IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# Web Structure Mining Using Breadth First Search with PageRank

[1]Priyanka S. Bandagale, [2] Swati Powar

[1] Assistant Professor, [2] Assistant Professor
Information Technology,
[1]Finolex Academy of Management & Technology, Ratnagiri

*Abstract:* Web Mining is an application of Data mining techniques to extract useful and knowledgeable information from web data. Web data may be web content, web structure and web usage data. Depending upon the type of data to be mined, Web Mining is divided into three types, i.e., Web Content Mining, Web Structure Mining and Web Usage Mining. Web Content Mining is used to extract useful information from the content of the web pages which contains different type of data, for e.g., audio, video, textual, metadata (data about data), image and hyperlinks. Web data may be Structured (in the form of table, tree, list and database), Unstructured (in the form of text-document) and Semi- Structured which do not have a predefined structure (XML and HTML).Web Structure Mining is used to extract structure information from the web and focuses on hyperlink structure. The approach called Breadth First search(BFS) for web structure mining with two algorithms Page Rank and HITS is considered .the implementation of web structure mining using BFS and Page Rank with results is presented

*Index Terms* -Web Structure Mining, Web Extraction, PageRank, HITS, Graphical User Interface.

## I. INTRODUCTION

With the rapid growth of the Web, users get easily lost in the rich hyper structure on the web. Providing relevant information to the users to supply to their needs is the primary goal of the owners of these websites. Web mining is one of the techniques that could help the websites owner in this direction. Web mining was categorized into three categories such as web content mining, web usage mining and web structure mining. Web structure mining plays an important role in this approach. Two page ranking algorithms such as PageRank and Hyperlink-Induced Topic Search (HITS) are commonly used in web structure mining. Both algorithms treat all links equally when distributing rank scores. A comparison between both algorithms was discussed in this paper as well.

## II. Web Mining Catagories

Web Mining consists of three main categories according to the web data used as input in Web Data Mining.
(1) Web Content Mining; (2) Web Usage Mining and (3); Web Structure Mining.

*a. Web Content Mining*

Web content mining is the procedure of retrieving the information from the web into structured forms and indexing the information to retrieve it quickly. It focuses on the structure within a web documents as an inner document level. Table I summarizes the type of concepts of web content mining.

B. Web Usage Mining is used to identify the browsing patterns by analyzing the navigational behavior of user [3]. It focuses to predict the user behavior while the user interacts with the web by using the data from the web. Through this mining technique we can determine what users are looking for on the Internet for the automatic recovery of data. While Web content mining and Web-structure mining utilize real or primary data on the Web. The data from Web server-access logs, proxy-server logs, browser logs, user profiles, registration data, user sessions or transactions, cookies, bookmark data, and any other data that is derived from a person's interaction with the Web.

**C.** Web Structure Mining Web structure mining is defined as the process by which we discover the model of link structure of the web pages. We classify the links; generate the ease of use information such as the similarity and relations among them by taking the advantage of hyperlink topology [2]. PageRank and hyperlink analysis also fall in this class.

## III. WEB STRUCTURE MINING

### A. BRIEF OVERVIEW

The aim of Web Structure Mining is to generate structured abstract about the website and web page. It attempts to discover the link structure of hyperlinks at inter document level. As it is very ordinary that the web documents contain links and they use both the real or primary data on the web so it can be accomplished that Web Structure Mining has a relation with Web Content Mining. It is quite frequently to join these two mining tasks in an application. Table I viewed the type of data that can be joined in web mining application.

| Parameters | |
|---|---|
| View of Data | Link Structures |
| Main Data | Link Structures |
| Representation | Graph,Web page hits |
| Method | Web Page Rank, Propritory algorithm |
| Application Catagory | Clustering, cataogrization |

Table I- Web structure Mining

The World Wide Web (WWW) is trendy and interactive intermediary to telecast in turn these days. It is an enormous, contrary diverse, dynamic and mostly formless data warehouse. As on today WWW is the prevalent information depository for awareness indication.

