ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 2, February 2019

Development of Web Pattern by Web Structure Mining – A Review

Shailendra Pratap Nirala[1], Abhishek Sharma[2], Brijesh Panday[3]

Pursuing Master of Technology, Dept. of Computer Science and Engineering, GITM, Lucknow, India[1]

Assistant Professor, Dept. of Computer Science and Engineering, GITM, Lucknow, India[2]

Assistant Professor & HOD, Dept. of Computer Science and Engineering, GITM, Lucknow, India[3]

ABSTRACT: The World Wide Web is a very useful and interactive resource of information like hypertext, multimedia etc. When we search any information on the Google, there are many URL's has been opened. The bulk amount of information becomes very difficult for the users to find, extract and filter the relevant information, so that some techniques are used to solve these problems. The objective of current manuscript is focus on processing of structured and unstructured data mining. With the tremendous growth in website, web portal to provide downloaded data to the user.The semantic web is about machine-understandable web pages to make the web more intelligent and able to provide useful services to the users. The data structure definition and recognition is to estimate the accurate page ranking and to produce better result while searching operation with web data**.**

KEYWORDS: Web Structure**,** Weighted PageRank, Topic Sensitive PageRank and TC-PageRank, Hypertext Induced Topic Search.

I. INTRODUCTION

Web structure Mining concentrates on link structure of the web site.The different web pages are linked in some fashion. The potential correlation among web pages makes the web site design efficient. This process assists in discovering and modeling the link structure of the web site. Generally topology of the web site is used for this purpose. The linking of web pages in the Web site is challenge for Web Structure Mining.
The structure of the web page is as shown below.

<html>
link
</html>

The WWW is a collection of various hyperlinked pages. The analysis of these linked pages is of very high importance. In addition to the text contents of a page, the link structure of such pages should be observed while searching for a particular resources. Consider the significance of a link A *B*: With such a link A recommends, that surfers visiting A follow the link and visit B. This may reflect the fact that pages A and B share a common topic of interest, and that the author of A thinks contents of page B. These links are called an informative link[1].There are a number of link structure algorithms. In this paper I have given just an introduction of three algorithms namely PageRank, Weighted PageRank and HITS.

The Page Rank method is used by the Google Web search engine to compute the importance of Web pages. The interpretation of the Page Rank method can be seen in two different view[2] and values

(a) Stochastic or random surfer method in this method the Page Rank values can be viewed as the steady-state distribution of a Markov chain, and

(b) Algebraic method where the Page Rank values taken as the eigenvector corresponding to eigen value of the Link structure matrix. The quality of the search results has been immensely improved by analyzing link structure of web pages[3].The search engine Google uses an iterative algorithm that determines the importance of web pages based on the importance of its parent pages.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 2, February 2019

The logs include information about the referring pages, user identification, time a user spend at a site and the sequence of pages visited. There are number of algorithms proposed based on link analysis.But none of these algorithms talks about the temporal interest shown by a user. As the page becomes older, it earns more links. The analysis of the link structure shows about the user behavior and the amount of time spent by a user on a specific web page. This temporal interest shown by a user which changes according to the time spent by a user affects the importance of a web page. In the present paper, a different interpretation of Page Rank is proposed, namely, a dynamic systems viewpoint, by showing that the PageRank method can be formally interpreted as a particular case of the Interaction information Retrieval method for a particular time.

Table 1.The statistical information collected from Log File

Statistical Information	Parameters which can be measured
History of the website and users	Number of navigators to the website for a given Threshold Number of hits to a particular page Time spent on each page Users visiting a particular web page Average time spent on each page The longest web pages traversal path Association between web pages Redirected or Failed or Successful hits Number of bytes transferred Various Browsers used by clients Usage of GET/POST method
Status Codes which help for troubleshooting	400 Series- Failure Eg:404-File Not Found 300 Series-Redirect 200 Series-Success 500 Series-Server failures
Grade of a Web Page can be measured based on number of hits, time spent, location of web page, in degree and out degree.	Excellent Medium Low

II. LITERATURE REVIEW

A Hyperlink is a structural unit that connects a location in a web page to a different location, either within the same web page or on a different web page. . There are number of algorithms proposed based on ranking. Some algorithms are Hypertext Induced Topic Search (HITS), extension of HITS, PageRank,Weighted PageRank, Topic Sensitive PageRank and TC-PageRank is dicussed below

A. HITS (Hyper-link Induced Topic Search)

A HIT is a purely link-based algorithm. It is used to rank pages that are retrieved from the Web, based on their textual contents to a given query. Once these pages have been assembled, the HITS algorithm ignores textual content and focuses itself on the structure of the Web only.

B. Weighted Page Rank (WPR)

The more popular webpages are the more linkages that other webpages tend to have to them or are linked to by them. The proposed extended PageRank algorithm–a Weighted PageRank Algorithm assigns larger rank values to more important (popular) pages instead of dividing the rank value of a page evenly among its outlink pages. Each outlink page gets a value proportional to its popularity (its number of inlinks and outlinks).

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 2, February 2019

C.Page Rank Algorithm

Page ranking algorithms are the heart of search engine and give result that suites best in user expectation. Need of best quality results are the main reason in innovation of different page ranking algorithms, HITS, PageRank, Weighted PageRank, DistanceRank, DirichletRank Algorithm , Page content ranking are different examples of page ranking used in different scenario. Since GOOGLE search engine has great importance now days and this affect many web users now days, so page rank algorithm used by GOOGLE become very important to researches [4].

