Scientific Journal of Impact Factor (SJIF): 4.14
e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 3, Issue 2, February -2016

A Review Paper On Various Web Page Ranking Algorithms In Web Mining

Palak Thacker[1], Asst.Prof. Chintan Thacker[2]

[1]Department of Computer Engineering HJD Institute of Technical Education And Research, Kera Kutch

[2]Department of Computer Engineering HJD Institute of Technical Education And Research, Kera Kutch

Abstract — *The WWW consist a large numbers of interconnected web pages that provide information to user. When user performs a query, search engine returns numbers of web pages as a search result. So ranking of page is very important so that user will get requested search result in top of few links. Page Rank is one of many factors that decide where your web page appear in search result ranking. There are many technique such as PageRank, HITS, Weighted PageRank used to rank website as per their importance and relevancy of query. As numbers of web pages increase day by day traitional PageRank algorithm undergo several enhancement and modification. In this paper we discuss how traditional PageRank algorithm works and modification done in standard PageRank algorithm.*

Keywords— *Web mining, PageRank, Ranking, modified PageRank, Search Engine.*

I. INTRODUCTION

We know that web is largest source of data storing and retrieving. Searching on web is most widely used operation on World Wide Web. Nowadays there are hundreds of millions of web pages are online available and roughly millions of web pages are adding to them. The largest search engine Google receives over 200 million queries each day through its various services. So with rapid information sources available on internet, it has necessary for user to use automated tool to retrieve desired information. User may not see all the retrieved pages so PageRanking of web pages is very useful. Ranking is generally classified into three categories: content-based, web structure-based, web usage-based.

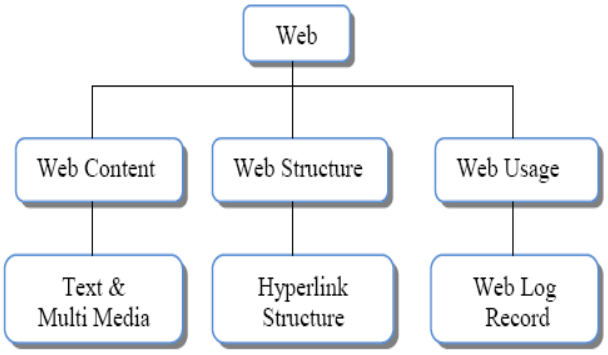

Fig 1. Categories of web mining

Web Content Mining is the process of discovering useful information from the contents of Web pages. Web content mining involves text, images, audio, video information. It is related to text mining because most of website contents are text. It is also related to image mining. Process of Image mining is quite difficult compare to text.

Web structure Mining deals with searching and modeling the web's link structure. Web structure mining consists of nodes (Web pages), as well as edges (hyperlinks) linking between two related pages it is the process of discovering structure information from the Web. Web structure also consist of In-degree and Out-degree Hyperlinks. A Hyperlink is used to connect a different Web page to other web page of different location.[15]

Web Usage mining has been used for various purposes:

1. A knowledge discovery process for mining marketing intelligence information from web data.

2. In order to improve the performance of the website, web usage logs can be used to extract useful web traffic patterns. Web usage mining provides valuable knowledge about user behaviour on WWW. One of the major goals of web usage mining is to reveal interesting trends and patterns which can be provide useful information about the user of a system. It includes web server log such as user's IP, referral URL, response status and HTTP request and other.

II. PAGERANK

In 1998, the founder of google Sergey brin and Larry Page developed the PageRank algorithm to decide the importance of millions of pages comprising on WWW.PageRank ranks the web page based on the web structure. Google most commonly used search engine use PageRank algorithm to rank the web pages. PageRank is developed by Google and named after Larry Page, Google's co-founder and president[1]. PageRank is numeric value which represent importance of web page on web. In PageRank when one page links to another it means it is casting a vote for that page. The more votes that are cast for a page, the more importance the page must be. Further, the importance of the page that is casting the vote determines how important the vote itself.Back Links can be described as links that are directed towards a website.

"The more back links website will have, the more frequently it will be visited by search engines robots, which means that new content will be indexed much faster.[2] "In short PageRank is a "vote", by all other pages of site on the web, about how important a page is.[3] If there's no link there's no support.[3]

2.1 PAGERANK ALGORITHM

PageRank is numerical value that represent the importance of web page based on number of inbound links. The main concept of PageRank is that the PageRank is directly proportional to number of web pages linking to that page.[4]

We assume page A has pages T1...Tn which point to it (i.e., are citations). The parameter d is a damping factor which can be set between 0 and 1. We usually set d to 0.85. There aremore details about d in the next section. Also C(A) is defined as the number of links going out of page A.[1]

The PageRank of a page A is given as follows:

PR(A) = (1-d) + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn)) (1)

Where,

PR(A)= PageRank og web page A.

d= Damping factor. Generally taken as 0.85

PR(Tn)=PageRank of pages n, that links to page A.

C(Tn)=Numbers of outbound links of web page n.

