Comparative Analysis of Focused Web Crawling Schemes

[1]Fatima Ziya, [2]Suresh Kumar
[1]Research Scholar, [2]Associate Professor
[1]Dept, CSE
[1]Ambedkar Institute of Advanced Communication and Research, New Delhi, India

Abstract : **Rapid development and availability of abundant data on web makes difficult in searching over the web. Moreover this problem becomes more complex due to increasing growth in web user. So there is a need to improve or design search algorithms that helps in efficiently and effectively searching the specific data from web. There are various kinds of crawlers used to fetch web pages like parallel crawler, Incremental crawler, Distributed Crawler and Focused web crawler. Apart from Focused crawler all crawlers are crawling irrespective of topic which result in more number of irrelevant pages , whereas Focused crawler crawls only topic specific web pages based on user query. Aim of focused crawler is to search the topic specific subset. In this paper, we have presented a comparative analysis of various Focused crawling schemes based on parameters as relevancy, refresh policy, effectiveness and complexity etc.**

IndexTerms Focused Crawler, Crawling Algorithms, Search Engine

I. INTRODUCTION

Traditional web search engine are based on breadth first or depth first traversal algorithms. They are commonly used to index the web. The elemental set of Uniform Resource Locator (URL) are used as a seed set and the schemes are used to connect the hyperlinks of pages. It begins with one or more URL's that constitute a seed set. The crawler select a URL from this seed set and then extract the web page from that URL. The main objective of the crawler is to cover the whole web. A crawler follow each link using breadth first strategy .It tries to identify the most relevant links , and ignore the irrelevant document[2].Focused Crawling was first introduced by Chakrabarti in 1994 [3].One of the first crawler web crawler was proposed by (Cho J et al) and they introduced the best first strategy. Search engine are totally depend on the crawling process, it fetches the quality of results .The crawlers have to work continuously to keep the page repository refresh. Web crawlers have been grouped into different categories as :-[4].

- Focused Crawler
- Parallel crawler
- Distributed crawler
- Incremental crawler

Parallel Crawling :In this crawling , search engine use multiple crawler in parallel to perform an efficient crawling.It downloaded the pages from the web store the pages locally and extract the URL's from the downloaded pages and then uses these URL's for further crawling.[4]

Distributed Crawler. In this crawling technique, the crawler uses the internet and crawl at distant locations. As the number of requests increases , the network traffic is distributed, in order to have the most coverage of the web. Distributed crawler is distributing geographically so that the center server transfer the communication and synchronization of the nodes easily. But the problem is that, if the center server fail to communicate then the nodes will not receive the information.

Incremental Crawler: In this crawling scheme a crawler continuously update the existing collection of pages by visiting them again and again. It periodically replaces old document with the newly downloaded. It resolve the problem of freshness but it does not retrieve those web pages which is related to domain specific or topic specific.

Focused Crawling: It is also known as Domain specific crawling. The crawler choose intelligently which links to follow and which links to discard. The main focus of this crawling scheme is to target the relevant page related to pre-defined topic and the way to proceed further. The topics not only using the keywords but also collecting the exemplary pages and indexing all the accessible web pages. It tries to download the pages which are interlinked to each other. The main goal of this Crawler is to target,how the given page is relevant to the particular topic and how to proceed forward [2 4] ?.

.A focused crawling has the following advantageous as compare to the others crawlers.

1) This crawling scheme fetches the most relevant web pages while the other crawlers are not capable of downloading the relevant page although of provide same seed.
2) It has a look-ahead strategy to fetch many web pages/ links away from the seed set, It can build high quality of web documents.

The brief description of article is as follows: Second section is about introduction and architecture of Focused crawler. Third section is about comparisons of existing focused crawling algorithms with their strength and weaknesses and time complexity as shown in Table 1.

