Survey on Different Page Ranking Algorithms Based on Links and Time

Jinal V.Patel[1], Prof. Rimi Gupta[2]

M.E. Student, Dept. of Computer Engineering, Sardar Vallbhbhai Patel Institute of Technology, Vasad, Gujarat, India

Assistant Professor, Dept. of Computer Engineering, Sardar Vallabhbhai Patel Institiute of Technology, Vasad,

Gujarat, India

ABSTRACT: Page Rank is extensively used for ranking web pages. There are many algorithms for page ranking such as Page Rank algorithm, Weighted Page Rank algorithm, Enhanced- ratio rank algorithm, Page Ranking Based on the Reading Tim**e** etc. The page ranking algorithm reflects the popularity of a web page in its page rank score. Due to the changing nature of web number of web pages are deleted and added newly. Every time a surfer searches web using the search engine, data should be fresh and relevant. So, to retrieving efficient, relevant and meaningful information from large sources of information is very challenging job and also it is very difficult to retrieving information which based on the user interest.The main aim of this paper is to discuss the various existing page ranking algorithms and the modification done to the standard page rank algorithm.

 KEYWORDS: Web Mining, Page Rank, Weighted Page Rank, Enhanced-Ratio Rank, Reading Time Based Page Rank.

I.   INTRODUCTION

        The Internet is the collection of large number of data that serves millions of users worldwide. It is a huge store of distributed documents. Internet is increasing day by day so there is a challenge for website owner to provide proper and relevant information to the internet user. But all of this information is not relevant to user. There are various challenges associated with the ranking of web pages such that some web pages are made only for navigation purpose and some pages of the web do not possess the quality of self-descriptiveness. For ranking of web pages, several algorithms are proposed including Page rank algorithm, Weighted Rank algorithm, Enhance-Ratio Rank algorithm. This Algorithm is Useful for Rank the pages. This three algorithm is based on the links which is related to the webpages. In Page Rank Algorithm Backlinks are used as a input parameter, In Weighted Rank algorithm Backlinks and Forward links are consider as a input parameter, In Enhanced-Ratio rank algorithm Inlinks, Outlinks and VOI is use as a input parameter.
        Web miningis the use of data mining techniques to automatically discover and extract information from Web data - including Web documents, hyperlinks between documents, usage logs of web sites, etc.It is the information service center for news, e-commerce, and advertisement, government, education, financial management, education, etc.We have developed Web mining framework for evaluating ecommerce web sites.
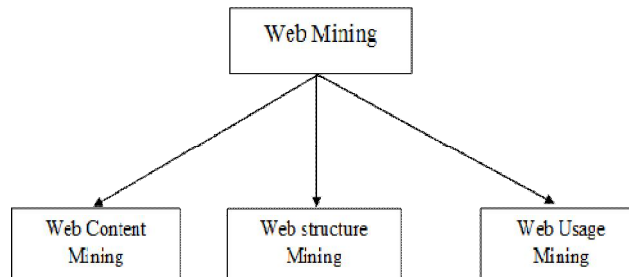
Figure: 1.1. Web mining framework

Web Content Mining: WCM is the process of extracting the useful information and knowledge from the web page content.

Web Structure Mining: WSM mainly analyses the node and link structure of a web site.

Web Usage Mining: WUM is the process of discovering the data stored in server access logs, user profile and pattern in user browsing the web pages.

II.   RELATED WORK

Since the early stages of the World Wide Web, search engines have developed different methods to rank web pages. For the purpose of better search results and especially to make search engines automatically generated web pages based upon the analysis of content specific ranking criteria, the concept of link popularity was developed. Using this concept, the number of inbound links for a document measures its general importance. Hence, a web page is generally more important, if many other web pages link to it.

So, the page rank is a numeric value that represents the importance of a page present on the web. When one page links to another page, it is effectively casting a vote for the other page. More votes implies more importance. A web page is important if it is pointed to by other important web pages. Google calculates a page's importance from the votes cast for it. Importance of each vote is taken into account when a page's Page Rank is calculated. Page Rank is Google's way of deciding a page's importance. It matters because it is one of the factors that determines a page's ranking in the search results. PageRank is based on the linking structure of the whole web.

There are many algorithms which is useful for the rank the pages.

A.   PAGE RANK ALGORITHM

The PageRank algorithm was developed by Google's founders, Larry Page and Sergey Brin, when they were graduate students at Stanford University [1]. Which is the first algorithm for the page rank. PageRank algorithm is simply to start from the perspective of hyperlinks-based structural analysis for measuring the relative importance of web pages. PageRank, a method for computing a ranking for every Web page based on the graph of the web, has applications in search, browsing, and traffic estimation. A page has high rank if the sum of the ranks of its backlinks is high. So, this algorithm is only based on the backlinks.

