Adapted Web Crawler for Mining Offline Web Data using AFHC

S.Amudha
Assistant professor, Department of BCA,
VLB Janakiammal College of Arts and Science,
Coimbatore, Tamil Nadu, India

ABSTRACT
Adaptive focused hyperlink crawler (AFHC) aim to search the entire inner level sub link of the web pages related to a specific topic and to download unique web pages to local disk. The download web page information searched in offline browsing and avoids the repeated searches in the web server to give a solution to problem. The major problem is to retrieve the maximal set of relevant and quality pages. Crawler software can be retrieve web pages by hyperlinks through internet. The number of web sites using web crawlers cannot retrieve the relevant pages. The system have browser and search engine. Browser using AFHC reduce time to finding accurate content in web pages in the hyperlinks and also restrict to download web pages to local disk. Search engine using extended cocitaion algorithm to retrieve accurate content in the local disk and search based on any word, all word and phrase matching in the local disk. It is useful to the student and organization. The research is easily found to personalize the crawl history and search history for knowing the user transactions.

Keywords: Web Crawler, Focused Crawler, Web Mining.

1. INTRODUCTION
A Web Crawler searches through all the Web Servers to find information about a particular topic. However, searching all the Web Servers and the pages, are not realistic, because of the growth of the Web and their refresh rates. To traverse the Web quickly and entirely is an expensive, unrealistic goal because of the required hardware and network resources [8].

Web is network of interconnected Hyperlinks on Internet. Web is accessed with the help of web browser. To reach particular page most of the user use search engines. Web Crawler is a main component of Search Engine. Web Crawler automatically browses Web and downloads information for Search engine [7].

The traditional process of focused web crawler is to harvest a collection of web documents that are focused on the topical subspaces [2]. They traverse the web collecting only relevant data to a predefined topic while neglecting on the same time off-topic pages. The crawler is kept focused through a crawling strategy which determines the relevancy degree of the web page to the predefined topic and depending on this degree a decision is made whether to download the web page or not [1].

In proposed approach, enter the URL to the browser. AFHC compute the URL score is based on topical relevancy of parent page block and all related information about topics refer to entire child pages. The relevant pages download to the local system. Search engine searched information in offline, the data retrieved locally and three option method of searching information.

2. LITERATURE REVIEW
Web search engines employ crawlers to continuously collect web pages from the web. The downloaded pages are indexed and stored in a database. This continuous updating of database renders a search engine more reliable source of right and updated information [4].

Focused crawler uses link structure of documents as well as keyword based similarity of pages to the topic in order to move slowly the web [6]. The internet contains hundreds of thousands of electronic collections that often contain high quality information. The basic aim is to select the best collection of information for a particular information need. The indexing phase of search engines can be viewed as a Web Content Mining process [5].

The crawler travels sample seed websites and their derived (children or linked) websites in sequence to collect experimental web pages, from which a set of relevant URL are grouped based on pre-defined topics. The relevancy is calculated between seed page and child page of seed page. Then URL in relevance group and irrelevant group based on its relevancy score with seed page. Focused crawler to crawl the Internet to find topic-related web pages for the end users based on distance score of URLs which is to be fetched and fetched related group URLs [9].

A working process of a focused crawler is composed of two main steps. The first step is to determine the starting URLs and specify user interest. The crawler is unable to traverse the Internet without starting URLs. The second step in a focused crawling process is the crawling method. In theoretical point of view, a focused crawler smartly selects a direction to traverse the Internet. A clever route selection method of the crawler is to arrange URLs so that the most relevant ones can be located in the first part of the queue. The queue will then be sorted by relevancy in descending order. The performance and efficiency of a focused crawler is mainly determined by the ordering strategy that determines the order of page retrieval [8].