The subsequent challenges [1] in Web Mining are:

1) Web is enormous.
2) Web pages are partially structured.
3) Web information stands to be miscellany in meaning.
4) Degree of quality of the in sequence extracted.
5) Winding up of knowledge from information extracted.

It is predictable that WWW has lingering by about 2000% since its evolution and is replication in size every six to ten keywords with the catalog proceeds the URLs of the pages to the months [6]. With the swift augmentation of WWW and the user's stipulate on knowledge, it is becoming more difficult to deal with the information on WWW and gratify the user desires. Therefore, the users are in search of improved information repossession techniques and tools to position, extract, and filter and locate the essential information. Most of the users use information reclamation tools akin to search engines to find information from the WWW. There are tens and hundreds of search engines obtainable but some are popular like Google, Yahoo, Bing etc., because of their swarming and ranking methodologies. The search engines download, index and store up hundreds of millions of web pages. They response tens of millions of queries every day. So Web mining and ranking mechanism becomes very significant for effective information retrieval. The sample architecture [4] of a search engine is shown in Figure. 1

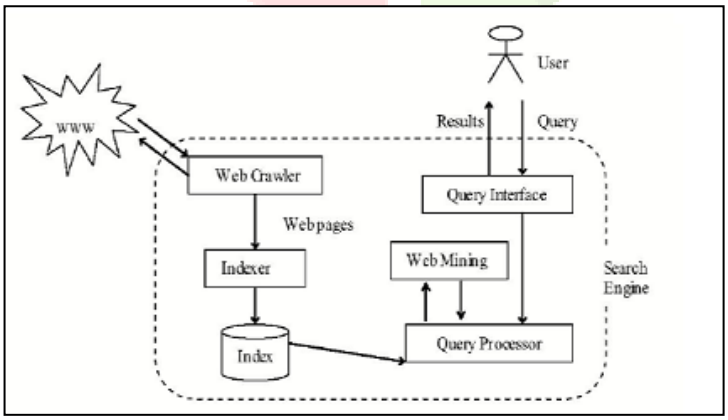


Figure 1: Flowchart for web page ranking on WWW

## IV. WEB PAGE RANKING ALGORITHMS

Web-page ranking is an optimization technique used for search engine, and basic page ranking algorithms can be briefly classified into two class

A. *HITS (HYPER-LINK INDUCED TOPIC SEARCH)*

Klienberg gives two forms of web pages called as hubs and authorities. Hubs are the pages that act as resource lists. Authorities are pages having important contents. A good hub page is a page which is pointing to many authoritative pages on that content and a good authority page is a page which is pointed by many good hub pages on the same content. A page may be a good hub and a good authority at the same time [4,5].

B. PageRank

*C. WEB PAGE RANKING*

PageRank is a link analysis algorithm, named after Larry Page. The Internet Search Engine Google assigns a numerical weighting to each element of a hyperlinked set of web pages. The numerical weight that it assigns to any given element E is also called the Page Rank of E and denoted by PR(E). The page rank within the set measures the relative importance of web pages. The initial PageRank will be evenly assigned to each node. When we starts updating, each page divides its current PageRank evenly across its outgoing links, and new PageRank will be the sum of PageRank of incoming link received. And at one specific value, the network will reach equilibrium where updated PageRank is identical as previous.

**V SYSTEM DESCRIPTION**

A. *JavaScript Object Notation (JSON)*

1. JSON is a syntax for storing and exchanging data. 2. JSON is built on two structures:

- A collection of name/value pairs. In various languages, this is realized as an *object*, record, struct, dictionary, hash table, keyed list, or associative array.
- An ordered list of values. In most languages, this is realized as an *array*, vector, list, or sequence.

