ISSN:2320-0790

An International Journal of Advanced Computer Technology

A Survey on Various Web Page Ranking Algorithms

Renu Gupta[1], Ankita Shah[2], Amit Thakkar[3], Kamlesh Makvana[4]

[1]M.E. Student, [2]Assistant Professor, [3]Associate Professor, [4]Assistant Professor
[1,2]Computer Engineering Dept., NSIT, Jetalpur, [3, 4]Information Technology Dept., CHARUSAT, Changa

Abstract: World is full of information and searching is most common task on web. As the amount of information available on web is increasing, it is difficult to acquire relevant information on web. User enters a query for retrieving required information from www and millions of web pages are fetched. These web pages or search results contain both relevant pages and irrelevant search results in response to query submitted by user. For this issue efficient Page Ranking algorithm is needed. Google uses very basic algorithm called Page Rank algorithm which uses web structure mining and has some limitations. In this survey paper we analyzed various improvements of Page Rank which uses web content mining for efficient ranking. Relative strengths and limitations of some algorithms are explored to find out further scope of research.

Keywords: Page Ranking Algorithm, Inlinks, Outlinks, Visit count, Weighted PageRank (WPR), Ratio rank, Weighted page content rank(WCPR), Topic Character, Time Factor.

I. INTRODUCTION

With the rapid development of web, Internet has become the world's richest source of information. Searching in World Wide Web (WWW) will give a large set of results for a single query, which can be up to millions in number containing both relevant and irrelevant results. The main issue on web is that how to get relevant and useful information from the large number of disorder information.

For efficient search results as according to user's query, many ranking algorithms are used which calculate Page Rank of web pages and the goal of the Page Ranking is to make the user get the desired result at the top of the list.

These ranking algorithms are either based on web structure mining or web content mining. For calculating the Page Rank value, web structure mining uses only link structure of web pages it does not take user query in account and web content mining calculate the rank value according to user query and does not use link structure for calculating rank value of web pages.

II. WEB MINING

Web Mining is used to extract useful information from web data. All the data mining technique which is applied on web than it is called web mining. Web mining is very helpful to find out relative information, to established relationship between no. of page which has a same concept, to identify user behaviour etc.
Web Mining can be classified into three categories:

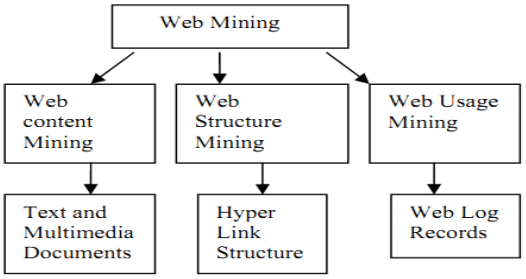

Figure 1: Web Mining Categories

1) Web Content Mining:

It is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. Web content mining is related but is different from data mining and text mining. It is related to data mining because many data mining techniques can be applied in web content mining [8].

It is related to text mining because much of web contents are text based. It is different from data mining because web data are mainly semi-structured and or unstructured. Web content mining is also different from text mining because of the semi-

structure nature of the web, while text mining focuses on unstructured texts [8]. The technologies that are normally used in web content mining are NLP (Natural language processing) and IR (Information retrieval).

2) Web Structure Mining:

We can define web structure mining in terms of graph. The web pages are representing as nodes and Hyperlinks represent as edges. Basically it's shown the relationship between user & web. The motive of web structure mining is generating structured summaries about information on web pages/webs [11]. It is shown the link one web page to another web page.

3) Web Usage mining :

Web usage mining is the application of data mining techniques to discover usage patterns from Web data in order to understand and better serve needs of Web based applications. It consists of three phases, namely pre-processing, pattern discovery, and pattern analysis. Web servers, proxies, and client applications can quite easily capture data about Web usage [8]. However, one of the major challenges faced by Web usage mining applications is that Web server log data are anonymous, making it difficult to identify users and user sessions from the data.

III. LITERATURE SURVEY ON VARIOUS PAGE RANKING ALGORITHMS

A. Page Rank Algorithm

Page Rank algorithm [1] is the very basic and first algorithm developed by Larry Page one of the founders of google and Sergery Brin which uses link structure i.e. web structure mining for ranking purpose. In Page Rank algorithm inlinks are used to count the rank of web pages. If more number of hyperlinks pointing to the page, the more important the page and hence it's Page Rank Value is high. A web page is important if it is pointed to by other important web pages.