2.1 General Architecture of Web Crawler
To identifies the most promising links that lead to target documents, and avoid off topic searches. In addition, it does not need to collect all web pages, but selects and retrieves relevant pages only. It starts with a topic vector, and for each URL, the relevance is computed for the contribution of web page in the selected domain. If it is found to be important, it gets added to the URL list else, gets discarded [8] [1].
The typical design of search engines is a "cascade", in which a Web crawler creates a collection which is indexed and searched. The basic web crawler is as shown in figure 1. [3].

2.2 Web Crawler Task

A Web crawler has four responsibilities:

1. It selects a URL from a set of candidates.
2. It downloads the associated Web pages.
3. It extracts the URLs (hyperlinks) contained within a Web page.
4. It adds those URLs that have not been previously encountered to the candidate set [10].

3. PROPOSED ARCHITECTURE

The proposed architecture is shown below in figure 2. The users enter the input as URL to the browser and using AFHC algorithm to retrieve relevant pages to local disk. The user can possible to setting their depth level and drive of local disk. Threading and exception handling method used in browser download locally.

Search engine using extended co citation algorithm to retrieve the relevant content on local disk. Searching text entered by the user and selects the three options for searches are any word, all word and exact phrase matching and display the related links as per Google search engine. The Google search engine work online but search engine work offline.