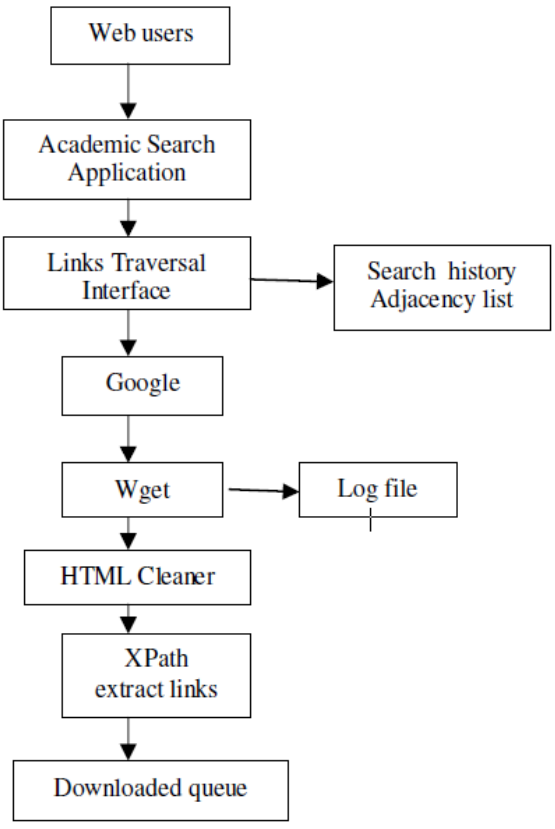


Figure 2 Links_Traversal Flow Diagram

B. The Flow diagram for the proposed system Links_Traversal is shown in Figure 3. The keywords are taken as input from the user through GUI. URL is generated from the search query and files are downloaded using JSON software . The HTML files downloaded are not well formed and not suitable for processing, so they are converted by HTML Cleaner into XML [6]. XPath tools are used to extract all the links and added to the download queue [7].

Each link is chosen by processing the queue and is downloaded. After downloading, the log file is checked to ensure that the file downloaded is of text/HTML MIME [8]. Keyword occurrence count is also done. The steps are repeated till the maximum number of links specified is reached. At each stage, appropriate output is tabulated on the GUI. When the process ends, a log file of the output is saved into a file. Breadth first search strategy has been chosen here..

*C. ALGORITHM: LINKS_TRAVERSAL*

The present work deals with implementation of an algorithm Links_Traversal.

*A. Reading of Inputs:*

Keywords, Number of links to crawl, Number of start links from Google.

*B. Generation of Output:*
Downloaded links in a file.

*C. Procedure:*
1. Enter keyword in GUI.
***Action*:** User enters the keyword in Links_Traversal interface and clicks the search button. A search query is generated and passed to the search engine Google using JSON software.

2. Download the HTMLfile.
***Action*:** JSON software is used to download the file. The downloaded HTML file is converted to XML using HTML Cleaner. HTML found on the web is ill-formed and not suitable for further processing because it contains unclosed tags and missing quotes in the document. HTML Cleaner is an open source HTML parser written in Java. It accepts HTML documents and produces well-formed XML.

3. Extract the links.
***Action*:** XPath is used to extract the links from the search page and each link is added to download queue. XPath is a language for finding information in an XML document.

4. Wget downloads one link.
***Action:*** One link is picked from the queue and downloaded using JSON. The downloaded HTML file is converted to XML.

5. The extracted links are added to the queue.
***Action:*** All the internal links are extracted from the search page using Xpath tools. These links are added to the download queue. Duplicate entries are removed.

6. Calculate the keyword occurrences.
***Action:*** Number of occurrences of keywords is calculated and the user interface is updated to reflect the keyword count for each link.

7. The link is marked as 'crawled'.
***Action:*** If the links are crawled successfully, it is marked as 'crawled', else any one of the following error messages like XML Parse error, Unknown error or Crawl error is displayed.

8. Repeat Steps 4 through 7, till maximum number of links are reached.
***Action:*** Repeat Step 4 to Step 7 till the user defined ' maximum number of links' is reached.

The JSON output is saved onto a log file. The history of the links searched and downloaded can be tracked.

## VI. IMPLEMENTATION

The Links_Traversal algorithm has been implemented using C#asp.net and Ajax Toolkit.

