [8] of web graph but if there is not any link present in the outward direction in the web graph then what will happen?? To overcome this problem a teleport operation is used. Teleport operation is defined. In this operation the surfer jumps from a node to any other node present in web graph. If N is the total number of nodes present in web graph the teleport operation takes the surfer to each node with probability 1/N. the formulae used to check the page rank of web page is given below and the formulae is:

$$\mathrm{PR(p)} = (1-\mathrm{d}) + \mathrm{d}\left(\frac{\mathrm{PR(t1)}}{\mathrm{outlink(t1)}} + \; - - - - + \frac{\mathrm{PR(tn)}}{\mathrm{outlink(tn)}}\right)$$

Where: PR(p)= page rank of p.
d= damping factor (usually set to 0.85).
outlink(t1) = number of outward link from page t1.

FISH SEARCH ALGORITHM:

This is the first heuristic search algorithm which searches the web pages dynamically. It is based on content analysis. In fish search algorithm web is considered as a directed graph and web pages are considered as nodes of a graph and hyperlinks are considered as edges of graph. This algorithm is known as fish search algorithm because here internet is assumed as a big sea and crawler is assumed as a fish which searches their foot to eat. If food is found to eat then we get the relevant pages and if food is not found it searches the food until food is found by fish. This algorithm predicts the relevancy of pages in terms of binary values only i.e. 0 and 1. Here queue [11] is maintained as a priority queue and list of URL's is maintained in queue according to their priority. So pages having higher relevancy is downloaded first and irrelevant pages gets discarded by the crawler. Limitation of this algorithm is that it gives only 0 and 1 as output. Sometimes it happens that some pages gets same relevancy and it consists of only 0 and 1 values to check the relevancy so it does not differentiate about relevancy of web pages and sometimes discard relevant pages also.

SHARK SEARCH ALGORITHM:

Shark search algorithm is a variation of fish search algorithm. It is a dynamic search algorithm. It is based on the concept of content analysis. It gives values between 0 and 1 instead of giving binary values. It overcomes the limitation of fish search algorithm i.e. fish search algorithm does not differentiate finely about relevancy of web pages but because of using values between the range of 0 and 1, it gives most relevant pages and performance of focused crawler get improved [11] by using this algorithm as compared to fish search algorithm.

T-PAGE RANK ALGORITHM:

This algorithm is an advance version of page rank algorithm. It is a topic sensitive crawler which searches the web page by checking its sensitivity [8]. This algorithm may help in determining that which site is a good match for a specific query. This algorithm can make difference between a first page position and ranking a site at the top of first page. This algorithm works more efficiently than page rank algorithm so Google also use this algorithm in ranking the web pages or web sites..

SHARK PAGE RANK ALGORITHM:

As the name suggests this algorithm is prepared by combining shark search algorithm and page rank algorithm [9]. As shark search algorithm computes the relevancy in the range of 0 and 1 and page rank algorithm search the ranking of adjacent web pages on the basis of link analysis so that if the links consist of relevant information so they can be downloaded and by combining both it has seen that this algorithm replace it's ancestor algorithms. So this algorithm works on the concept of content and link analysis. This is the latest search algorithm which can further be improved..

VII. CONCLUSION & FUTURE WORK

Here we consider various types of focused web crawler and how they work. In this paper we come across several algorithms of focused web crawler and limitations of these algorithms also. This paper consists of survey or detailed study of focused web crawler. So from this paper we conclude that we can further research in shark page rank algorithm by increasing its performance in terms of harvest ratio or we can also go into the deep study of semantic similarity vector space model crawler using the concept of ontology.

REFERENCES

[1]. Sameendra, samarawickrama ,Lakshman Jayaratne ,Automatic Text Classification and Focused Crawling, 978-1-4577-1539-6/11, 2011 IEEE.
[2]. Yajun Du, Wenjun Liu,Xianjing Lv, Guoli Peng,.An Improved focused Web Crawler Based On Semantic Similarity Vector Space Model 1568-494/Elsevier
[3]. Quan Bai, Gang Xiong, Yong Zhao, Longtao He. Analysis and Detection of Bogus Behaviour In Web Crawler Measurement. Procedia computer science 31(2014) 1084-1091.
[4]. Promila devi, Ravinder Thakur Comprehensive Review of Web Focused Crawling. IJICT, vol 5(5),2014,6035-6038.
[5]. Nidhi Jain et al,/(IJCSIT)International Journal of Computer Science And Information Technologies, Vol.4(3),2013, 398-40. A Study of Focused Web Crawlers for Semantic Web.
[6]. Xiaoyun Chen, Xin Zhang, HAWK: A Focused Crawler with Content and Link Analysis. DOI10.1109/ICEBE.2008.46 (2008 IEEE).
[7]. Mini Singh Ahuja, Dr, Jatinder Singh Bal, Varniea, Web Crawler: Extracting the Web Data. International Journal of Computer Trends and Technology(IJCTT)-volume,13 number, july 2014.
[8]. Fuyong Yuan, Chuniax Yin, Jian Liu, Improvement of Page Rank for Focused Crawler.0-7695-2909-7/07 (2007 IEEE), DOI 10.1109/SNPD.2007.458.
[9]. Jay Prakash, Rakesh Kumar, Web Crawling through Shark-Search using Page Rank. Department (ICCC-2015).
[10]. Yugandhara Patil, Sonal Patil, Review of Web Crawler with Specificatiom and Working. IJARCCE Vol. 5, Issue 1, January 16.
[11]. Andas Amrin, Chunlei Xia, Shuguang Dai, Focused Web Crawling Algorithms. Journal of Computers, Volume 10, Number 4, July 2015.