Volume 8, No. 3, March – April 2017
ISSN No. 0976-5697
International Journal of Advanced Research in Computer Science
REVIEW ARTICLE
Available Online at www.ijarcs.info

A Survey on Web Structure Mining

Kuber Mohan
M.Tech CS
Department of Computer Science
BBAU, Lucknow (U.P.) India

Jitendra Kurmi
Assistant Professor
Department of Computer Science
BBAU, Lucknow (U.P.) India

Sanjay Kumar
M.Tech CS
Department of Computer Science
BBAU, Lucknow (U.P.) India

***Abstract*:** These days, World Wide Web gets to be distinctly gigantic data asset. Research Engine likewise gets to be distinctly famous apparatus that help client find required data rapidly. Due to the enormous number of websites and further more website pages, web crawlers assume an essential part in nowadays. One of the fundamental elements that make the distinction of a web index with other is the ranking mechanism (page rank algorithm). In this paper, we will condense some noticeable page rank algorithm and after that we will exhibit our executions that actualize two page rank algorithm PageRank and Weighted Page Rank Algorithm.

Keywords: PageRank, algorithm, Weighted, mining

I. INTRODUCTION

These days, World Wide Web(WWW) gets to be distinctly colossal data asset. Alongside the advancement of WWW is the improvement of apparatuses that bolster data mining, for example, search instruments like google, yahoo, and so on. As a result of the colossal number of website pages, finding the helpful data is difficult with clients. Web search devices assume a vital part in nowadays.

Fundamentally, the administrator of web crawlers is as follows. Firstly, they read the website pages, remove substance of website pages and discover the connections. By those connections the crawler can take after and handle another website pages. In the wake of gathering the web pages, index module will parse the substance of site pages and fabricate index table in view of the catch phrases utilized as a part of those web pages. When a client fires a pursuit inquiry, web search tools will coordinate the watchwords in the question and in the list table and give back the related website pages. Before give back the outcomes to the client, a ranking mechanism is executed to arrange the outcomes and give the best website page request to the client. These days, there are many web crawlers, the opposition is high. In this way, ranking mechanism of web indexes turns into the fundamental element that decides the accomplishment of web crawlers.
In this paper, we will study some primary ranking technique ,and then we will show a review in that we attempt to execute a few methods with some genuine website page. The arrange of this paper is as per the following. Area II introduce the Web Mining and its arrangement. In segment III, we will outline some conspicuous methods. Area IV talk about the examination between those methods and few dialog. In area V, we will demonstrate our execution of two methods. Segment VI finishes up the paper.

II. WEB MINING

Ranking system is firmly identified with web mining.
Therefore ,we will present Web Mining and its characterization in this area. Web mining is the utilization of data mining procedures to consequently find and concentrate the data of the WWW. Web Mining comprises of the accompanying undertaking [1]:

- Asset finding: the task of recovering planned web records.
- Data choice and pre-handling: naturally choosing and pre-handling particular data from got web assets.
- Speculation: consequently finds general examples at individual websites and also over numerous locales.
- Examination: approval or potentially elucidation of the mined designs.

Asset finding is the way toward recovering the on the web on the other hand disconnected information which is content asset accessible on the web for example, email, substance of HTML archive, and so forth. Data choice and pre-preparing is the change undertaking keeping in mind the end goal to get the primary substance of record. For instance, HTML record is evacuated HTML tag to get the fundamental substance of HTML record. In the third step, machine learning or data mining method are normally used to find the general example of single site or various destinations. Examination approve the mined example and might translate it.

There are three sort of Web mining: Web Content Mining, Web Usage Mining and Web Structure Mining. Web Content Mining is the way toward extricating helpful data
From the substance of web archive. Web Content Mining concern with the recovery of data from substance of WWW. Web Content Mining can be separate from two unique perspectives : Information Retrieval view and Database see. The objective of Web Mining from the Information Retrieval view is to help then again to enhance the finding or separating data. While the objective of Web Mining from Database view is attempting to display information on the web and coordinate them.

Web Usage Mining is the way toward separating valuable data from the auxiliary information got from the cooperation of client while collaborating with the web. The web utilization information incorporates web server get to logs, client profiles, client session, etc.