Formula for Page Rank calculation:

$$PR(A) = 1 - d + d\left(\frac{PR(T_1)}{C(T_1)} + \frac{PR(T_2)}{C(T_2)} + \cdots \frac{PR(T_n)}{C(T_n)}\right)$$

Where, *PR(A)* is Page Rank of page *A*, *n* is total pages accounted, d is dampening factor which is generally set to 0. 85 and used to give some Page Rank value to those pages which has no inlinks. $C(T_1)$, $C(T_2)$, … , $C(T_n)$ are number of outlinks of pages T_1 , T_2 , … , T_n which links to page A respectively.

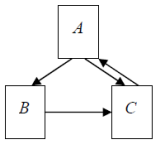

Figure 2: Example of Hyperlinked Structure

Example illustrating working of Page Rank:

Suppose that a small website consists of three pages as page A, page B and page C. Page A links to page B and page C. page B links to page C and page C links to page A. The value of dampening factor is usually taken as 0.85. We can calculate the Page Rank value by following equation.

PR(A) = 0.15 + 0.85 PR(C)
PR(B) = 0.15 + 0.85 (PR(A)/2)
PR(C) = 0.15 + 0.85 (PR(A)/2 + PR(B))

Iterative Calculation of above example is as follows:

Table 1: Iterative calculation of each page

Iteration	PR(A)	PR(B)	PR(C)
0	1	1	1
1	1.00	0.58	1.06
2	1.05	0.60	1.10
3	1.09	0.61	1.13
4	1.11	0.62	1.15
5	1.13	0.63	1.17
6	1.14	0.63	1.17
7	1.14	0.63	1.17

Page Rank value of page A, B and C is:
PR(A) = 1.14
PR(B) = 0.63
PR(C) = 1.17
It may be noted that in this example, PR(c) > PR(A) > PR(B).
As Page Rank algorithm uses link structure to counting the rank value of the webpage some results produced by Page Rank algorithm are not relevant to user's query this problem is called theme drift.

B. Hypertext Induced Topic Search

HITS (Hypertext Induced Topic Search) Algorithm was introduced by Kleinberg [2] which uses web structure mining. HITS ranks the webpage on the basis of inlinks, outlinks of any webpages. HITS takes web as a web graph and divide the webpages into two types' hubs and authorities. Hubs are pages which points towards the data related to the query and Authorities are pages which are related to user's query. Figure 3 depicts the hubs and authorities created by HITS.

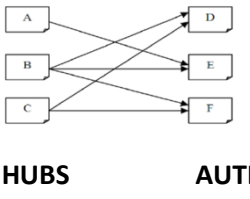

Figure 3: Hubs and Authorities

HITS algorithm considers the entire World Wide Web as a graph (V, E) where V are the web pages and E are the hyperlinks. It has two steps: Sampling step and Iterative step.

***Sampling Step**:* The relevant pages are gathered based upon the given query.

***Iterative Step**:* The relevant pages that are found in the basis step are used in this iterative step. Hubs and Authorities are determined in these pages. This is done iteratively such that the Hub weight and the Authority weight results converges. The Hub weight of a page p is given by *Hp*. The Authority weight of a page p is given by *Ap*. Initially the values of *Hp* and *Ap* are set to be 1.Then its values are calculated based upon the iterative formula [12].

$$H_p = \sum_{q:p\to q} A_q$$

$$A_p = \sum_{q:p\to q} H_q$$

The Hub weight is the sum of all authority weights that is pointed by the Hub. And the Authority weight is the sum of all hub weights that points to the authority. Although HITS provide good results, it has certain drawbacks [12]. Hubs and Authorities show mutual relationship with each other and weight of each depends on the other. And sometimes HITS can contain pages that are not relevant to a query. And also HITS algorithm is not very efficient for real time usage.

HITS give same importance to all pages this is major limitation of HITS and problem of theme drift still exist in HITS.