D. Page Rank Based on VOL

We have seen that original Page Rank algorithm, the rank score of a page p, is equally divided among its outgoing links or we can say for a page, an inbound links brings rank value from base page, p(rank value of page p divided by number of links on that page)[5]. Which more rank value is assigned to the outgoing links which is most visited by users. In this manner a page rank value is calculate based on visits of inbound links.

E. Result Analysis

This section compares the page rank of web pages using standard Weighted PageRank (WPR), Weighted PageRank using **VOL (*WPR*$_{VOL}$)** and the proposed algorithm. We have calculated rank value of each page based on WPR, ***WPR***$_{VOL}$ and proposed algorithm i.e. ***EWPR***$_{VOL}$for a web graph shown in Table2

Table 2.Comparision Of DifferentAlgorithm.

Algorithm	Page Rank	Weighted Page Rank	Page Rank with VOL	Weighted Page Rank with VOL
Web mining technique used	Web Structure mining	Web Structure mining	Web Structure mining, web usage mining	Web Structure mining,web usage mining
Input parameters	Blacklinks	Blacklinks, Forward links	Backlinks and VOL	Backlinks and VOL
Importance	More	More	More	More
Relevancy	Less	Less	More	More

values of page rank using **WPR**, **WPR**$_{VOL}$ and **EWPR**$_{VOL}$have been compared [6]. The values retrieved by EWPRVOL are better than original **WPR** and **WPR**$_{VOL}$.

Table 3.Shows Comparison of Web Page Ranking Algorithm

System	PageRank	Weighted Page Rank	Page Content Rank	HITS	Link Editing	General Utility Mining	Topological Frequency Utility Mining
Web mining Activity	Web Structure Mining	Web structure Mining	Web content Mining	Web Structure Mining & Web Content Mining	Web structure Mining &Web Usage Mining	Web Usage Mining	Web Structure Mining &Web Usage Mining
Rank Assigned Considering	Pages on the Web	Pages on the Web	Pages on the Web	Pages on the web	Pages in the website	Pages in the website	Pages in the website

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 2, February 2019

Topology	Partial	Partial	Not Considered	Partial	Partial	Not considered	Complete
Process	The page rank for each page is Computed During indexing but not during query time. The relevance of a page plays an important role to sort	The distribution of score is Unequal among its out links. The Computation of scores is done at indexing time.	New grades are computed on the fly for top n pages. Whenever query is Posted relevant web pages will be displayed.	Figures out hub and Authority grades of top n Highly Appropriate pages on the fly. Applicable as well as key pages are returned	The grade for each page is Computed offline. The pages with high in degree and more time spent are Important	The grade of each page is computed considering frequency of each page & the utility of each page.	The grade of each page is Computed Using Frequency, Utility along with Topology Parameters.

The **WPR** uses only web structure mining tocalculate the value of page rank, $\mathbf{WPR_{VOL}}$ uses both web structure mining and web usage mining to calculate value of page rank but it uses popularity only from the number of inlinks not from the number of outlinks.

Table 3. Shows Comparison of Web Page Ranking Algorithm

	the results.			to the users	Pages		
Input Arguments	In links	In links, Out links	Content	In links, Out links And Content	In links, Time spent on Each page, depth of a Page	Frequency and Utility of each page	In links, out Links, Level, Number of pages in the web site
Weighting Factor	Not Considered	Not Considered	Not Considered	Not Considered	Not Consider Ed	Considered only for a specific input	All values Ranging from 0 to 1
Time Complexity	O(log N)	<O(log N)	<O(log N)	O(log N) (higher than WPR)	O(log N)	O(log N)	O(log N)
Result Analysis	Medium	Higher than Page Rank	Approximate or equal to WPR	Less than PR	Equal to PR & less accuracy than TFU	Less Accuracy (than TFU)	High Accuracy

The proposed algorithm $\mathbf{EWPR_{VOL}}$ method uses number of visits of inlinks and outlinks to calculate values of page rank and gives more rank to important pages.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 2, February 2019

III. CONCLUSION AND FUTURE SCOPE

It was concluded that using clustering an algorithm can be designed which is based on partitioning model. The algorithm shows that each partitioned cluster contains average of 80% of all referred pages refer by the pages within that cluster even the percentage remains approximately unchanged from the variations of cluster numbers. The processing time increases linearly as the no of processing node increases.

As the proposed technique is good for clustering, but the time consumed in clustering increases with increase in number of clusters. The reason behind is property of k-means clustering which is sensitive to initial condition. It causes the non uniform cluster information for which clustering is repeated. Hence this problem can be overcome by the use of k-means ++ clustering.

REFERENCES

[1] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine",Computer Networks and ISDN System,30(1-7),,pp. 107-117,1998

[2] Dominich, Sandor and Skrop, Adrienn , "PageRank and Interaction Information Retrieval", JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, 56(1),pp. 63-69, 2005.

[3] Jon Kleinberg, "Authoritative Sources in a Hyperlinked Environment", In Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, 1998.

[4]KaushalKumar, Abhaya and Fungayi Donewell
Mukoko, "PageRank algorithm and its variations: A Survey report", IOSR-JCE., Vol 14, Issue 1, Sep. - Oct. 2013, PP 38-45.

[5] Sonal Tuteja," Enhancement in Weighted PageRank Algorithm Using VOL," in IOSR-JCE, Volume 14, Issue 5 Sep. - Oct. 2013), PP 135-141.

[6]ALLAN BORODIN, GARETH O. ROBERTS , JEFFREY S. ROSENTHAL , and PANAYIOTIS TSAPARAS "Link Analysis Ranking: Algorithms, Theory,and Experiments", ACM Transactions on Internet Technology, Vol. 5, No. 1, February 2005, Pages 231–297.