Web Structure Mining is the way toward creating the basic synopsis about the site and page. The challenge for Web Structure Mining is to concentrate on the hyperlink structure of the Web. At the end of the day, Web Structure Mining is thought to be a procedure by which the model of connect structures and site pages are found [2]. A definitive reason for Web Structure Mining is to create auxiliary rundown about the Web webpage and Web page. This model can be used to arrange website pages or produce helpful data for example, the relationship between the websites. It is utilized to imagine that Web Structure Mining and Link examination are one. With the developing enthusiasm of Web Mining, Web Structure Mining exploration is forming into various systems. Inside itself, Web Structure Mining is arranged into various sub-categorizations as indicated by other researchers[3], [4]. It is prescribed Web Structure Mining to be ordered into two sub-orders: Document Structure Mining and Link Mining. While Link Mining is meant to produce the data of Web pages, for example, the comparability and relationship between various Web destinations, Document Structure Mining opens another course inquire about: uncover the structure (pattern) of Web pages with a specific end goal to think about or incorporate Web page plans [5]. However in the extent of this review, the sub-class Link mining is focused on its procedures and issues.

Ranking mechanism can utilize the procedures of three classifications above. Be that as it may, practically systems identified with the Web Structure Mining to assess the significance of website pages. There are number of calculations proposed to comprehend issues in Link mining point. In the following section, four vital calculations: Pagerank algorithm, Weighted Pagerank algorithm, Weighted content pagerank algorithm (WCPR), Hyperlink-Induced Topic Search (HITS) are talked about and contrasted with make it clear about their techniques and issues.

III. PAGE RANKING ALGORITHM

A. PageRank Algorithm:

Larry Page and Sergey Brin created Pagerank algorithm named PageRank [6]. PageRank accepts that a page that has high rank if the total of the rank of its backlinks is high. It implies that if a page has backlinks from the other high rank pages or has numerous backlinks will have high rank. PageRank algorithm is used by Google search engine. After client ask for a hunt inquiry, Google joins pre-figured static PageRank scores with substance coordinating score to acquires a general ranking score for each website page. The PageRank condition is characterized in 1.

$$R(u) = (1-d) + d \sum_{v \in B(v)} \frac{R(v)}{N(v)} \quad (1)$$

here, u is a website page that we need to ascertain rank score, B(v) is an arrangement of site pages which indicate u, R is PageRank score of a website page. N_v is number of website page which is pointed by website page v. d is a damping element that can be considered as the likelihood of client's taking after direct connections i.e the web chart and is generally set to 0.85.

$$R_A = (1 - d) + d \left(\frac{R_B}{2} \right)$$

$$R_B = (1- d) + d \left(\frac{R_A}{2} + R_C \right) \quad (2)$$

$$R_C = (1- d) + d \left(\frac{R_A}{2} + \frac{R_B}{2} \right)$$

In order to illustrate the working of PageRank algorithm. Let consider the example in the figure 1.

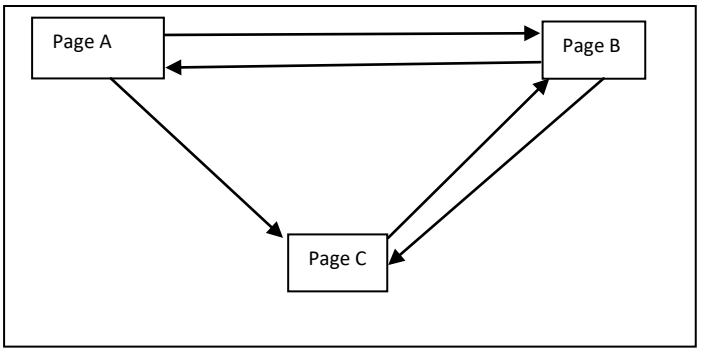

Figure 1. Hyperlink Structure for 3 pages

The Rank Score of each page are calculated as equation system 2. Solve the system of equation with d = 0.82 we can get RA = 0.702,RA = 1.298, RC = 1. However, with the huge number of web pages in reality, solving the equation system is impossible. Another solution for calculating the Rank score is use iterative calculation. In this method, each page is assigned starting rank value of 1 and then Rank score is iteratively calculated by new values. The method is illustrated in Table I

