An Enhanced Page Ranking Algorithm Based on Weights and Third level Ranking of the Webpages

Prahlad Kumar Sharma*[1] , Sanjay Tiwari#[2]
M.Tech Scholar, Department of C.S.E, A.I.E.T Jaipur Raj.(India)
Asst. Professor, Department of C.S.E, A.I.E.T Jaipur Raj.(India)

Abstract— *Web is the large collection of changeable documents which are changing every second by means of deletion and insertion of webpages and websites. Data retrieved with respect to the user query should be fresh and most relevant according to the query, relevancy means most matching or meaningful for the user or surfer. When a user searches a specific topic or say query by using search engine lakhs of pages, documents are being retrieved on the basis of different searching technic of the search engine and crawler, hence that lakhs of documents retrieved are to be numbered or indexed in such a way that the most relevant or meaningful webpage, document should be at the top in the list. In this paper few of page ranking algorithm are being discussed and are also compared which one another and a new page ranking algorithm is also proposed "An Enhanced Page Ranking Algorithm Based on Weights and Third level Ranking of the Webpages". The proposed algorithm ranks the webpages by comprising the third level importance of pages linked with the retrieved webpage and hence due to the third level importance matching of the webpages the relevancy of the ranking algorithm is more then that discussed in literature.*

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

I. Introduction

World Wide Web consists of huge data in the form of webpages, websites or can say simply as websites, and hence the data is for data gathering purpose of the user. Search Engine(Web Crawlers) are used for gathering the information from the web, for this data gathering process the web crawler performs a numbers tasks (crawling, searching, indexing, sorting, etc) based on the architecture of the type of crawler is being incorporated for searching the web. Sorting is actually the page ranking of the retrieved webpages, a number of mathematical algorithms are being used. As the www is the collection of large number of webpages or documents and the webpages are hyperlinked with each other. The structured web is partitioned on the basis of three mining technics as web structure mining, web usage mining and web content mining. Page Ranking Algorithm uses all three mining technic to sort or to rank the webpages, the algorithm discussed in this paper are based on web structure mining and also web content mining. Some of the link based page ranking algorithm are Page Ranking Algorithm [1], HITS (Hypertext Induced Topic Selection)[2], Weighted Page Ranking Algorithm[4], Weighted Page Ranking Algorithm based on visit of link of webpages[5], Page Ranking Algorithm based on visit of links[6], etc.

The paper is formatted in various sections as section II describes the related work for proposed algorithm or the comparison and strength and weaknesses of various link based page ranking algorithm are being discussed, section III describes the proposed algorithm "An Enhanced Page Ranking Algorithm Based on Weights and Third level Ranking of the Webpages", section IV describes the results and experiments, section V describes the strength and limitations of the proposed algorithm and section VI provides a short conclusion of the paper.

II. Related Work

In this section related work with respect to the proposed algorithm is being presented which uses two of the mining technics of the web mining as web structure mining (inlinks, outlinks), and the web usage mining (visit count of links).

Standard Page Ranking Algorithm [1] is the first sorting which was incorporated by google to rank the webpages developed by S. Brin and L. Page. The page ranking algorithm incorporates the inbound links to rank the retrieved webpages, means the inlinks of the current webpage have given more importance then the others and hence on the basis of number of inbound links decides the order of webpages which are being retrieved with respect to the user query.

HITS[2] Algorithm is one of major used link based page ranking algorithm, inlinks and outlinks are being incorporated to rank the retrieved webpages in HITS Algorithm. Hubs and Authorities are major factors in HITS algorithm, hubs of any webpage represents the inlinks to the webpage or which are very related to the query made by the user and the Authorities are links going away from the current webpage. And if we say practically A Hub is considered as good if it is pointed to many other Authorities and the an Authorities is considered as good if it is pointed by many other Hubs. Theme Drift and topic drift are the major limitation of the HITS algorithm as it gives equal weightage to all the webpages which is later on removed in the extension to the HITS algorithm as PHITS[12] and another limitation as topic drift is removed in the another extension to the HITS algorithm as I-HITS[13] which assigns unequal weightages to the retrieved webpages.

Page Ranking Algorithm based on VOL [6] is the extension to the Standard Page Ranking Algorithm[1], webpages with more number of inlinks are again given more importance over others and also the web usage mining is used in terms of visit of links, the importance or rank to any webpage is decided on the basis of the number of inlinks and the number of user visits the link to that webpage. The Page Ranking Algorithm based on VOL is more relevant in terms of the higher ranking to the importance webpages than the Standard Page Ranking Algorithm[1] because of the usage factor consideration for the calculation of the rank, need of specialized crawler for visit count are the major limitations of this algorithm.

Weighted Page Ranking Algorithm [3] is one step ahead of the Standard Page Ranking Algorithm[1] which only incorporates the inlinks to rank the webpages. Weights of the inlinks and outlinks is factor which is being used to compute the ranking of the retrieved webpages means the in the Weighted Page Ranking Algorithm[3] weights to links to any webpage is decided on the basis of the importance of the webpage or unequal distribution of weights among retrieved webpages. The weights to the links is assigned on the basis of the popularity, popularity in terms of the number of inlinked pages and number of outlinked pages to that link, if the sum of number of inlinked pages and outlinked pages is more than the popularity is also more of that webpage. The $W^{in}_{(v.u)}$ and $W^{out}_{(v,u)}$ are used to record the popularity of the webpages.