II. ARCHITECTURE OF FOCUSED CRAWLER

A Focused Web Crawler (FWC) follows a procedure in such a way that each link connects with another link along with its score in the pages. It is designed in a way, it gather only the relevant documents on a specific topic. FWC designs its crawl boundary to find the useful hyperlinks for the crawling and avoid the useless hyperlinks of the web. It entails very less investment in the network and hardware resources .The architecture of FWC is shown in Fig 1 below. It consists of three main components known as Classifier, Distiller and Crawler Manager [5].

The role of Classifier is to decide the page relevancy to determine its further link expansion. It directly affects working efficiency of the crawler. Classifier gives instructions to FWC related to relevancy and taxonomies of the pages. The second component is Distiller, it identifies topical viewpoint of the pages and assign the authority of the pages [3].

Crawler: It is also known as spider of the web. It continuously sends request to the Web to fetch the web page against a particular URL. Then, the requested downloaded pages are stored in crawler repository. This repository further sent to the indexer for indexing. It is a continuous process of sending URLs to web and downloading web pages in a circular manner. Search engine and many websites make use of crawling for providing the latest data. Focused crawler uses the best strategy to download the most relevant pages based on some criteria [5].

Seed Detector: The role of seed detector is to collecting the seed URL's with the help of specific keyword by fetching the list of URL's. After retrieving the pages it assign a priority based on page rank algorithm. FWC makes indexing more effectively. From the large repository of web, FWC achieves the more retrieval information [10].

Crawl Manager: This component of search engine acts as a manager. The role of crawl manager is to overcome the speed of the crawler as well as balancing the load. It controls all the crawlers also checks duplicate web pages. The crawler fetches the URL's from URL repository and store into the buffer. If the size of the buffer is full then the crawler manager dynamically generates the replica for the crawlers, which will download the documents. To maintain the efficiency, Crawl Manager create a crawling pool.

The most essential analysis of focused crawler is to quantify the harvest ratio. Harvest Ratio is defined as the rate in which relevant pages are obtain and irrelevant pages are effectively filtered off the crawl. FWC spends more time to discard, the irrelevant pages while the others crawlers are not spending efficient time to eliminate the irrelevant information. To improve the relevancy of the pages Focused crawler uses many algorithms, some algorithm are discussed in section III [5 6].

- Fish Search Algorithm
- Shark Search Algorithm
- Context Graph Algorithm
- Info-Spider Algorithm
- Page Rank Algorithm

III. FOCUSED CRAWLER SCHEMES

Some web crawling algorithms are discussed below and also compare these algorithms along with their strength and weakness in table 1.

A. **Fish Search Algorithm**. It was the earliest algorithm used by focused crawler. The aim of using this scheme is to create the efficient web pages. Fish search algorithm was implemented for dynamically search information. The search agents continuously found the more relevant information. It automatically navigates which webpage to crawl. The key principal of the algorithm is as follows [2].

- The first step is to take the input which is a seed URL.
- After taking the URL,it dynamically create the priority list of the next URL called nodes.
- At every step the first node is extract from the list and processed. As the text becomes available, it is assigned and processed using a scoring tool the relevant information is assigned to (1) or irrelevant to (0) .The source document is fetched and scanned for links. The nodes pointed to these existing links (denoted as children).
- If the parent URL is relevant, the depth of the children URL is assign to predefined value. Else, the depth of children URL is set to be one less than the depth of the parent URL.
- When the depth reaches zero, the direction is dropped and none of the children URL is inserted into the list.

The Fish search has some drawbacks of the search:

- The algorithm consume more time for crawling and downloading web documents by www which is unacceptable for users.
- This algorithm cannot search documents from the hidden web. It retrieve only those pages for which the URL's are found.
- The Time complexity of this algorithm is terrifically high and the efficiency of this algorithm depend on the ability to find more relevant pages and avoid irrelevant pages.