Table I. PAGERANK ITERATION METHOD

Iteration	R_A	R_B	R_C
1.0	1.0	1.0	1.0
0	0.575	1.425	1.0
1	0.756	1.244	1.0
2	0.679	1.321	1.0
3	0.711	1.289	1.0
4	0.698	1.302	1.0
5	0.704	1.296	1.0
6	0.701	1.299	1.0
7	0.702	1.298	1.0

The PageRank algorithm has two primary components. Firstly, PageRank considers 3 figures the rank of website pages that point to the site page, number of it's active connections and number of approaching connection of the website page. Also, the joining time could be expansive.

In [8], the creators proposed another algorithm named to enhance the execution of PageRank algorithm in view of the enhanced standardize procedure. In every cycle, after rank of each page is recalculated. A mean value is computed by isolating summation of rank of all website pages by the number

of website pages. And afterward, the rank of each page is standardized by dividing past rank by mean value. The algorithm is delineated in figure 2.

B. Weigted PageRank Algorithm

Weighted PageRank Algorithm appoints bigger rank qualities to more mainstream pages as opposed to isolating the rank value of a page among its outlink pages. The mainstream web page is the more linkages that other website page have a tendency to need to them or are connected to by them. The rank score is ascertained as condition 3.

$$R(u) = (1-d) + \sum_{v \epsilon B_{(v)}} R(v)\, W^{in}_{(v,u)}\, W^{out}_{(v,u)} \quad (3)$$

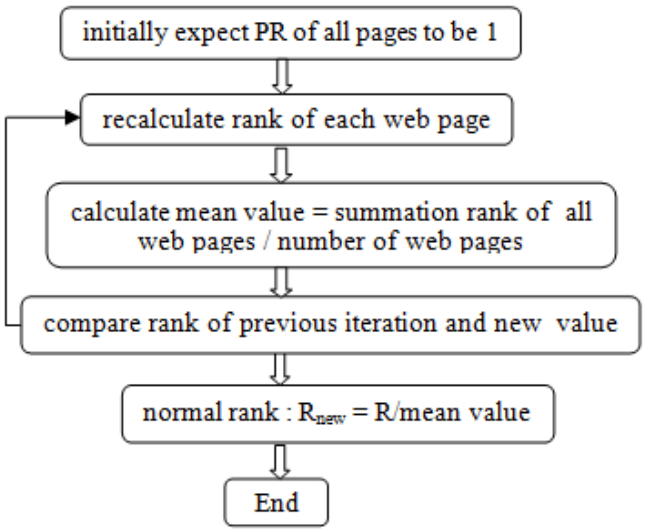

Figure 2. PageRank algorithm based on normalized technique

Win(v,u) is the weight of link(v,u), which is calculated based based on the quantity of inlinks of page u and the quantity of inlink of all reference pages of page v.

$$W^{in}_{(v,u)} = \frac{I_u}{\Sigma_{v \epsilon R}\, f_{(v)}\ I_p} \quad (4)$$

Wout(v,u) is the weight of link(v,u), which is computed in based on the number of outlinks of page u and the number of outlink of all reference pages of page v.

$$W^{out}_{(v,u)} = \frac{O_p}{\Sigma_{v \epsilon R}\, f_{(v)}} \quad (5)$$

Let's illustrate by example in figure 1.

$$W^{in}_{(B,A)} = \frac{I_A}{I_A+I_C} = \frac{1}{3}$$

$$W^{out}_{(B,A)} = \frac{O_A}{O_A+O_C} = \frac{2}{3}$$

Similarly, we have:

$$W^{in}_{(A,B)} = \frac{1}{2} \text{ ;and } W^{out}_{(A,B)} = \frac{2}{3}$$
$$W^{in}_{(B,C)} = \frac{2}{3} \text{ ;and } W^{out}_{(B,C)} = \frac{1}{3}$$
$$W^{in}_{(C,B)} = 1 \text{ ;and } W^{out}_{(C,B)} = 1$$
$$W^{in}_{(C,A)} = \frac{1}{2} \text{ ;and } W^{out}_{(C,A)} = 1$$
$$W^{in}_{(A,C)} = \frac{1}{2} \text{ ;and } W^{out}_{(A,C)} = \frac{1}{3}$$

Solve the system equation we can get RA = 0.234, RB = 0.284, RC = 0.237.

