International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 8 5268 - 5272

A Noval Approach for Web Page Ranking Based on Weights of Links

Prahlad Kr.Sharma
Department of C.S.E.
A.I.E.T.
Jaipur, Rajasthan
Prahlad_sharma6@yahoo.co.in

Sanjay Tiwari
Asst. Prof.
A.I.E.T.
Jaipur, Rajasthan
sanjay76_tiwari@yahoo.com

***Abstract*—**As the web is the large collection of the information and also due to the changing content/nature of the web (plenty of pages or documents and pages are newly added and deleted on the time basis).The information present on the web is of great need, the world is full of questions and the web is serving as the major source of gaining information about specific query made by the user. As per the search engine for the query made a number of pages are retrieved among which the quality of the page that are retrieved is questioned. On the pages retrieved the search engine apply the certain algorithms to bring a order to the pages retrieved so that the most relevant document or pages are displayed at the top of list. Page ranking is done on the basis of the different approaches as the content based approaches, link based approaches. This paper will provide a review to few of the linked based page ranking algorithms.

Keywords-*Inlinks, outlinks, Page Ranking,Inbound links, outbound links , Visit count, Information Retrieval, World Wide Web.*

*****

I. INTRODUCTION

WWW is the large collection of the data and is expanding or changing every minute WWW is the large collection of the webpages or web documents and also contains the structured and unstructured hyperlinking in the terms of inlinks and outlinks of the webpages. And using the hyperlinkg structure of the webpages Page Ranking Algorithm computes a numerical value that decides the importance of the webpages that has been retrieved corresponding to the user query. In this paper we will be surveying the different linked based algorithms. In the paper section II will describes various page ranking algorithms and the sub-sections will describe as, section A will describe the "Standard Page Ranking Algorithm"(by Google)[1] make use of the inlinks of the webpages whose ranking is to be calculated. The number of links of any webpage will play a vital role to decide ranking or importance of that webpage. Sub-section B describes "HITS" (Hypertext Induced Topic Search)[2]which uses the inlinks and oulinks of the webpages in the terms of the hubs and authorities and hence the number of hubs and number of authorities will decide the importance of any webpage. Sub-section C describes the "SALSA" (The Stochastic Approach For Link-structure Analysis)[3] uses the linking structure of the web or webpages in terms of the inlinks and outlinks and uses the mutual reinforcement approach to decide the importance or ranking of any web page, Sub-section D describes the "Weighted Page Ranking Algorithm"[3] uses the inlinks and outlinks of the retrieved webpage in the terms of the inlink and outlink weights which decides the popularity of the links. And hence the popularity alone with the page ranking of the inlinked webpage is incorporated to compute ranking of the web page in Weighted Page ranking Algorithm, Sub-section E will describes the "Page Ranking Based on

Number of Visit of Links Of Webpage"[5].The Page Ranking based on number of visit of links to webpages will also shows the improvement to the basic Page Ranking Algorithm by L. Page in which the weights are evenly distributed to all inlinks to the retrieved page but in case of the Page Ranking Based of Number of Visit of Links of Webpages, number of times the particular page is visited by the user is taken into the consideration for ranking the web pages. Sub-section F will describes the "Weighted Page Ranking Based on Visit of Links"[4] which uses the inlinks and outlinks of the webpages in the terms of weights as WPR[4] to decide the popularity of the links and also uses the number visit of links to decide the ranking of any webpage along with the inlink and outlink weights, Sub-section G will contributes for the discussion of normalization Page Rankin Algorithm [6] which replaces the rank value of the initial round calculated using the standard Page Ranking Algorithm [1], in the algorithm the computational complexity of the overall ranking process will be reduced because it reduces the number of iterations. Sub-section H will contributes towards the explanation of FlexiRank [7] in which the certain parameters of the page is taken and weights corresponding to those parameters are computed and mean of it is rank of the page. Section II. will discuss the brief comparison of different page ranking algorithms based on the various parameters to compare them, the discussed parameters are as the Web Mining Technique(Web Structure Mining, Web Content Mining and Web Usage Mining), Input parameters(Inlinks, Outlinks and the number of visit count of webpage) and Relevancy of the webpages which are retrieved on corresponding to the query made by the user and section IV will conclude the overall review.

II. LINK BASED PAGE RANKING ALGORITHMS

A. PAGE RANKING ALGORITHM

Page Ranking Algorithm[1] is the most widely used algorithm, developed by S.Brin and L. Page[1](co-founder of the Google).web is the collection of hyperlinking structure in the terms of inlinks and outlinks, to rank out the webpages. Ranking means importance the webpage with respect to the query over other retrieved webpages and according to the Page Ranking Algorithm[1] more is the number of inlinks, higher is the ranking of the webpage.