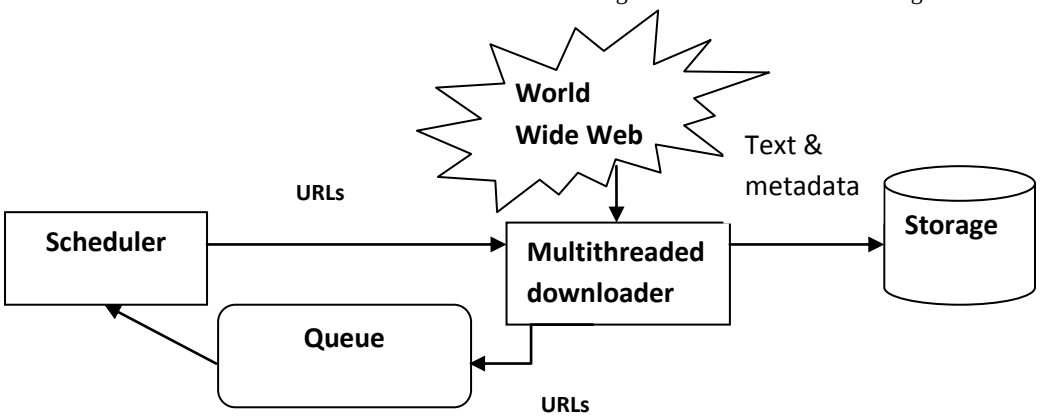


Figure 1: General architecture of web crawler

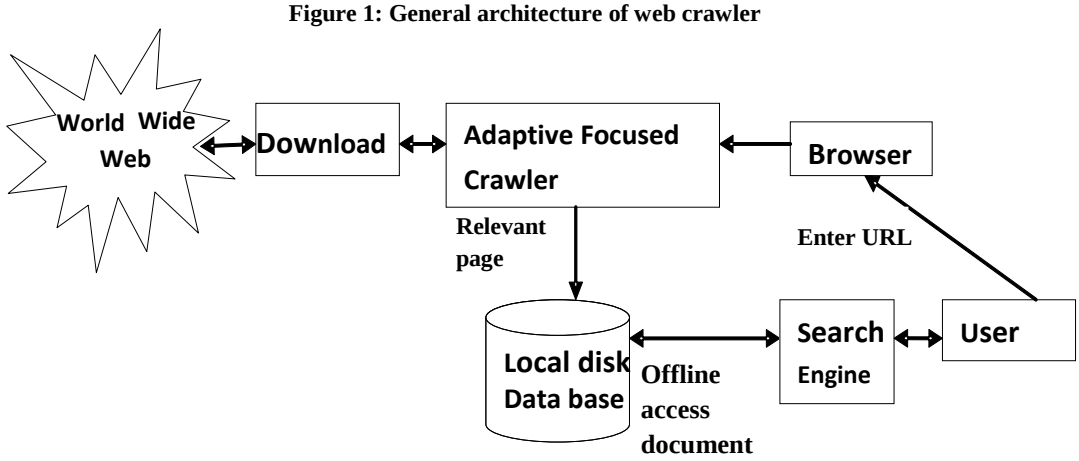


Figure 2: The Proposed Architecture

4. PROPOSED APPROACH

4.1Adaptive Focused Hyperlink Crawler

1. Extract seed pages from User Input.
2. Extract all terms and links from seed page and all sub links.
3. To find the total number of links in web pages
4. To find the total number of parent pages of each link.
5. Calculate the relevancy score of each parent page based on subject keywords until find relevancy page.
6. Calculate Relevancy Score URL _description on entire sub links.
7. Calculate Anchor Relevancy Score.
8. Calculate link score is all out links in a particular page.
9. Based on link score to find relevant page and retrieved first using frontier.

5. METHODOLOGY

5.1 Seed URL Extraction

Seed URLs are extracted by search engine known to entire sub link of URL and URL put a query in the search engine. The search engine result is only relevant to query and these URLs information stored to local disk.

5.2 Frontier

Frontier is initialized to seed URLs. It contains only unvisited URLs. The unvisited URL uses the priority queue and top of the queue is higher priority. A URL which has higher URLs score is given higher priority to the web page downloader.

5.3 Web Page Downloader

A web page downloader is used to enter input URLs. The URL has higher priority from frontier and downloads the web page from internet in online.

5.4 Page Segmentation

The page segmentation is used AFHC algorithm to represent a set of blocks in fetched page. The segmentation process is web partition from suitable extracted blocks.

5.5 Parser and Extractor

The parser and extractor are allows parsing and extracting the terms and the hyperlink from each block of downloaded page.

5.6 Relevance Calculator

To calculates relevance score of block with respect to topic and stores relevance score of blocks in relevance block database. The relevance calculation of all blocks is finished, and then it goes to relevance calculation of page step. Otherwise, it again returns to relevance calculation step to calculate the relevance score of rest of the blocks in particular page.

5.7 Relevance Analyzer

Relevancy analyzer analyses the relevance score of total blocks and then calculates the summation of relevance score of all blocks this is the relevance score of page. Pages have relevance score greater than user Specified limit, only that page is stored in Relevant page DB. Otherwise, the page is discarded. From Block DB, the block's URLs extract. The flow of extracting information from web and offline access is show in figure 3.