C. Weighted Page Rank Algorithm

Weighted Page Rank Algorithm [3] is an extension (enhanced version) of Page Rank Algorithm in which the rank value of the web pages is calculated using weight of inlinks and weight of outlinks. WPR gives better results as compared to Page Rank Algorithm according to user's query.

The weight of inlinks and outlinks of web pages are calculated as following equations [3]:

$$W^{in}_{(v,u)} = \frac{I_u}{\sum_{p\in R(v)} I_p}$$

Here, Iu and Ip are the number of inlinks of webpage u and webpage p and R (v) is the reference page list of v.

$$W^{out}_{(v,u)} = \frac{O_u}{\sum_{p\in R(v)} O_p}$$

Where, Ou and Op are the outlinks of page u and p.

Modified Equation of Weighted Page Rank is as follows [3]:

$$PR(u) = (1-d) + d\sum_{v\in B(u)} PR(v)\, W^{out}_{(v,u)} W^{in}_{(v,u)}$$

Example illustrating working of Weighted Page Rank:

To illustrate the working of WPR refer again to Figure 2.The value of dampening factor is usually taken as 0.85. We can calculate the Weighted Page Rank value by following equation.

$$WPR(A) = 0.15 + 0.85(WPR(C)\cdot W^{in}_{(C,A)}\cdot W^{out}_{(C,A)})$$
$$WPR(B) = 0.15 + 0.85(WPR(A)\cdot W^{in}_{(A,B)}\cdot W^{out}_{(A,B)})$$
$$WPR(C) = 0.15 + 0.85(WPR(A)\cdot W^{in}_{(A,C)}\cdot W^{out}_{(A,C)} + WPR(B)\cdot W^{in}_{(B,C)}\cdot W^{out}_{(B,C)})$$

The weights of incoming as well as outgoing links can be calculated as follows:

$W^{in}_{(C,A)} = I_A = 1$
$W^{out}_{(C,A)} = O_A = 2$
$W^{in}_{(A,B)} = \frac{I_B}{(I_{B+I_C})} = 1/(1+2) = 1/3$
$W^{out}_{(A,B)} = \frac{O_B}{(O_{B+O_C})} = 1/(1+1) = 1/2$
$W^{in}_{(A,C)} = \frac{I_C}{(I_{B+I_C})} = 2/(1+2) = 2/3$
$W^{out}_{(A,C)} = \frac{O_C}{(O_{B+O_C})} = 1/(1+1) = 1/2$
$W^{in}_{(B,C)} = I_C = 2$
$W^{out}_{(B,C)} = O_C = 1$

After substituting above calculated weight values in WPR equation, page ranks of A, B and C become:

WPR(A) = 1.85
WPR(B) = 0.41
WPR(C) = 1.37

Here WPR(A) > WPR(C) > WPR(B). It shows that the resulting order of pages obtained by original PageRank and WPR is different.The problem of theme drift is exist in WPR.

D. An Improved Method for the computation of PageRank

An Improved method for PageRank was introduced by Huang Wei, and Bin Li [10]. Traditional Page Rank algorithm has some limitations such as more emphasis on old pages, an absolute average of PR distribution and topic bias (theme drift). Hence Wei and Li proposed an algorithm which removes limitations of Page Rank algorithm. The improved algorithm taking topic character and time factor into account. The proposed formula for improved Page Rank is as follows [10]:

$$[PR(u) = dT_u\left[\sum_{v\in B(u)} PR(v)\left(\frac{m}{N_v} + (1-m)W_v\right) + (1-$$

d)

Where, *Wv* denotes the topic weight between page *u* and page *v*. *Tu* denotes the time factor of page *u*. *Nv* is the number of elements in the intersection of the page u topics set and page *v* topics set. *m* is split factor, for distributing a

certain proportion of the page's Page Rank value to those pages that is completely irrelevant with linked page**.**

After implementation authors conclude that improved computation of Page Rank shows more excellent search performance compared to the classic Page Rank algorithm [10].