C. Weighted content PageRank algorithm (WCPR)

Although Weighted Page Rank likewise takes the significance of the inlinks and outlinks of the pages, it was understood that the rank score to all links is not similarly appropriated, for instance the unequal circulation is performed. Weighted content pagerank algorithm (WCPG) which in view of web content and structure mining is acquainted all together with demonstrates the pertinence of the pages to a given question so it make clients to be effortlessly get the pertinent and essential pages in the rundown . Weighted Page Content Rank Algorithm (WPCR) is a proposed page ranking algorithm which is utilized to give a sorted request to the web pages returned by a web crawler in response to a client inquiry. WPCR is a numerical value based of which the web pages are given an order. This algorithm utilizes web structure mining and also web content mining techniques. Web structure mining is utilized to figure the significance of the page and web content mining is utilized to discover what amount important a page is? Significance here means the prominence of the page, e.g. what number of pages are indicating or are alluded by this specific page. It can be computed in view of the quantity of inlinks and outlinks of the page. Relevancy implies coordinating of the page with the let go inquiry. In the event that a page is maximally coordinated to the question, that turns out to be more important. The entire of this algorithm can be condensed as the two stages underneath: Input for the algorithm: Page P, inlink and outlink Weights of all backlinks of P, Query Q, d (damping element). Output of the algorithm: Rank score Step :

Step 1: Relevance figuring:
a. Locate all significant word strings of Q (say N)
b. Find whether the N strings are happening in P or not?
c. Z = Sum of frequencies of all N strings.
d. S = Set of the most extreme conceivable strings happening in P.
e. X = Sum of frequencies of strings in S.

f. Content Weight (CW) = X/Z
g. C = No. of question terms in P
h. D = No. of all question terms of Q while overlooking stop words.
i. Probability Weight (PW) = C/D

Step 2: Rank calculation:
a. Discover all backlinks of P (say set B).
b. Compute ranks score as condition 6 .
c. Output PR(P) as the Rank score

$$PR(u) = (1-d) + d \sum_{v \in B_{(*)}} R(v)\, W^{in}_{(v,u)}\, W^{out}_{(v,u)}\, (C_{\omega}+P_{\omega}) \quad (6)$$

D. Hyperlink-Induced Topic Search (HITS)

Hyperlink-Induced Topic Search (HITS) is a link algorithm, presented by Jon Michael Kleinberg . He said webpages : 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. 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 decent expert at the same time.

The HITS algorithm can be compressed two fundamental steps as following: Input with a search topic, indicated by one or more query terms. Step1 -Sampling: A sampling component, which builds an engaged gathering of a few thousand Webpages prone to be rich in important specialists; Step2 - Weight propagation: A weight-propagation component, which determines numerical estimates of hub and authority weights by an iterative procedure.

Outputs of HITS are hubs and authorities for the search. A few constraints of HITS calculation are:

a. Hubs and authorities: It is difficult to recognize amongst hubs and authorities because many websites are hubs as well as authorities.
b. Topic drift: Sometime HITS may not create the most significant archives to the client questions due to proportionate weights.
c. Automatically generated links: HITS gives with significance for consequently produced links which.
d. Productivity: HITS algorithm is not proficient continuously

IV. WEB STRUCTURE MINING ALGORITHMS COMPARISON

See table II.