Note that the PageRanks form a probability distribution over web pages, so the sum of all web pages' PageRanks will be one.[1]

PageRank or *PR(A)* can be calculated using a simple iterative algorithm, and corresponds to the principal eigenvector of the normalized link matrix of the web.[1]

PageRank is also calculated by power method. In power method it use matrix to calculate the PageRank. To calculate the PageRank the equation is as below

$$\mathbf{P=(1-d)e + dA^T P} \qquad \mathbf{(2)}$$

Let begin by taking web net as a directed graph, with nodes represented by web pages and edges represented by web pages and edges represented by the links between them.

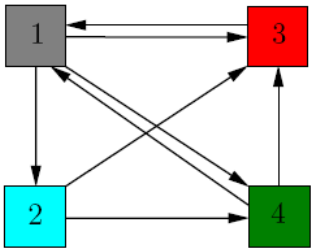

Fig 2 Web graph structure

Here each page should transfer evenly its importance to the pages that it links to. Node 1 has 3 outgoing links, node 2 and 5 have 2 outgoing links and node 3 has only 1 outgoing link. So here for this graph A is as below

$$\begin{bmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{bmatrix}$$

A^T is conversion of row to column and column to row of A.

Here there is 4 node so e is define as below

$$e = \begin{pmatrix} ¼ & ¼ & ¼ & ¼ \\ ¼ & ¼ & ¼ & ¼ \\ ¼ & ¼ & ¼ & ¼ \end{pmatrix}$$

d is generally taken as 0.85. So by putting all this value in eqn (2)we will get the answer as two consecutive iteration match.

The PageRank theory holds that an imaginary surfer who is randomly clicking on links will eventually stop clicking. The probability, at any step, that the person will continue is a damping factor d. Various studies have tested different damping factors, but it is generally assumed that the damping factor will be set around 0.85.[1]

The major disadvantage of PageRank algorithm is that it stores the rank value at indexinx time not at query time. PageRank calculated by PageRank algorithm depends on stored database which is updated after decided interval.

2.2 Ranking Web Page Based On User Interaction Time

In 2015 Nandnee jain[5] proposed an algorithm which ranks web pages based on user interaction time. The idea is to compute access time of each page and then compute average access time on basis of which ranking of web page is calculated.[5]

In this algorithm when user submit a query, on basis of that submitted query a no of URL retrieves. Then for each of access URL access time of visited URL is calculated. A web log data is created for each of visited web page with their respective access time. Then Comparison of access time with minimum and maximum threshold access is take place. If the access time is greater than maximum threshold time set the access time as maximum threshold time. If the access time is less

than minimum access time set access time as zero. At last PageRanking is calculated based on access time of each web page. Formula for calculating this is as below(3)

PR(Urli)= (1-d) + d* (PR(Urli)/C(Urli)) **(3)**

Where,

Urli is retrieved URL's

Comparison of access time for various web pages using traditional PageRank algorithm and proposed algorithm, proposed algorithm provides much efficient and accurate prediction of ranking of web pages over internet.[5] In case of computational time and goodness factor this algorithm performs better compare to existing algorithm.

2.3 A Novel User Preference and Feedback Based Page Ranking Technique

In 2015 Pooja Devi[6] proposed an algorithm based on user preference and feedback. Traditional approach do not at all consider user preference, feedback and user interest while ranking the web page. In this approach algorithm that combines web structure, web usage and web content mining. This algorithm include content based features like user preference and feedback. Content comes from how much query match the text. User preference comes from domain profile. User feedback comes from average time spent by user. Formula of this algorithm is as below(4)

PageRank Score= 0.2*PR + 0.2* PH score + 0.2*PC score + 0.3*Domn (4)

Where,

PR= Traditional PageRank algo

PH score= Average user time spent on web page.

Domn= 1-If web page matches user preferred domain 0-If web page does not matches user preferred domain.
PC score= 0.5*title text + 0.6* body text (content of doc).

A score value of 1,0.5 and 0 was assigned to pages marked as more relevant, less relevant, irrelevant respectively. The main advantage of this algorithm is that user get full information in very first few URLs. This proposed system seems to provide us a mechanism that can help retrieve high quality documents with maximized user satisfaction.

2.4 Spread Influence algorithm Of News Website Based On PageRank

In 2015 Guo Wei Chen[7] proposed a PageRank algorithm with website attention factor. Experiments shows that the algorithm based on PageRank algorithm effectively reflecting spread influence of the news site.[7] Spread influence is key factor to achieve business purpose or to complete elements of their mission. Influence is ability to be with others in process of communication , influence or change other people's mentality and behavior. NWRank(News Website Rank) the algorithm adopts a similar PageRank iteration mechanism by constant iteration each user's NWRank value is obtained s as to find the most influence news site.

It found in process of analysis and research on the dynamics of human behavior and link analysis, PageRank value of page are evenly distributed to all associated pages. This allocation will result in the old pages have more higher PageRank value than the new pages. However on the importance of web pages, real-time, new web pages should be better than old pages having high value.