 ISSN: 2321-8169
5268 - 5272

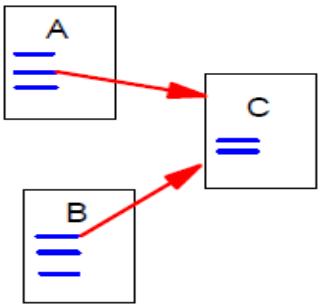


Figure: 1 A and B are the backlinks to C

The mathematical equation for Page Ranking Algorithm is:

$$PR(u) = c \sum_{v=B(u)} PR(v) / Nv \quad (1)$$

Where u depicts the current webpage whose ranking is to be calculated, v is the inlinked webpage of page u. PR(u) and PR(v) are the corresponding ranking of the webpages u and v, B(u) is the group of webpages having inlink to the webpage u, N_v represents the outgoing links from webpage v and c is the normalization factor.

The equation(1) later on modified due to the indirect link of the webpages, where d in equation(2) is the dampening factor and is defined as the probability of following the direct link to any webpage, and (1-d) is the probability of following the indirect link to any webpage. In most of the experimental analysis of page ranking of webpages dampening factor is set to be 0.85.

$$PR(u) = (1-d) + d \sum_{v=B(u)} PR(v) / Nv \quad (2)$$

The Page Ranking Algorithm is limited to the fact that the ranking is done at the time of indexing of the webpages not at the time of retrieval of the webpages.

B. *HITS(Hypertext Induced Topic Selection)*

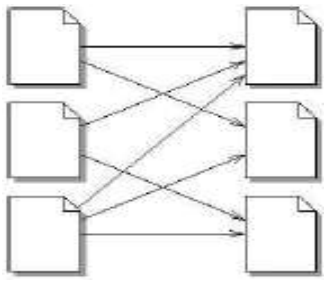
HITS[2] Algorithm uses the hyperlinking structure of the webpage that are being retrieved with respect to the query made by the user. HITS Algorithm divides the retrieved webpages into two types of pages as hubs and authority. Hubs and Authorities are differentiated on the basis of the inlinks and outlinks of the retrieved webpages. Mutual reinforcement approach is used detect the hubs and authorities using the matrix formulation. Good hubs are meant by the number authorities and good authorities are meant by the number of hubs, weights of the hubs and authoritiesdecides the ranking of the webpages.

In the equation(3),(4) a_p and h_p represents the hubs and authority correspondingly, p is the current webpage whose hubs and authorities are to be computed. B(p) and I(p) are referrer and reference webpages corresponding to webpage p.

Mathematical equation for computing the hubs and authorities are:

$$a_p = \sum_{q=B(p)} h_q \quad (3)$$

$$h_p = \sum_{q=I(p)} a_p \quad (4)$$



hubs authorities

Figure: 2 Hubs and Authorities

HITS[2] Algorithm ranks the webpages on the basis of the textual comparison between the query and webpage retrieved with respect to the user query. After retrieval the textual part is ignored and the hyperlinking content is used in the rest of the process. Limitations of the HITS algorithm are topic drift, less efficient and distribution of weights as equal weights are assigned to all retrieved webpages hence the algorithm fails the popularity of the links.

C. *SALSA[8](Stochastic Approach for Link-Structure Analysis)*

In HITS Algorithm the retrieved webpages are divided into hubs and authorities and to overcome this the author in SALSA(Stochastic Approach for Link structure Analysis) replaced the Reinforcement Approach for hyperlinking analysis by a stochastic for doing the same. In Stochastic Approach the interdependency of hubs and authorities is tight as compared to that in the case of HITS algorithm.

The webpages(ℓ) which are retrieved with respect to the user query(t) are the input to the approach defined. Output to the Approach is the bipartite graph with no coupling between the nodes at the same level. From the bipartite graph G if the direct link between nodes as hubs and authorities exist that means there exist a informative link between r and s, where r and s describes the hubs and authorities correspondingly.