B. **Shark Search Algorithm** : It is a refined version of fish search . In the same exploration time shark search algorithm discover and retrieve the more relevant information. It is based on the schools of fish metaphor. This algorithm uses a binary value rather than binary evaluation. This approach gives better result, it uses a score between 0 and 1 (0 for no match and 1 for perfect match). Shark search algorithm uses inheritance method of node's children,that have a great impact on relevance score, it gives an inherited score of the children and the children's children. It overcome the drawback of fish search algorithm [8 10].

C. **Best Search Algorithm** : As the name implies , this algorithm can be used to search the best URL which explore the URL queue to find the most efficient URL. It's an algorithm that uses a score to define which page has a best score. It is based on a heuristic function f(n).The value of a heuristic function is present in a queue accessed for each URL .The highest priority list and the URL with the highest priority can be fetched . Best Search Algorithm fetched the optimal URL but it consume more memory and time [8].

D. **Context Graph Algorithm** : It is based on the concept graph of the semantic content of relevant pages .The data of this algorithm is represented as a graph which is related to each other and cover a minimum distance to move from one graph to another .It uses the limited capacity of search engines. This crawling scheme has two main concepts (i) In the starting phase s set of context graph and classifiers are constructed for each of the seed documents. (ii) A crawling phase which is using classifier to lead a crawling process and update context graphs [2 13 16].

E. **Info Spider Algorithm**: are a dynamic and multi-agent system. It is based on online agent system that can search on behalf of the user. These agents communicate independently to each other and then try to cover a specific area of the relevant document with the help of artificial intelligence techniques .When a user enter a query, it obtain a set of seed links and then agent examine the links by computing the similarity of the text around [1 2].

F. **Page Rank algorithm** (PR): is an algorithm used by Google Search to rank the web pages . Page Rank Algorithm was proposed by Sergey Brin and Larry Page founders of Google. Page Rank Algorithm is used for measuring the rank of website pages. The basic concept of the algorithm is that importance of the page is directly proportional to the number of web page. According to google page rank,Page is important if more number of other web page linking to that page. The links to a page can be classified into the following types: Inbound links which are links into the given site from outside. Outbound links which are links from the given page to pages in the same site [14].

The rank of a page is calculated as a sum of the PageRanks of all pages linking to it divided by the number of outgoing links on each page.

- PR[M]: PR of page M
- PR[Z_i] : PR of pages Z_i which link to page M.
- C[Z_i] : is the number of outbound links on page Z_i.
- D: is the damping factor which can be set between the value 0 and 1.It depend on the no of clicks , where n is the number of inlinks of page M [9 10].

$$PR(M) = (1-d) + d\sum_{i=1}^{n}\frac{PR(Zi)}{C(Zi)}$$

Eq.1

Page Rank algorithm does not rank the whole website, but it is determined for each page individually .The main disadvantage of page rank algorithm is that it is calculated and stored at the time of indexing and not the time of query so that the relevancy of resultant page to the user query is very less.

IV .Comparative Analysis of Focused Crawling Schemes Table 1

In table 1, we have analyzed all the Focused Crawling algorithms based on difference parameters like efficiency, relevance, Speed, Refresh policy and Complexity.

- **Efficiency** : In these algorithms efficiency is measure how long the algorithm finds the more relevant pages and avoid the irrelevant pages.
- **Relevance**: Relevancy of these algorithm is defined as to cover a good coverage of links or to cover the whole content of web page.
- **Refresh Policy**: Refresh policy is defined as the freshness of the pages.
- **Time Complexity**: Time Complexity is defined as the total time to which the crawler finds the more relevant pages.

V .Conclusion and Future work

A Focused crawler is essential for a topic based search. It downloads web pages that are related to the specific topic. Various crawling algorithms are being used for searching process. In this paper various crawling schemes are discussed along with their strength and weakness. Fish Search algorithm is the first algorithm used for crawling but it cannot search the hidden web. The main challenge in Focused Crawler is to maintain the efficiency and freshness of the pages. Google search engine using PageRank algorithm to improve the efficiency but in PageRank algorithm has its own limitations as relevancy of the resultant pages to the user query is very less. The future work is to implement and design the novel algorithm to improve the searching result obtained from the focused web crawler. Thus the focused web crawler proves to have better performance than any other web crawlers.