PageRank value PR (A) of page Acan be calculated by the following two steps:

$$\textbf{PR (A) = (1-d) + d} \left(\frac{PR(T1)}{C(T1)} + \frac{PR(T2)}{C(T2)} + \ldots + \frac{PR(Tn)}{C(Tn)} \right) \ldots\ldots \textbf{(1)}$$

Where,

- d is a damping factor which can be set between 0 and 1, usually set its value to 0.85.
- PR (T1) is the incoming link to page A and C (T1) is the outgoing link from page T1 (such as PR (T1)).

B. WEIGHTED PAGE RANK ALGORITHM

This weighted page rank algorithm was proposed by Wenpu Xing and Ali Ghorbani which is an extension of PageRank algorithm [2].This Algorithm assigns rank values to pages according to their importance rather than dividing it evenly. The importance is assigned in terms of weight values to incoming and outgoing links. WPR takes into account the importance of both the Inlinks and the Outlinks of the pages and distributes rank scores based on the popularity of the pages. The popularity from the number of Inlinks and Outlinks is recorded as $W_{(v,u)}^{in}$ and $W_{(v,u)}^{out}$ respectively.

The $W_{(v,u)}^{in}$ and $W_{(v,u)}^{out}$ are used to record the popularity of the Inlinks and Outlinks based on the Inlinks and Outlinks of that link.

The mathematical equations of the weights are given as follows:

$W_{(v,u)}^{in}$ is the weight of link(v, u) calculated based onthe number of inlinks of page *u* and the number of inlinksof all reference pages of page *v*.

$$W_{(v,u)}^{in} = {I_u}\Big/{\sum_{p \in R(v)} I_p}$$

Where,
- Iuand Iprepresent the number of inlinks of page *u* and page *p*, respectively.
- R (v) denotes the reference pagelist of page *v*.

$W_{(v,u)}^{out}$ is the weight of link(v, u) calculated based onthe number of outlinks of page *u* and the number of outlinksof all reference pages of page *v*.

$$W_{(v,u)}^{out} = {O_u}\Big/{\sum_{p \in R(v)} O_p}$$

Where,
- Ouand Oprepresent the number of outlinks of page *u* and page *p*, respectively.
- R(v) denotes the reference page list of page *v*.

Considering the importance of pages, the original PageRank formula is modified as:

$$PR(u)=(1-d)+d*\sum_{v \in B(u)} PR_{(v)} W_{(v,u)}^{in} W_{(v,u)}^{out} \ \dots\dots (2)$$

Where,
- PR (u) and PR (v) are ranking of the webpages u and v respectively,
- $W_{(v,u)}^{in}$ and $W_{(v,u)}^{out}$ are used to record the popularity of the Inlinks and the Outlinks,
- d is the dampening factor,
- B(u) are the pages which points to webpage u.

C. PAGE RANK ALGORITHM BASED ON VISIT OF LINKS

This Page Ranking based on number of visits of links of web pages is improvement of the page rank algorithm by takes number of visits of inbound links of Web pages. In this algorithm we 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.This concept is very useful to display most valuable pages on the top of the result list on the basis of user browsing behavior, which reduces the search space to a large scale.

The modified version based on VOL is given in equation [4]

$$PR (u) = (1-d) + d * \sum_{v \in B(u)} L_u (PR (v))/TL (v) \ \dots.(3)$$

Where,
- d is a dampening factor that is usually set to 0.85
- Lu denotes number of visits of link which is pointing page u from v.

- B(u) is the set of pages that point to u,
- PR (u) and PR (v) are ranking of the webpages u and v respectively,
- TL (v) denotes total number of visits of all links present on v.

D. WEIGHTED PAGE RANK ALGORITHM BASED ON THE VISIT OF LINKS

This weighted page rank algorithm based on number of visits of links of web pages is improvement of the weighted page rank algorithm by including visit of links [5]. The main purpose of this algorithm is finding more relevant information according to user's query. So, this concept is very useful to display most valuable pages on the top of the result list on the basis of user browsing behavior, which reduce the search space to a large scale.This algorithm is improvement of the page ranking algorithm by adding visit of links in page rank algorithm and also give the briefly about the weighted page rank algorithm and the briefly describe the weighted page rank algorithm based on visit of links.In this weighted page rank based on visit of links algorithm they assign more rank value to the outgoing links which is most visited by users and received higher popularity from number of inlinks. They do not consider here the popularity of outlinks which is considered in the original algorithm. The advanced approach in the new algorithm is to determine the user's usage trends. The user's browsing behavior can be calculated by number of hits (visits) of links.Calculate the $W_{(v,u)}^{in}$ for each node present in web graph by applying the same equation which is use in weighted page rank algorithm.