A undirected bipartite graph G=(V_h ,V_a ,E) drawn on the basis of the webpages retrieved and the linking structure of that webpages. Where

$$V_h = \{s_h \mid s \in \ell \text{ and out degree(s)>0}\} \text{(the hub side of G)} \quad (5)$$

$$V_a = \{s_a \mid s \in \ell \text{ and in degree(s)>0}\} \text{(the authority side of G)} \quad (6)$$

$$E = \{(s_h, r_a) \mid s \rightarrow r \text{ in } \ell\}. \quad (7)$$

Two adjacency matrix are obtained as hubs matrix and authority matrix from the obtained bipartite graph:

$$h_{i,j} = \sum_{\{k|(i_h,k_a),(j_{h,k_a})=G\}} (1/\deg(i_h))(1/\deg(k_a)) \quad (8)$$

The Authorities matrix is shown as follows:

$$a_{i,j} = \sum_{\{k|(k_h,i_a),(k_h,j_a)=G\}} (1/\deg(i_a))(1/\deg(k_h)) \quad (9)$$

Hubs and Authorities can be obtained from the hubs and authorities matrix and the webpages are ranked accordingly. Stochastic Approach for deriving the hubs and authorities is the major advantage of the SALSA Algorithm for webpage ranking and TMC(Tightly Knit Community) is one of the disadvantage of SALSA which occurs when large number of irrelevant webpages with respect to the user query are retrieved.

D. WEIGHTED PAGE RANKING ALGORITHM

Weighted Page Ranking[4] is the improved version of the standard page ranking algorithm[1]. The equal distribution of the weights was the limitation of the Page Ranking Algorithm[1] which was improved in WPR[4] as the weights of the links is assigned on the basis of the popularity of any particular link and the assigned weights decides the ranking of the webpage. The popularity of any link is decided with the help of inlink and outlink count.

Efficiency of WPR[4] is higher then that of Page Ranking Algorithm[1] in the terms of number of relevant webpages and is limited popularity method for deciding the ranking of the webpage because of which the ranking may also be intentionally incearesed.

Mathematical formulae to decide the weights of inlinks and outlinks are defined in equation (10), (11).

$$W^{in}_{(v,u)} = I_u / \sum_{p=R(v)} I_p \quad (10)$$

Where, I_u and I_p are the number of inlinks of webpage u and webpage p and R(v) is the reference page list of v.

$$W^{out}_{(v,u)} = O_u / \sum_{p=R(v)} O_p \quad (11)$$

Where, I_u and I_p represents the inlinks of webpage u and p, R(v) are the reference webpages, and O_u and O_p are the corresponding outlinks to u and p.

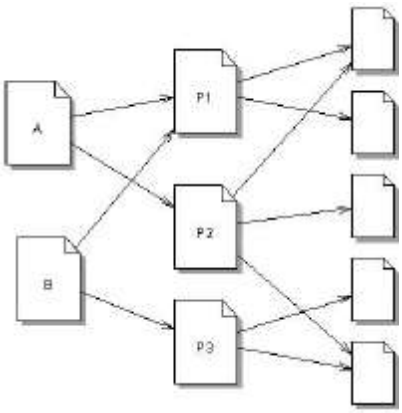


Figure: 3 Links of a hypothetical website

By considering the weights in terms of popularity of links the Weighted page Ranking Algorithm is formulized as:

$$PR(u) = (1-d) + d \sum_{v=B(u)} PR(v) W^{out}_{(v,u)} W^{in}_{(v,u)} \quad (12)$$

Where d defines the dampening factor, B(u) defines the webpages with inlinks to webpage u.

E. WEIGHTED PAGE RANKING ALGORITHM BASED ON NUMBER OF VISITS OF LINKS OF WEB PAGE

Weighted Page Ranking Algorithm based on Numberis improvement over WPR[4] in which the inlinks and outlinks weights are defined based on the popularity of links. In WPR[4] author has also considered the number of times the user visiting the particular link, which is the extra web feature added to the algorithm defined under the web usage mining to rank the retrieved webpages. Hence a user choice is also being considered to the algorithm in the terms of the visit of links. Efficiency of Weighted Page Ranking Algorithm Based on Visit of Links[] in the terms of the number of relevant webpages is higher as compared to WPR[4]because the algorithm manipulates over user choice in terms of visit of links along with inlinks and outlinks weights.
The requirement of specialized search engine for manipulating the ranking algorithm is the major disadvantage of WPRVOL[]. The weights of inlinks and outlinks in terms of popularity are considered for ranking the webpages while WPRVOL[] assigns more weights to the inlinks which are frequently visited by the user.

The mathematical formulae for Weighted Page Ranking Algorithm based on VOL is:

$$WPR_{vol}(u) = (1-d) + d \sum_{v=B(u)} L_u WPR_{vol}(v) W^{in}_{(v,u)} / TL(v) \quad (13)$$

Where d describes the dampening factor, u webpage whose ranking is to be computed, v is inlinked webpage to u, B(u) are the total incoming links to webpage u, $WPR_{VOL}(u)$ and $WPR_{VOL}(v)$ defines the ranking of the webpages u and v, L_u is the visit count of the link from u to v, TL(v) is the total number visit of all links.