E. PageRanking Based on Number of Visits of Links of Web Page (VOL)

Kumar Gyanendra, Neelam Duhan, and A. K. Sharma [4] analyzed various link based ranking algorithms. In this analysis, they focus on traditional Page Rank algorithm and extension of Page Rank algorithm called Weighted Page Rank algorithm (WPR). Authors conclude that basic Page Rank algorithm ranks the pages on the basis of inbound links. WPR considers both inlinks weight and outlinks weight. Authors presented a modified PageRanking algorithms which is more target oriented than original PageRank. This modified algorithm calculates Page Rank value or importance of web pages based on the visits of incoming links on a page. It is not only considering link structure but also focuses on a particular page. In this algorithm, they assign more rank value 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. The modified version based on VOL is given in equation [4]:

$$PR(u) = (1\text{-}d) + d \sum_{v \in B(u)} \frac{L_u(PR(v))}{TL(v)}$$

Where, L_u is the number of visits of link which points from v to u, $TL(v)$ is the total number visit of all links present on v, B(u) are the pages which points to webpage u. d is damping factor usually set as 0.85.

VOL [4] calculates rank value of a web page based on the user visits on incoming links of that page. The ordering of pages in this way increases the relevancy of pages and thereof provides the user with quality search results. As a result, user may find the desired content in the top few pages, thus search space can be reduced to a large scale.

F. RATIORANK:Enhancing The Impact Of Inlinks And Outlinks

Singh Ranveer and Dilip Kumar Sharma [6] proposed new algorithm called Ratiorank. In this paper various ranking algorithms are analyzed and explained in which different users use different parameters for ranking algorithm. Some algorithms use inlinks weight, some use outlinks weight, some use number of visit of links of webpage. Some use both inlinks weight as well as outlinks weight. Then the new algorithm is presented which is called RatioRank, in which the inlinks weight and outlinks weight are used with the consideration of number of visit count and is compared with some algorithms by using certain parameters.

The equation for the Ratiorank is as follows [6]:

$$RR(u) = (1-d) + d \sum_{v \in B(u)} \frac{\left(V_u * x * W^{in}_{(v,u)} + y * W^{out}_{(v,u)}\right) RR(v)}{TL(v)}$$

Where, *RR(u)* and *RR(v)* are ranking of the webpages u and v respectively $W^{in}_{(v,u)}$ and $W^{out}_{(v,u)}$are used to record the popularity of the inlinks and the outlinks and calculated as per Weighted Page Ranking Algorithm[3]. d is the dampening factor. V_u is the number of visits of link which points from v to u. *TL(v)*is the total number visit of all links present on v, *B(u)* are the pages which points to webpage u, x is the ratio of inlink weight and y is the ratio of outlink weight. The values of **x** and **y** will be set between 0 to 1 on the basis of empirical results, while the ratio of inlink weights will be higher than that to the ratio of outlink weights because inlinks are considered to be more important to rank the webpages.
This algorithm gives better results in terms of relevance of webpages, because no other then the RatioRank Page Ranking algorithm uses all the features together as the inlinks and outlinks of any webpage and the visit of link count [6].

G. Enhanced-Ratiorank:Enhancing The Impact Of Inlinks And Outlinks

Again Singh Ranveer and Dilip Kumar Sharma [7] proposed Enhanced-RatioRank algorithm which considers ratio of weight of the inlinks and weight of outlinks for calculation. They check which ratio gives the best result which ratio helps to give better relevancy of the web pages and proposed new equation called Enhanced RatioRank is given as follows [7]:

$$RR(u) = (1-d) + d \sum_{v \in B(u)} \frac{\left(V_u * .7 * W^{in}_{(v,u)} + .3 * W^{out}_{(v,u)}\right) RR(v)}{TL(v)}$$

The parameters of above equation are same as RatioRank equation. As in equation only 70 percent of the weight of inlinks and the 30 percent of the weight of the outlinks is being used. Compare to other this ratio gives the better result.

By using all three parameters for computing the Page Rank and taking particular ratio of weight of inlinks and outlinks gives the better relevancy of web pages. But the problem of theme drift still exists in this algorithm [7].