Table II: VARIOUS WEB PAGERANK ALGORITHMS COMPARISON TABLE

Algorithm	**PageRank**	**WPR**	**WPCR**	**HITS**
Author/Year	S. Brin et al., 1998	Wenpu Xing et al, 2004	P. Sharmar et al., 2000	Jon Kleinberg, 1998
Mining Technique Used	WSM	WSM	WSM and WCM	WSM and WCM
Description	Computes scores at indexing time, not query time. Results are sorted according to importance of pages.	Assigns large value to more important pages instead of diving the rank value of a page evenly among its outlink pages	Gives sorted order to the web pages returned by a search engine as a numerical value in response to a user query	Computes hub and authority scores of n highly relevant pages on the fly. Relevant as well as important pages are returned.
Input / Output Parameters	Backlinks	Backlinks,Forward links	Backlinks,Forwardlinks,Contents	Backlinks,Forwardlinks,Contents
Complexity	O(logn)	¡O(logn)	¡O(logn)	¡O(logn)
Advantages	Providing important pages according to given query.	Providing important pages according to given query. Assigning importance in terms of weight values to incoming and outgoing links.	Providing important pages and relevant pages according to query by using web structure and web content mining	Providing more relevant authority and hub pages according to query
Limitation	Query independent	Query independent	Importance of page is ignored	Topic drift (topic unrelated to the original query)Cannot detect advertisements
Search Engine	Google	Research model	Research model	Clever

V. CASE STUDY: AN IMPLEMENTATION OF PAGERANK AND WEIGHTED PAGERANK ALGORITHM

Keeping in mind the end goal to see unmistakably about PageRank algorithm and Weighted PageRank algorithm, we choose to actualize them and apply it to reality web site. In this segment we will introduce our execution and the outcome.
The website was chosen to analyze is altair.chonnam.ac.kr/ kbkim/". In this segment, the crawler we utilized crawler4j an open source for Java language. The first experiment we compare the result of two algorithm. In this experiment we choose value of d is 0.85 and the threshold is 0.001. Table III shows 10 pages in which initial 5 pages is top 5 page of page rank list that is obtained by PageRank algorithm and last 5 pages is top 5 pages acquired by Weighted PageRank algorithm. Figure 3 demonstrates the connection between consequence of PageRank and Weighted PageRank. The order of x axis follows the rank order of PageRank algorithm. We can observe that, The difference between two algorithm results is very small in the left area. However, there is big difference between two results in the right area. It is because the difference of considering important page. PageRank considers three factors inlinks, source pages of inlink, outlink of source page. But Weighted Page Rank considers inlinks, source pages of inlinks, outlinks, relation between outlinks of source pages.

Table III. TOP TEN OF HIGH RANK PAGES

URL	PR		WPR	
	score	order	score	order
...3/overview-summary.html	25.942	1	2.651	4
...2/overview-summary.html	25.91	2	2.303	6
...3/deprecated-list.html	14.846	3	0.234	41
...3/help-doc.html	14.846	4	0.246	37
..3/index-files/index-1.html	14.846	5	0.725	19
...2/allclasses-frame.html	12.789	12	4.336	1
...3/allclasses-frame.html	12.308	14	3.898	2
../package-summary.html	9.841	17	2.746	3
3/overview-summary.html	8.921	18	2.59	5
...game/package-summary.html	7.771	20	2.253	7

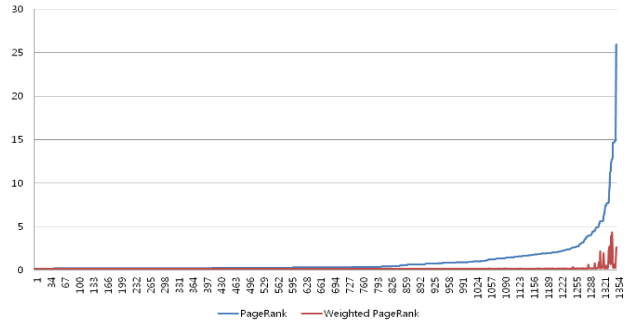

Figure 3. Relation Between PageRank Result and Weighted PageRank Result

With a specific end goal to assess the execution of pagerank and weighted pagerank, we run two algorithms with different values of threshold and then we calculate the convergence time and number of round. Figure 4 show the comparison of convergence time and iteration number of two algorithms. From the figure, we can watch that, the execution of Weighted Page Rank is superior to PageRank. The number of iteration and the convergence time of Weighted PageRank is smaller than PageRank's.