F. PAGE RANKING ALGORITHM BASED ON NUMBER OF VISITS OF LINKS OF WEB PAGE

Page Ranking Algorithm based on VOL is the improvement over Standard Page Ranking Algorithm[1]. In the algorithm the author has also undertaken or added the user choice in the terms of visit of inbound links and the weightage is assigned accordingly.
The mathematical formulae for the ranking algorithm is as in equation (14).

$$PR(u) = (1-d) + d \sum_{v=B(u)} L_u (PR(v)) / TL(v) \quad (14)$$

Where L_u is the visit count of the inlinks towards page u, TL(v) represents the total visit count of all links.
A client side script is being used for adding the visit count. A counter is being maintained in search web crawler which is raised by one for each click of visit and the counter is fetched out to complete the ranking process of the webpages. The algorithm is complex when we talk about the design issues and is limited to the inbound links for computing the page ranking.

G. *AN IMPROVED PAGE RANKING ALGORITHM BASED ON OPTIMIZED NORMALIZATION TECHNIQUE*

The Improved Page Ranking Algorithm based on Optimized Normalization Technique[6] is the improvement over Standard Page Ranking Algorithm[1] in which a normalization feature is being added to rank the webpages which result in the less number of iterations in the complete ranking process.
The normalized Page Ranking Algorithm normalizes the ranking of the page by dividing the actual rank of the webpage with the mean value of rank of all the pages which are retrieved on correspondence to any user query. The normalization was done so as to reduce the number of iteration required to reach the convergence point. Steps of the Page Ranking Algorithm based on optimized normalization technique are:

- Assign 1 as the initial rank to all webpages,
- Now apply the page ranking algorithm as in equation (2)

$$PR(u) = (1-d) + d \sum_{v=B(u)} PR(v) / Nv$$

- Ranking value obtained from above step is replaced by the mean value of the ranking of all retrieved webpages
- Then replace the page rank of each webpage by the normalized value as:

$$\text{Norm PR(u)=PR(u)/mean value} \quad (15)$$

Where, norm PR(u) represents the normalized value of the webpage u.

- Assign PR(u)=normPR(u)
- The process is repeated until any two consecutive value are not same.

The discussed algorithm is less complex as compared to SPR[1] because of the reduction in the number of iteration which reduces the computation complexity. Again the concept of standard page ranking algorithm is being used hence only inbound links of any page is only used to calculate the ranking which is the major disadvantage of this algorithm.

III. PROBLEM STATEMENT

Dangling Links, surfer jamming and hyperlinking loops are major cause of similarity ranking of the webpages. Dangling links are those links from which no other pages can be reached means with zero oulinks and hence results in the similarity ranking of the webpages while ranking the webpages. Similarity ranking is the situation while ranking the webpages when same ranking is assigned to many of the webpage due to above mentioned problems which affects the ranking or deciding the relevancy of the webpages.

IV. PROPOSED ALGORITHM

In this paper a new page ranking algorithm is discussed on the basis of hyper-linking of the pages in the terms of the inlinks and the outlinks. The algorithm considers the weights of the inlinks and outlinks which decides the popularity of the links. The proposed algorithm incorporated the page ranking of the inlinked and outlinked webpage to decide the ranking of current webpage. The proposed Page Ranking Algorithm uses the inlinks and outlinks computation part from the Weighted Page Ranking Algorithm[6] in which weights of the inlinks and outlinks are taken as the factor to compute the page ranking of any webpage where weight defines the popularity of the links.The problem of the dangling links is the major limitation of the weighted page ranking algorithm and the similarity ranking of the webpages is also one of the limitation of the algorithm. The proposed page ranking algorithm considers the weight of inlinks, weight of outlinks along with the ranking of the webpages to which the link goes or comes from different in the case inlinks and outlinks. As the inlinks and outlink weights are taken in additive nature in the page ranking expression to the ranking which helps in avoiding the dangling link problem and also avoids the similarity ranking of the webpages. If either factor is zero then the ranking to the webpage is assigned on the basis of another factor means the additive nature helps in avoiding the similarity ranking of the webpages. The proposed page ranking algorithm also increase the relevancy as the ranking is on the basis of the ranking of the third level connected webpage and hence avoids the intentionally increasing the ranking of the webpage.