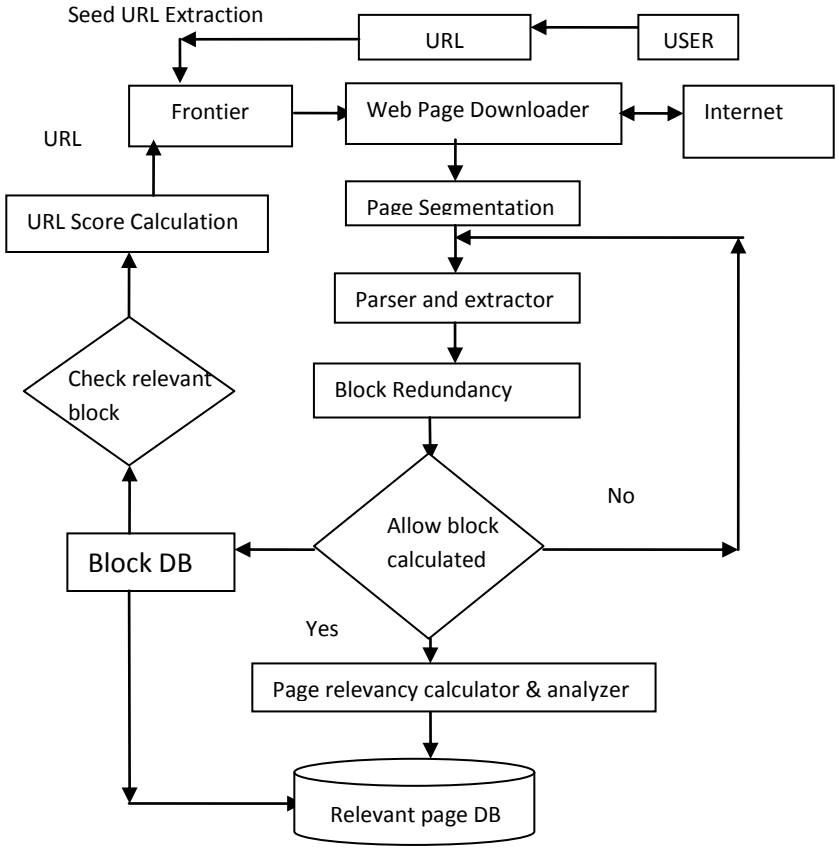


Figure 3: The flow of extracting information from web and offline access

6. IMPLEMENTATION

The implementation has been done using VB.NET and C#. VB.NET used to implement search engine and they have three options. There are any words, all words, and phrase. The search engine is shown in figure 5. The browser implements using C# and work online is shown in figure 4.

6.1 Proposed Browser Work

Enter the particular URL and click go button to download the all type of data like image, text, audio, video, pdf etc. Before downloading data from the Browser user set the depth level of downloading and has based on hyperlink of page. User set the target folder in local computer to store all the download data. Browser support threading at the same time download different type of data.

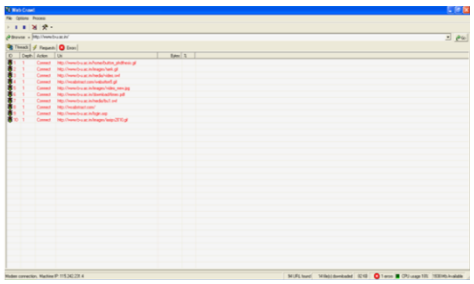


Figure 4: Browser

Examples download image data means create an image folder automatically to our local computer and store all images in our specified URL. Browser display the number of files downloads depth level of the particular link, action of the particular link for connect, download, and complete. The total

bytes download in single link and percentage of downloading content on the internet [figure 4 and 5].

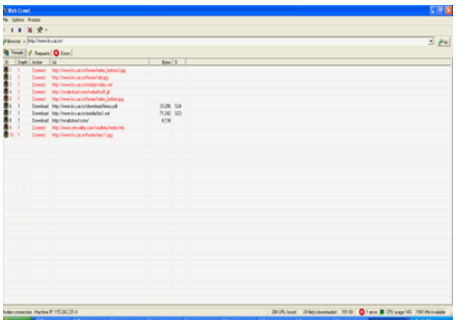
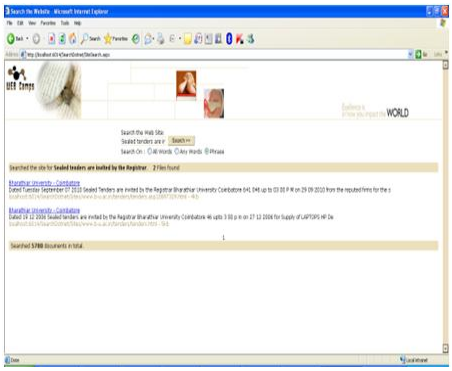


Figure 5: Download Files to Local Disk

6.2 Proposed Search Engine Work

Our search engine using three options are all words, any words and phrase. To search the content in offline refer to local disk content already download the content from the internet through our browser. Existing Search engine works online and searching content using any words option. Our search engine works offline to search any words matching to our local disk content display the links and all words matching, exact phrase matching to local disk content.



Search Engine Figure 6

7. RESULT

The user give input URL to the browser and retrieve all the information from URL like pdf, txt, doc and also specify the path to store the download information from based on URL. They before accessing of browser enter to the user detail in the registration form. The system is easily to known the crawl history and user transactions. The browser is known about depth of sub link, actions like connect or disconnect, URL name display, bytes are download and then finally known about percentage of download information from URL. The browser is shown in figure 6.