H. Weighted Page Content Rank for Ordering Web Search Result utlinks

WCPR was introduced by Sharma Pooja and Pawan Bhadana [8] which uses both web structure mining (WSM) and web content mining (WCM) for efficient ranking of web pages. In this paper basic search engine architecture is explained. After that basic Page Rank algorithm and weighted Page Rank algorithm are analyzed and by this analysis they came to know that both Page Rank and weighted Page Rank algorithm have some limitations. In

Page Rank Algorithm's search results, some links are not related to the user query because Page Rank is equally distributed to outgoing links and it is purely based on the number of inlinks and outlinks. In Weighted Page Rank algorithm provides important information about a given query by using the structure of the web. While some pages may be irrelevant to a given query, it still receives the highest rank because it has many inlinks and many outlinks. There is a less determination of the relevancy of the pages to a given query. To improve these limitations improved algorithm called Weighted Page Content Rank (WPCR) which uses both web structure mining as well as web content mining for provide efficient search results. The proposed algorithm is as follows [8]:

Step 1: Relevance calculation:

a) Find all meaningful word strings of Q (say N)
b) Find whether the N strings are occurring in page P or not?
Z= Sum of frequencies of all N strings.
c) S= Set of the maximum possible strings occurring in P.
d) X= Sum of frequencies of strings in S.
e) Content Weight (CW) = X/Z
f) C= No. of query terms in P
g) D= No. of all query terms of Q while ignoring stop words. h) Probability Weight (PW) = C/D

Step 2: Rank calculation:

a) Find all backlinks of P (say set B).

b)$PR(P) = (1-d) + d[\sum_{v\in B}^{n} \mathrm{PR(V)}\, W_{(p,v)}^{in} W_{(p,v)}^{out}](CW + PW)$

c) Output *PR(P)* i.e. the Rank score.

In above equation, two parameters are used which are necessary to understand.

Probability Weight: It is the probability of the query terms in the web page. This factor is the ratio of the query terms present in the document and the total number of terms in the fired query.

Content Weight: It is the weight of content of the web page with respect to query terms. This factor is the ratio of the sum of frequencies of highest possible query strings in order and sum of frequencies of all query strings in order. The maximum possible strings are selected in such a way that all such strings represent a different logical combination of words.

WCPR algorithm [8] employs Web structure mining (WSM) as well as Web content mining (WCM) techniques. This algorithm is aimed at improving the order of the pages in the result list so that the user may get the relevant and important pages easily in the list.

I. An Effective Content Based Web Page Ranking Approach

Shalya Nidhi, Shashwat Shukla, and Deepak Arora [9] gave a new ranking method which uses web content mining (WCM). In this paper, traditional Page Rank algorithm is analyzed and limitation of it is considered for improving. The limitation of the Page Rank algorithm is that the rank score of a web page is divided evenly over its outlinked pages and pages that are not relevant to the user query may get the higher rank. To overcome this problem, they proposed a new query dependent algorithm which is based on the web structure mining and web content mining.
The content based Page Rank algorithm is given as [9]:

1) Initially, let PAGE RANK of all web pages to be 1.
2) Calculate PageRanks of all pages by following formula:

$$\mathrm{PR(u)} = (1-\mathrm{d}) + \mathrm{d}\sum_{v\in B}^{n} \mathrm{PR(V)}\,.\,\mathrm{WL(v,u)}.\mathrm{Wc}$$

Where, *PR(u)* and *PR(v)* are the Page Rank scores of page *u* and *v* respectively, *B(u)* is the set of pages that point to u, *Wc* is the content weight [9] of the web pages with respect to the query terms.

3) Repeat step 2 until values of two consecutive iterations match.

By implementation of proposed algorithm authors conclude that proposed algorithm provide efficient search results then traditional Page Rank algorithm.

IV. COMPARITIVE ANALYSIS OF VARIOUS WEB PAGE RANKING ALGORITHMS

Various Web Page Ranking algorithms are summarized in Table 2 with their strength and limitations and technique used to rank the retrieved webpages.