Application also performs spell check and spelling suggestion as shown in figure5 and figure 6:

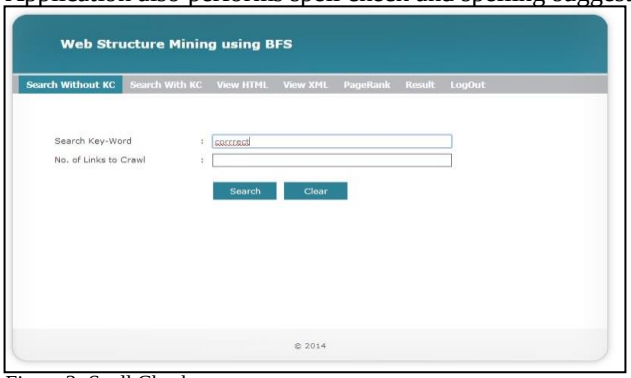


Figure 3: Spell Check

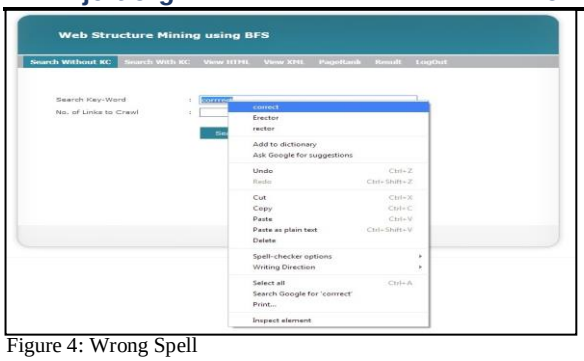


Figure 4: Wrong Spell

The Figure 4 shows the Screen shot of the opening menu of the Links_Traversal module.

The user has to enter all the information like *Keywords*, *Number of links to crawl*, *Number of start links from Google* and click *Search* button. The maximum number of links is limited to 100. The option *Keyword Count* displays the number of occurrences of keyword.The status shows the Page Rank and crawled status.

Output without keyword count is as shown:

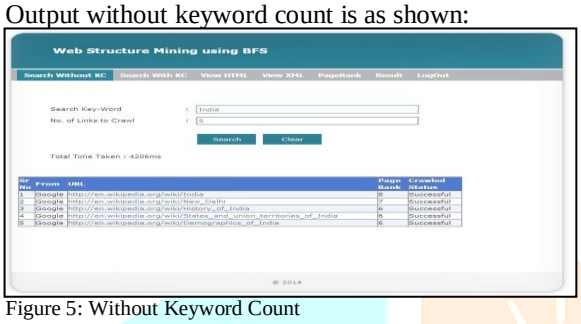


Figure 5: Without Keyword Count

On Clicking the Link in the output it opens in a browser as shown in Figure 6:


Figure 6: Output of Browser

Output with keyword count is as shown in Figure 7

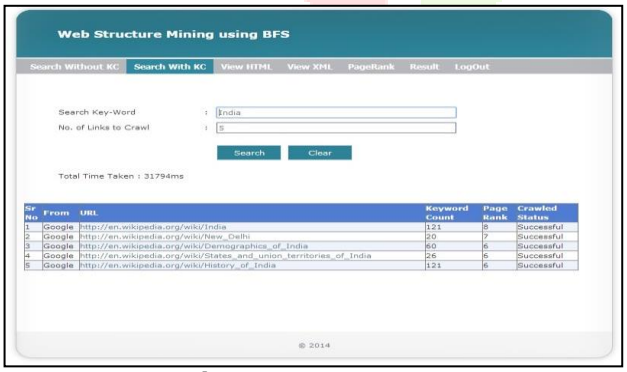


Figure 7: Keyword count

The XML and HTML for the link can be seen as given in Figure 08:

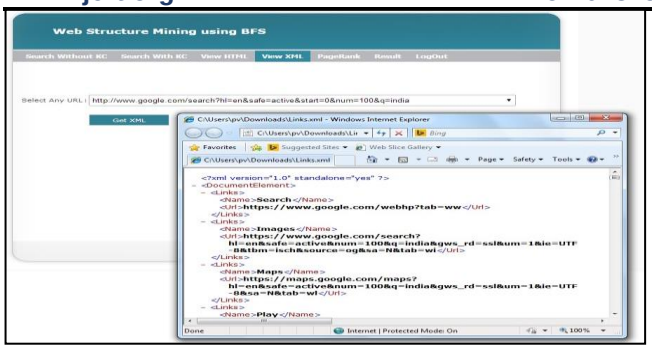


Figure 08: Page structure in XML Format

The result is time required to execute with and without keyword count in Figure 09 and Figure 10.