The search engine has three options any words, all words, and phrase. It is used options to find relevant topic to the user input. The proposed search engine has accessed web pages on offline. Normal search engine only work online any problem an internet is not possible to access but proposed system work offline reduce the searching timing is shown in figure 7.

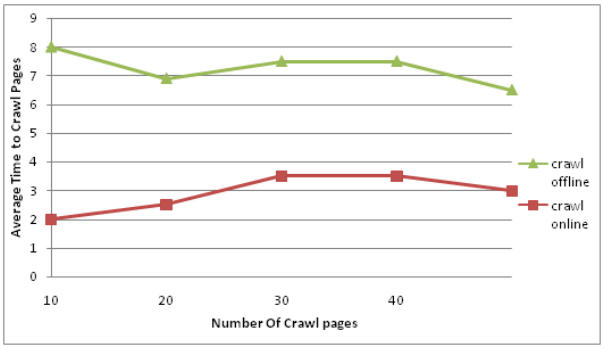


Figure 7: online crawler vs. offline crawler

8. CONCLUSION AND FUTURE WORK

Proposed system using Adaptive focused hyperlink crawler to crawl entire sub link of URL and to store local disk. The local disk do not store more than one time because of apply the relevant score. This information accessed by our search engine it is possible to find phrases, any words, all words. The search engine retrieved information from local disk and it can access offline.

In future, search engine will be enter input is image not a text. The image matches the depth of the pixel, width, height and color matching like eye, skin hair to search from local disk in any files like video or pictures.

9. REFERENCES

[1] Debashis Hati, Amritesh Kumar, Lizashree Mishra, 2010, "Unvisited URL Relevancy Calculation in Focused Crawling Based on Naïve Bayesian Classification", International Journal of Computer Applications (0975 – 8887), Volume 3 – No.9, July 2010.PP: 23-30

[2] Debashis Hati, Amritesh Kumar, 2010, "An Approach for Identifying URLs Based on Division Score and Link Score in Focused Crawler", International Journal of Computer Applications (0975 – 8887),Volume 2 – No.3, May 2010.PP:48-53

[3] M. Sunil Kumar, P.Neelima, 2011, "Design and Implementation of Scalable, Fully Distributed Web Crawler for a Web Search Engine ",International Journal of Computer Applications(0975 – 8887), Volume 15– No.7, February 2011.PP:8-13

[4] Niraj Singhal, Ashutosh Dixit, Dr. A. K. Sharma, 2010, "Design of a Priority Based Frequency Regulated Incremental Crawler",2010 International Journal of Computer Applications (0975 – 8887) Volume 1 – No. 1

[5] Parul Gupta, Dr. A.K.Sharma, 2010 ,"Context based Indexing in Search Engines using Ontology", ©2010 International Journal of Computer Applications (0975 – 8887), Volume 1 – No. 14.PP:49-52

[6] S.Thenmalar, T. V. Geetha, 2011," Concept based Focused Crawling using Ontology", International Journal of Computer Applications (0975 – 8887),Volume 26– No.7, July 2011.PP:29-32

[7] Shekhar Mishra, Anurag Jain,Dr. A.K. Sachan, 2011,"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.PP:8-14.

[8] Swati Mali, B.B. Meshram, 2011, "Focused Web Crawler with Page Change Detection Policy", 2nd International conference and workshop on Emerging Trends in Technology (ICWET) 2011, Proceedings published by International Journal of Computer Applications (IJCA).PP:51-57.

[9] Debashis Hati, Amritesh Kumar, 2010," UDBFC: An Effective Focused Crawling Approach Based On URL Distance Calculation", 978-1-4244-5539-3/10/$26.00 ©2010 IEEE

[10] Swati Mali, B.B. Meshram, 2011," Focused Web Crawler with Page Change Detection Policy", 2nd International Conference and workshop on Emerging Trends in Technology (ICWET) 2011, Proceedings published by International Journal of Computer Applications® (IJCA)