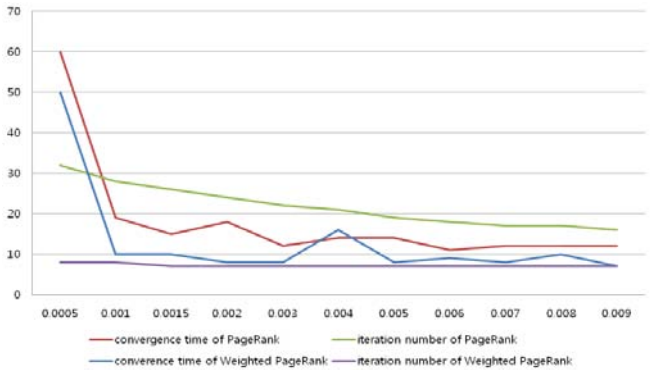

Figure 4. Comparison of convergence time and iteration number between two algorithms

In the following experiment, we run two algorithms with different values of d parameter and then we calculate the convergence time and number of round. Figure 5 show the comparison of convergence time and iteration number of two algorithms. A similar outcome with the past analysis. The execution of Weighted Page Rank is superior to pagerank, and we can observe that the number of iteration get smaller when d parameter is smaller.

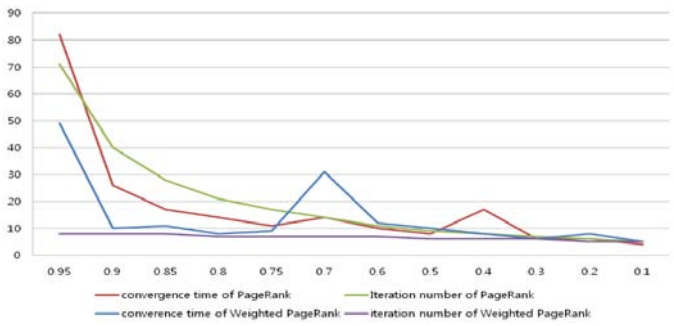

Figure 5. Comparison of convergence time and iteration number between two algorithms

VI. CONCLUSION

In this review, Web Structure Mining was inspected in the relationship to Web Mining. Additionally, the two sub-classifications of Web Structure Mining were stated through writing analysis work. The four critical algorithms in Web Structure Mining were inspected and abridged consolidated to make it effortlessly comprehend the idea of those algorithms. The most critical some portion of this review is the tabular comparison content about four vital techniques utilized as a part of Web Structure Mining. The comparison substance was inspected precisely through writing analysis work, and afterward it was tried by contrasting the outcome picked up from an execution by and by some of those techniques. Be that as it may, this review still has some limitation. We did not have enough time to execute those methods to test the correlation content experimentally. One thing we still missed in this review when despite everything we couldn't contribute any oddity to those techniques.

REFERENCES

[1] R. Kosala and H. Blockeel, "Web mining research: A survey," SIGKDD Explore. Newsl., vol. 2, no. 1, pp. 1–15, Jun. 2000. [Online]. Available: http://doi.acm.org/10.1145/360402.360406.

[2] T. Bhatia, "Link analysis algorithms for web mining," IJCST, vol. 2,no. 2 2 2 9 - 4 3 3 3, pp. 243–246, jun 2011.

[3] WebStructure mining: an introduction, 2005. [Online]. Available: http://dx.doi.org/10.1109/icia.2005.1635156

[4] M. A. Preeti Chopra, "A survey on improving the efficiency of different web structure mining algorithms," International Journal of Engineering and Advanced Technology (IJEAT), vol. 2, no. 2249 8958, pp. 296–298,feb 2013.

[5] S. K. Madria, S. S. Bhowmick, W. K. Ng, and E.-P. Lim, "Research issues in web data mining." in DataWarehousing and Knowledge Discovery, First International Conference, DaWaK '99,1999,pp.303–312.[Online]. Available:http://www.springerlink.com/content/yar775kx05pgnj93

[6] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," in Proceedings of the 7th International World Wide Web Conference, Brisbane, Australia, 1998, pp. 161–172. [Online]. Available:citeseer.nj.nec.com/page98pagerank.html

[7] "Weighted pagerank algorithm," in Proceedings of the Second Annual Conference on Communication Networks and Services Research, ser. CNSR '04. Washington, DC, USA: IEEE Computer Society, 2004, pp. 305–314. [Online]. Available: http://dl.acm.org/citation.cfm?id=998669.998911

[8] H. Dubey and P. B. N. Roy, "An improved page rank algorithm based on optimized normalization technique," pp. 2183–2188, 2011.