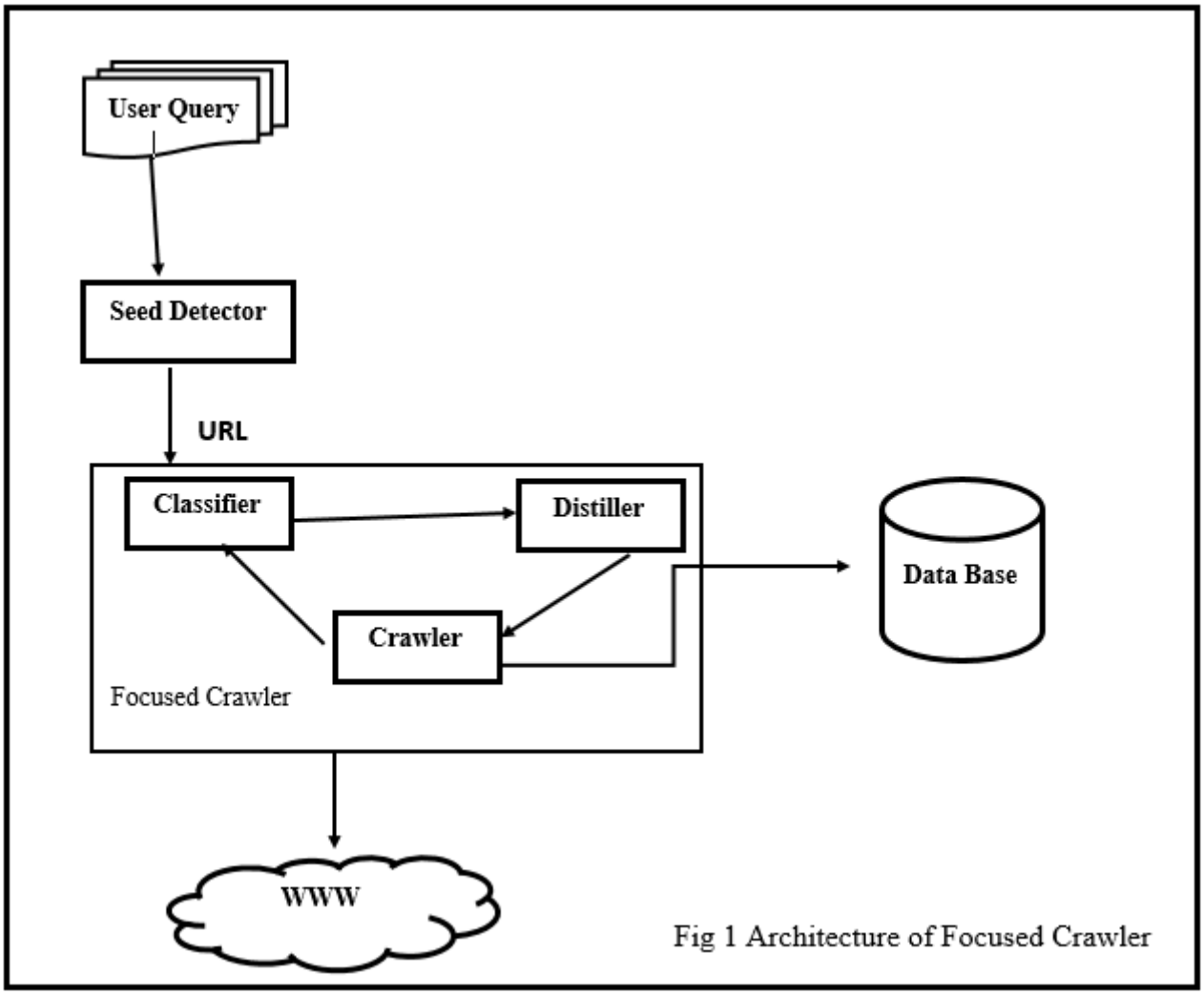

Fig 1 Architecture of Focused Crawler

Table 1 Comparative Analysis of Focused Crawler Schemes

Crawling Algorithms	Principal	Strength	Weakness
Fish Search Algorithm	Fish School Metaphor	• High Dimensional Search • Local Computation Autonomy	• Time Consuming. • Cannot search hidden web. • Usage is terrifically high.
Shark Search Algorithm	Fish School Metaphor	• Improved Version of FSS • Better Relevance • Improvement of inheritance	• Insufficient Context trapping.
Best Search Algorithm	Best Score Pages	• Search the best URL • Simple to implement	• It is a Blind Search when the search space is large. • Time Consuming.
Context Graph Focused crawling	Minimum Distanced Based	• Limited Capacity of traditional search • High Relevance	• Less Guidance. • Context is too large.
Page Rank Algorithm	Page Rank	• It takes less time and fast. • More flexible	• Give priority to older pages. • Relevancy is low. • Dangling links.

VI. REFERENCES

[1] Andas Amrin , Chunlei Xia and Shuguang Dai "Focused Web Crawling Algorithms" journals of computer , May 21 2015.
[2] Blaž Novak and Jozef Stefan Institute "A SURVEY OF FOCUSED WEB CRAWLING ALGORITHMS"
[3] Vruksha Shah, Riya Patni , Vivek Patani, Rhythm Shah "Understanding Focused Crawler"International Journal of Computer Science and Information Technologies , 2014
[4] Satinder Bal Gupta "The Issues and Challenges with the Web Crawlers" International Journal of Information Technology & Systems, Vol. 1 2012.
[5] Dvijesh Bhatt, Daiwat Amit Vyas and Sharnil Pandya "Focused Web Crawler" Advances in Computer Science and Information Technology (ACSIT), 2015