The modified version based on WPR (VOL) is given in

$$WPR_{vol}(\mathbf{u}) = (\mathbf{1\text{-}d}) + [\mathbf{d} * \sum_{\mathbf{v \in B(u)}} \frac{\mathbf{WPR\ vol(u)\ W_{(v,u)}^{in} * L_u}}{\mathbf{TL(v)}}] \ ....(4)$$

Where

- u represents a web page,
- B(u) is the set of pages that point to u,
- d, is the dampening factor.
- WPRvol (u) and WPRvol (v) are rank scores of page u and v respectively,
- Lu denotes number of visits of link which is pointing page u form v.
- TL(v) denotes total number of visits of all links present on v.

E. ENHANCED-RATIO RANK ALGORITHM

The Enhanced- Ratio Rank is the extension to the Weighted Page Ranking Algorithm in which more weights are given to the Inlinks and Outlinks on the basis of the popularity of the links [6]. In this algorithm the page is considered to be more important if the Inlinks of that webpage is visited by the user more than any other webpage and many other good pages out linked by it, means in overall it may be said that all the features which are considered will come together to rank the webpage. The Enhanced-Ratio Rank both the Inlinks and the Outlinks are being considered for computing the page ranking, and the number of times the user visit the particular link.The $W_{(v,u)}^{in}$ and $W_{(v,u)}^{out}$ are used to record the popularity of the Inlinks and Outlinks based on the Inlinks and Outlinks of that link. The mathematical equations for calculating weights of $W_{(v,u)}^{in}$ and $W_{(v,u)}^{out}$ is same as the use in weighted page rank algorithm [6]

Mathematical Equation for Enhanced-Ratio Rank Algorithm

$$\mathbf{PR\ (u)} = (\mathbf{1\text{-}d}) + \mathbf{d} * \sum_{\mathbf{v \in B(u)}} \frac{[(\mathbf{V_u * 0.7 * W_{(v,u)}^{in} + 0.3 * W_{(v,u)}^{out}}).\mathbf{PR(v)}]}{\mathbf{TL(v)}} ....(5)$$

Where,

- PR (u) and PR (v) are ranking of the webpages u and v respectively,
- $W_{(v,u)}^{in}$ and $W_{(v,u)}^{out}$ are used to record the popularity of the Inlinks and the Outlinks,
- 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 of visits of all links present on v,
- B(u) are the pages which points to webpage u.

In algorithm Inlinks of any webpage votes more to increase the rank of the webpage, less of that by Outlinks, that's why more weightage is given to the Inlinks than to the Outlinks and the ratio as 70/30 is defined on the basis of the better results than all other ratios.

F. PAGE RANKING BASED ON THE READING TIME:

Many Page Ranking algorithm gives the results based on the links of the web structure. On the basis of interest of user, the importance of a page is determined. If the page is interested by the user, the reading time of that page will also be larger than the pages don't accord with user's interest. It means the content of that page is most probably the user want to search.The main aim of this algorithm is to rank those pages at higher position that are most likely by the user. This can be done by including the reading time as a time factor into the computation of the ranking algorithm. Time factor of a page shows how much a user like a page. So, using reading time, calculate the rank of those pages which has a high Inlinks and the high reading time. Calculation of the reading time of a web page based on client side script. When a user clicks on a webpage, the script will be loaded on the client side from web server and starts counting time the user spends on a web page. When the user close the web page, than script will send the message to the web server with the information about the reading time or the time which is spends by the user on that pages and the hyperlink.On the server side, a database will used to store all the information about the pages including visit of links. Using this information calculate the rank of that page. And the result will the based on the Inlinks of that pages and the reading time which is spend by the user, so in general results is given based on the user interest of that pages.

The mathematical equation for calculating the rank of the pages using reading time is [7]:

$$\textbf{PR (u)} = (\frac{\textbf{1}-\textbf{d}}{\textbf{N}}) + [\textbf{ d} * \sum_{\textbf{v} \in \textbf{B(u)}} \frac{\textbf{PR(v)} * \textbf{L}_\textbf{u}}{\textbf{TL(v)}}] * \textbf{RT(u)} \text{ .... (6)}$$

Where,
- d is the dampening factor,
- u and v represents the web pages,
- B[u] is the set of pages that points to page u,
- PR[u] and PR[v] are the page ranks of page u and v respectively,
- Lu is the total of visits of link which is pointing page u from page v,
- TL[v] represents total number of visits of all links present on v,
- RT[u] is the maximum of the time that user's take to read a page u,
- N is the total number of web pages.