The earlier equation was as follows[6]:

$$W^{in}_{(v,u)} = I_u / \sum_{p=R(v)} I_p \quad (9)$$

$$W^{out}_{(v,u)} = O_u / \sum_{p=R(v)} O_p \quad (10)$$

Where $W^{in}_{(v,u)}$ and $W^{out}_{(v,u)}$ are used to record the popularity of the inlinks and the outlinks, d is the dampening factor, I_u and I_p are the number of inlinks of webpage u and the webpage p, O_u and O_p are the number of outlinks of the webpage u and p.

The mathematical formula to compute the ranking :

$$PR(u) = (1-d) + d \sum_{v \in B(u)} ((W^{in}_{(v,u)} PR(I_v) + W^{out}_{(v,u)} PR(O_p)) PR(v))/TL_v \quad (11)$$

Where d represents the dampening factor means the probability that user will follow the direct link, PR(i_v) and PR(O_v) represents the ranking of the webpages which are followed by the inlinks and outlinks whose weights are computed, PR(u) and PR(v) are ranking of the webpages u and v, TL_v represents the total number of outgoing links from webpage v, B(u) are the webpages which points to webpage or having outlink towards webpage u.

Step by step execution of the proposed page ranking algorithm.

Step 1: Hyperlinking structure of the webpages is obtained in the terms of inlinks and outlinks.
Step 2: Assign 1 as the initial ranking to the retrieved webpages.
Step3: Calculate the ranking of the webpages using the proposed webpage ranking formula in equation (11).
Step 4: The complete ranking process is iteratively repeated until a stability in the ranking of the webpages is not reached means doesn't shows much of the change in the ranking during the iterative cycle of the computation.

IV. CONCLUSION

In the paper a review and the a new page ranking algorithm is being discussed for webpage ranking based on different parameters, in the parameters hyperlinking is used as the major parameters in the form as inlinks and outlinks to any webpage. Some of the algorithm includes the user behavior in terms of the number of times the user visits the page and the links. The proposed algorithm will consider the inlinks, outlinks and the number of times a user visits the inlink of a webpage. The algorithm will give better results than the standard Page Ranking Algorithm[1], WPR[3] and WPR VOL[4], in the terms of the similaritry ranking amoung the retrieved webpages,also works for avoiding the looping conditions,etc. The proposed algorithm reduces the number of iteration to reach the normalized ranks to the pages..

REFERENCES

[1] S.Brin and L.Page, "The Antonomy of a Large Scale Hypertextual Web Search Engine,"7th Int.WWW Conf. Proceedings,Australia ,April 1998.

[2] J.Kleinberg,"Authoritative Source in a Hyperlinked Environment,"Proc.ACM-SIAM Symposium on Discrete Algorithm,1998, pp. 668-677.

[3] W.Xing and A.Gorbani,"Weighted PageRank Agorithm," Proceedings of the Second Annual Conference on Communication Networks and Services Research,May 2004,pp. 305-314.

[4] N.Tyagi and S. Sharma,"Weighted Page Rank Algorithm Based on Number of Visits of Links of Web Page,"International Journal of Soft Computing and Engineerig(IJSCE),July 2012..

[5] G.Kumar, N. Duhan and A.K. Sharma,"Page Ranking Based on Number of Visits of Web Pages,"International Conference on Conputer & Communication Technology(ICCCT, 2011,pp. 11-14.

[6] H. Dubey and Prof. B.N. Roy,"An Improved Page Rank Algorithm based on Optimized Normalization Technique,"International Journal of Computer Science and Information technologies(IJCSIT),2011,pp.2183-2188.

[7] D. Mukhopadhyay and P. Biswas, " FlexiRank: An Algorithm offering Flexibility and Accuracy for Ranking the Web Pages, Berlin Heidelberg New York, pp. 308-313, 2005.

[8] R.Lempel and S. Moran,"SALSA: The Stochastic Approach for Link-Structure Analysis," ACM Tracsactions on Information Systems,Vol. 19,April 2001,pp. 131-160.

[9] N. Duhan, A.K. Sharma and Bhatia K.K., "Page Ranking Algorithm : A Survey", Proceeding of the International Conference on Advance Computing, pp. 128-135, 2009.

[10] D. K. Sharma and A. K. Sharma ", A Comparative Analysis of the Page Ranking Algorithms" International Journal of Computer Science and Engineering(IJCSE), pp. 2670-2776, 2010.

[11] C. Ding, X. He, H. Zha, P.Husbands and H. Simon ", Link Analysis: Hubs and Authorities on the World," Technical Report: 47847, 2001.

[12] L. Page, S. Brin, R. Mtvani and T. Winogard ", The Page Ranking Citation Ranking: Bring Order to the Web," Technical Report, Stanford Digital Libraries, SIDl-WP, 1999.