Table 2: Comparison of Various Web Page Ranking Algorithms

Algorithm	Mining Technique	Input Parameter	Advantage	Limitation
Page Rank Algorithm	Web Structure Mining	Inlinks	Ranking is done at indexing time not at query time	Theme Drift
HITS	Web Structure Mining and Web Content Mining	Inlinks, Outlinks, Content	Relevancy of the pages is high	Theme Drift
Weighted Page Rank Algorithm	Web Structure Mining	Inlinks and Outlinks Weight	Higher Relevancy than traditional Page Rank Algorithm	Theme Drift
Page Ranking Based on Visits of Links of Pages	Web Structure Mining and Web Usage Mining	Inlinks , Outlinks, Visit Count of Links	User Input is considered hence Relevancy of Pages is Higher	Theme Drift
An improved Method for Computation of PageRank	Web Structure Mining	Inlinks, Topic Character, Time Factor	Reduces Computational Complexity	Theme Drift
Ratio Rank And Enhanced-RatioRank	Web Structure Mining	Inlinks, Outlinks, Visit Of Links From User	Avoid Similar Ranking	Theme Drift

V. CONCLUSION

In this survey paper, we have analyzed various Page Ranking algorithms such as Original PageRank algorithm HITS, WPR, VOL. These algorithms have limitation of theme drift (some pages not give result related to user's query). We then analyzed the improved Page Rank algorithms. After Studying various link based ranking algorithms, we understood that these algorithms use only link structure i.e. web structure mining so problem of theme drift is exist in all these algorithms. We have analyzed various content based algorithms which use web content mining. Some content matching method is needed to reduce this problem. So by proper analysis of both WSM and WCM, we conclude that by using web structure mining parameters and web content mining parameter i.e. Content Weight in calculation of Page Rank can improve the relevancy and reduce the problem of theme drift.

VI. REFERENCES

[1]. Brin, Sergey, and Lawrence Page. "Reprint of: The anatomy of a large-scale hypertextual web search engine." Computer networks 56.18 (2012): 3825-3833.

[2]. Kleinberg, Jon M. "Authoritative sources in a hyperlinked environment."*Journal of the ACM (JACM)* 46.5 (1999): 604-632.

[3]. Xing, Wenpu, and Ali Ghorbani. "Weighted Page Rank algorithm."Communication Networks and Services Research, 2004. Proceedings. Second Annual Conference on. IEEE, 2004.

[4]. Kumar, Gyanendra, Neelam Duhan, and A. K. Sharma. "Page Ranking based on number of visits of links of Web page." Computer and Communication Technology (ICCCT), 2011 2nd International Conference on. IEEE, 2011.

[5]. Tyagi, Neelam, and Simple Sharma. "Weighted Page Rank algorithm based on number of visits of Links of web page." *International Journal of Soft Computing and Engineering (IJSCE) ISSN* (2012): 2231-2307.

[6]. Singh, Rajdeep, and Dilip Kumar Sharma. "RatioRank: Enhancing the impact of inlinks and outlinks." Advance Computing Conference (IACC), 2013 IEEE 3rd International. IEEE, 2013.

[7]. Singh, Rajdeep, and Dilip Kumar Sharma. "Enhanced-RatioRank: Enhancing impact of inlinks and outlinks." *Information & Communication Technologies (ICT), 2013 IEEE Conference on*. IEEE, 2013.

[8]. Sharma, Pooja, and Pawan Bhadana. "Weighted page content rank for ordering web search result." International Journal of Engineering Science and Technology 2.12 (2010): 7301-7310.

[9]. Shalya, Nidhi, Shashwat Shukla, and Deepak Arora. "An Effective Content Based Web Page Ranking Approach." *International Journal of Engineering Science and Technology (IJEST)* 4.08 (2012).

[10]. Huang, Wei, and Bin Li. "An improved method for the computation of PageRank." Mechatronic

Science, Electric Engineering and Computer (MEC), 2011 International Conference on. IEEE, 2011.

[11]. Sharma, Kavita, Gulshan Shrivastava, and Vikas Kumar. "Web mining: Today and tomorrow." *Electronics Computer Technology (ICECT), 2011 3rd International Conference on*. Vol. 1. IEEE, 2011.

[12]. Sangeetha, M., and K. Suresh Joseph. "Page Ranking algorithms used in Web Mining." *Information Communication and Embedded Systems (ICICES), 2014 International Conference on*. IEEE, 2014.

[13]. Devi, Pooja, Ashlesha Gupta, and Ashutosh Dixit. "Comparative Study of HITS and Page Rank Link based Ranking Algorithms." *International Journal of Advanced Research in Computer and Communication Engineering* 3.2 (2014).