[6] Vruksha Shah, Riya Patni , Vivek Patani and Rhythm Shah "Understanding Focused Crawler" International Journal of Computer Science and Information Technologies , 2014.
[7] Simarjeet Kaur "Search Engines and SEO: Need and Working" International Journal of Research and Cultural Society ,2018
[8] Bireshwar Ganguly and Rahila Sheikh "A Review of Focused Web Crawling Strategies" International Journal of Advanced Computer Research,2012
[9] Apoorv Vikram Singh , Vikas , and Achyut Mishra "A Review of Web Crawler Algorithms" International Journal of Computer Science , 2014
[10] Pavalam S M, S V Kashmir Raja, Felix K Akorli and Jawahar M "A Survey of Web Crawler Algorithms" International Journal of Computer Science,2011
[11] Promila Devi and Ravinder Thakur "Comprehensive Review of Web Focused Crawling"International Journal of Computer Science and Information Technologies, Vol. 5 (5) , 2014
[12] Bin Wanga, Junhao Wanga, Xiaohua Sunb, Na Wangc "An Improved Shark-Search Algorithm for Agriculture Web Search Engine"Chemical Engineering Transactions ,Vol 51,2016
[13] Md. Hijbul Alam · JongWoo Ha · SangKeun Lee "Novel approaches to crawling important pages early", 2010
[14] Sanjay and Dharmender Kumar"A Review Paper on Page Ranking Algorithms" International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Vol 4, 2015
[15] K. R. Srinath" Page Ranking Algorithms – A Comparison" International Research Journal of Engineering and Technology (IRJET) ,2017
[16] Sung Jin Kim, Sang Ho Lee "An Improved Computation of the PageRank Algorithm "
[17] M. Diligenti, F. M. Coetzee, S. Lawrence, C. L. Giles and M. Gori "Focused Crawling Using Context Graphs" International Conference,2000
[18] Soumen Chakrabati Martin van den Berg and Byron Dom "Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery"
.