Figure 09: for Result without Keyword count

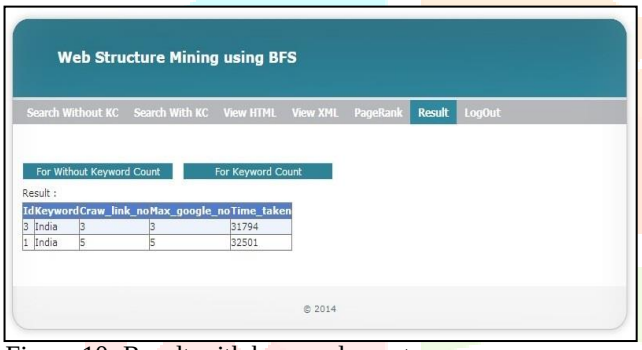


Figure 10: Result with keyword count

Page Rank for the URLS can also be seen by giving input as the URL as shown in figure 11

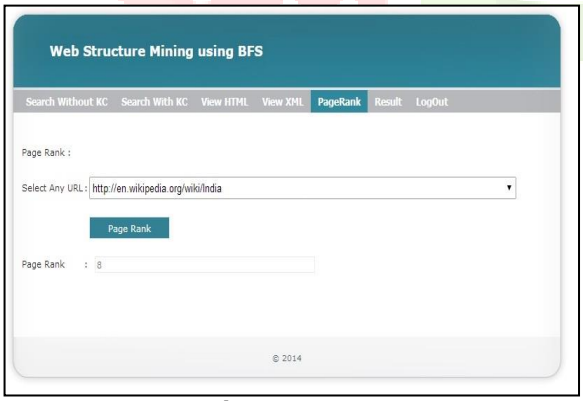


Figure 11: Page Rank

## VII. CONCLUSION

Web Mining is powerful technique used to extract the information from past behavior of users. Various algorithms are used in Web Structure Mining to rank the relevant pages. PageRank, Weighted PageRank and HITS treat all links equally when distributing the rank score. PageRank and Weighted PageRank are used in Web Structure Mining. HITS is used in both structure Mining and Web Content Mining. PageRank. And BFS with PageRank gives more accurate result and takes less time to produce results.

**REFERENCES**


[1]. Search Engine Optimization services and articles, inc. PageRank explained, search engine optimization forum. UK based, worldwide clients, Available: http:// www.webworkshop. net/

[2]. Zakaria Suliman Zubi, Marim Aboajela Emsaed. 2010. Sequence mining in DNA chips data for diagnosing cancer patients. In Proceedings of the 10th WSEAS international conference on Applied computer science (ACS'10), Hamido Fujita and Jun Sasaki (Eds.). World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA, 139-151.

[3]. Zakaria Suliman Zubi. 2009. Using some web content mining techniques for Arabic text classification. In Proceedings of the 8th WSEAS.

[4]. C. Ding, X. He, P. Husbands, H. Zha, and H. Simon, Link analysis: Hubs and authorities on the world. Technical report: 47847, 2001.

[5]. J. M. Klienberg, Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5):604-632, September 1999.

[6]. P Ravi Kumar, and Singh Ashutosh kumar, Web Structure Mining Exploring Hyperlinks and Algorithms for Information Retrieval, American Journal of applied sciences, 7 (6) 840-845 2010.

[7]. 6th International Conference on Internet Technology and Secured Transactions, 11-14 December 2011, Abu Dhabi, United Arab EmiratesDesign and Implementation of a Web Structure

[8]. Mining Algorithm using Breadth First Search Strategy for Academic Search Application S. Jeyalatha, B. Vijayakumar Department of Computer Science BITS Pilani, Dubai Campus Dubai, U.A.E, jeylatha@yahoo.com, bv_uma@yahoo.com