III. COMPARISON OF PAGE RANKING ALGORITHM

On the basis of analysis, a comparison of various page ranking algorithms is done on the basis of some vaults such as main technique use, methodology, parameter, relevancy, quality of results, and limitations. On the basis of parameters we can find the powers and limitations of each algorithm. This comparison is show in Table 1.

IV. CONCLUSION

In this paper, we discussed the various algorithms and techniques mainly used by search engines in ranking web pages on the internet. The traditional page rank algorithm has been modified by adding many different factors. Google Page Rank Algorithm computes the page ranks of web pages only using inlinks of web pages. There is a need of some kind of modified algorithm that can give results at the time of indexing as well as at the time of user query. Using many factor that is useful for improving the tradition page rank algorithm which include using inlinks and outlinks and also consider the visit of links. Based on the user interest time factor can also add in the page ranking algorithm for the better result and result give according to user interest.

ACKNOWLEDGEMENT

REFERENCES

1. Sergey Brin, Lawrence Page, "The anatomy of a large-scale hyper textual Web search engine", Computer Networks and ISDN Systems-1998.
2. Wenpu Xing, Ali Ghorbani, "Weighted PageRank Algorithm", IEEE- 2004.
3. Neelam Duhan, A. K. Sharma, Komal Kumar Bhatia,"Page Ranking Algorithms: A Survey" IACC- 2009.
4. Gyanendra Kumar1, Neelam Duhan2, A. K. Sharma," Page Ranking Based on Number of Visits of Links of Web Page", ICCCT-2011.
5. Neelam Tyagi, Simple Sharma, "Weighted Page Rank Algorithm Based on Number of Visits   of Links of Web Page" IJSCE)-July 2012.
6. Ranveer Singh, Dilip Kumar Sharma, "Enhanced-RATIORANK: Enhancing Impact of Inlinks and Outlinks", IEEE-2013.
7. Shweta Agarwal, Bharat Bhushan Agarwal, "Reading Time: A Method for Improving the Ranking Scores of Web Pages", International Journal of Computer Applications, August 2013.
8. Sanjay, Dharmender Kumar, "A Review Paper on Page Ranking Algorithms", IJARCET,June 2015.
9. Dilip Kumar Sharma, A. K. Sharma, "A Comparative Analysis of Web Page Ranking Algorithms", IJCSE-2010.
10. Shweta Agarwal, Bharat Bhushan Agarwal, "Comparative Analysis of Various Web Page Ranking Algorithms", International Journal of Computer Science & Information Technology, 2013.
11. Ms.M.Sangeetha, Dr.K.Suresh Joseph, "Page Ranking Algorithms used in Web Mining" ICICES2014.
12. Shikha Goel, Sunita Yadav, "Search Engine Evaluation Based on Page LevelKeywords", IEEE 2012

BIOGRAPHY

Jinal Vinodbhai Patel is a M.E. student in the Computer Engineering Department, Sardar Vallbhbhai Patel Institute of Technolgy , Vasad. Her research interest are page ranking and web mining .

TABLE: 1COMPARISON OF PAGE RANKING ALGORITHM

	Main Technology	Input Parameter	Methodology	Relevancy	Quality of Results	Importance	Limitation
Page Rank	WSM (Web structure Mining)	Back Link	This algorithm computes the score for pages using Inlinks of pages.	Less(Rank the Pages on the Inlinks)	Medium	High. Back links are considered.	-It favors older pages, -Relevancy of the resultant pages to the user query is very less as it does not consider the content of web page.
Weighted Page Rank	WSM (Web structure Mining)	Back Link & Forward links	Weight of web page is calculated on the basis of input and outgoing links and on the basis of weight the importance of page is decided.	Less as Ranking is based on the calculation of weight of the web pages at the time of indexing	Higher than PR	High. The pages are Sorted according to the importance.	-It is based only on the popularity of the web page. -Less Relevant

Enhance-Ratio Rank	WSM (Web structure Mining)	Inlinks, Outlinks, Visit Of Links From User	Rank the pages based on the Inlinks and Outlinks and visit of links using Ratio between Inlinks and Outlinks.	High relevancy	Higher than PR and WPR	High	Theme drift
Reading Time Based PR Algorithm	Web Structure Mining, Web Usage Mining	Server Log	Rank is evaluation on the basis of number of visits of inbound links as well as maximum time taken by a user to read a page.	High	Higher than PR	High	-Theme drift, -Use only inlinks of the web pages.