In the first step according to the website attention model calculate for each focus attention value at time t. The second step, to set initial value of all website NWRank. The third step is to calculate the value of each site NWRANK according with NWRank. Formula for this step is as below(5)

$$\mathbf{NWR_{(v,t)}} = 1 - \gamma_v(t) + \gamma_v(t) \sum_{\cdot u(v,u)\epsilon E} A(u,v) NWR(u,t) \quad \textbf{(5)}$$

Where

r(t) is average probability that time t, v website to be forwarded to other websites.

$W_{(u,v)}$(t) is attention value at time t between website u and website v. Website V followers number at time t.

Comparing to classic algorithm this algorithm results in rela time to better reflect the influence of site ranking.

2.5 Page Ranking Based On Numbers Of Visits Of Links Of Web Page

In 2011 Gayendra K., Neelam D.and A.K.Sharma[8] proposed an PageRank algorithm based on numbers of visits of links. This concept is used to display most valuable pages on the top of the result list based on user behavior, which reduce the search scale. The algorithm assign more rank value to the outlinks which is most visited by the user. It assume that more popular the web pages are, most linkages other web pages tend to have to them or are linked to by them. This algorithm assigns large rank values to more important pages instead of dividing the rank value of page evenly among its outgoing linked pages. Each outgoing page gets a value proportional to its popularity. PageRank is calculated based on inbound links because we assign more rank value to the outgoing link which is most visited by user. Formula of PageRank based on Visits Of Links is as below(6)

$$\mathbf{PR(u)= (1-d) + \sum L_u(PR(v))/TL(v)} \qquad \mathbf{(6)}$$

Where,

L_u denotes number of visits of links which is pointing page u from v

TL(v) denotes total number of visits of all links present on v.

d is damping factor.

Database or web file access by crawler at the time of crawling. The crawl's information stored in search engines database, which is used to calculate Rank value of different pages.

This algorithm provide more relevant result than traditional algorithm. As a result user may find the described content in top of few pages , so search scale can be reduced to large scale.

Scalability and more experiments and evolutions are the future work of this algorithm.

III. CONCLUSION

In this paper, we discussed how traditional PageRank algorithm calculates the PageRank. With course of time traditional PageRank algorithm has also modified by adding many components. The main issue of traditional PageRank algorithm is relevancy, because it calculates the PageRank at the time of indexing. As we seen above, the overall focus is to modify algorithm which can give result at the time of indexing and at time of query. With increasing data or information on every web page day by day more and more work is going on relevancy so user can get more relevant pages on top of the result list. Besides of this no more work or modification is done on traditional PageRank algorithm to reduce its numbers of iterations. Increasing efficiency by reducing time complexity is also an area of research with relevancy.

REFRENCES

[1] L. Page, S. Brin, R. Motwani, and T. Winograd. The Pagerank citation ranking: Bring order to the web. Technical report, Stanford University,1998.

[2] Meng Cui, Songyun Hu, 2011: 'Search Engine Optimization Research for Website Promotion', IEEE computer society.

[3] Ali H. Al-Badi, Ali O. Al Majeeni, Pam J. Mayhew and Abdullah S. Al-Rashdi, 2011: 'Improving Website Ranking through Search Engine Optimization', Journal of Internet and e-business, 2011, Article ID 969576.

[4] Sanjay* and Dharmender Kumar, "A Review Paper on Page Ranking Algorithms", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4 Issue 6, June 2015.

[5] Nandnee jain and Upendra Dwivedi "Ranking Web Pages Based on User Interaction Time ", 2015 International Conference on Advances in Computer Engineering and Applications (ICACEA) IMS Engineering College, Ghaziabad, India.

[6] Pooja Devi," A Novel User Preference and Feedback Based PageRanking Technique" 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)
[7] GuoWei Chen, "Spread Influence Algorithm of News Website Based on PageRank" 2015 IEEE ICIS 2015, June 28-July 1 2015, Las Vegas, USA.
[8] Gyanendra K., Duhan N., and A. Sharma, 2011: 'Page Ranking Based on Number of Visits of Links of Web Page', proc. India International Conference on Computer & Communication Technology (ICCCT).
[9] Sanjay* and Dharmender Kumar, "A Review Paper on Page Ranking Algorithms", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4 Issue 6, June 2015.
[10] Seema Rani, "A Review Paper On Web Page Ranking Algorithms", International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 3 Issue 8 August, 2014 Page No. 7946-7949.
[11] Rekha Jain, "Page Ranking Algorithms for Web Mining", International Journal of Computer Applications (0975 – 8887) Volume 13– No.5, January 2011.
[12] N. V. Pardakhe, "Analysis of Various Web Page Ranking Algorithms in Web Structure Mining", International Journal of Advanced Research in Computer and Communication Engineering Vol.2, Issue 12, December 2013
[13] http://en.wikipedia.org/wiki/PageRank

[14] SEO Tips, 2011: 'Backlinks', accessed January 2013, http://www.seotipsy.com/backlinks
[15] Punit Patel,"Research of Page ranking algorithm on Search engine using Damping factor", *International journal of Advance Engineering and Research Development (IJAERD) Volume 1 Issue 1, February 2014, ISSN: 2348 - 4470*