Weighted Page Ranking based on VOL[4] is the extension of the Weighted Page Ranking [3], Weighted Page Ranking Algorithm based on VOL[4] also incorporates the weights of the inlinks which is unequally distributed among retrieved webpages and the number of times the user visits the link to that webpage. In], Weighted Page Ranking Algorithm based on VOL[4] more importance is given to the webpages with more number of inlinks but instead of assigning equal weights as that in Standard Page Ranking Aglorithm[1] unequal distribution of weights is being done in], Weighted Page Ranking Algorithm based on VOL[4]. The weights to the inlinks are assigned on the basis of the popularity (in terms of the number of inlinked pages and outlinked pages) of that inlink, more weights are given to those inlinks whose popularity is more and is visited by user by user more frequently. Due to the consideration of the user factor in terms of visit count the, Weighted Page Ranking Algorithm based on VOL[4] is more relevant than, Weighted Page Ranking Algorithm[3]. Assignment of similar weights to all non-visited webpages and the consideration of inlinks are two major limitation of the algorithm.

FlexiRank[7]: An Algorithm Offering Flexibility and Accuracy for Ranking the Web Pages, in which the retrieved webpages are divided as Index Page, Home Page, Article and Advertisement Page. Relevancy weight(relevance weight is the weight of the query term with respect to the relevant document retrieved), Hubs, Authorities weights, the link analysis of the webpage, number of images in a webpage and the special tags (such as header, title tags) are the parameters on the basis of which the retrieved webpages ranked.

The step by step execution of the FlexiRank[7] is as follows:

- Property selection from the retrieved webpages on the basis of user demand.
- Measurement of weights among selected properties of the webpages.
- Ranking the webpages by computing the mean of the weights of all properties of the webpage.

Property selection and assigning weights to that properties provides flexibility to the user and is count among the major advantages of the FlexiRank[7] and a small change in terms of content, structure and link analysis affects the ranking of a page to large extent.

The Improved Page Ranking Algorithm based on Optimized Normalization Technique [6] is he extension to the Standard Page Ranking Algorithm[1], the The Improved Page Ranking Algorithm based on Optimized Normalization Technique [6] adds a normalization factor to the process of ranking the webpages. The normalization process reduces the number of iteration in the ranking process hence results in the reduction in the overall complexity of the ranking process. Step by step formalization of the Improved Page Ranking Algorithm based on Optimized Normalization Technique [6]
is as follows:

Step 1: Assign 1 as the initial rank to all the retrieved webpages using the mathematical equation given in Standard Page Ranking Algorithm[1], *Step 2:* Calculate the mean value of all the rank of the webpages that are retrieved, as equivalent summation of all rank value/ total number of webpages retrieved,

Step 3: Apply the normalization factor means replace the ranking values of the webpages by the mean values as PR(u)=normPR(u) and same is being repeated until the ranking values are not same in two consecutive iterations.

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 outlinks 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 "An Enhanced Page Ranking Algorithm Based on Weights and Third level Ranking of the Webpages" is discussed based on hyper-linking of the pages in the terms of the inlinks and the outlinks. The algorithm takes the weights of the inlinks and outlinks based on the popularity of the links. The proposed algorithm also considers the ranking of the webpages to which the inlinks and outlink whose weights are calculated represents. The proposed Page Ranking Algorithm is the extension to the Weighted Page Ranking [3] 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.

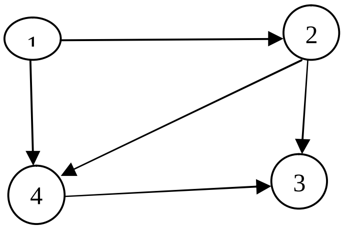


Figure 1: Shows inlinks and outlinks

Figure 1 depicts the flow of inlinks and outlinks from one node or webpages to another in the form of directed links.

The $W_{(v,u)}^{in}$ is the inlink weight which defines the popularity of a inlink on the basis of number of inlinks and outlinks.

$W_{(v,u)}^{out}$ is the inlink weight which defines the popularity of a outlink on the basis of number of inlinks and outlinks associated with those links. The mathematical expression for inlink and outlink weights are given as :

$$W_{(v,u)}^{in} = I_u \,/\, \sum_{p=R(v)} I_p \tag{1}$$

$$W_{(v,u)}^{out} = O_u \,/\, \sum_{p=R(v)} O_p \tag{2}$$

The equation for the proposed algorithm is as follows:

$$PR(u) = (1-d) + d\sum_{v\in B(u)}((W_{(v,u)}^{in} PR(I_v) + W_{(v,u)}^{out} PR(O_p))PR(v))/TL_v \tag{3}$$

Where d is the dampening factor means the probability that user will follow the direct link, PR(i_v) and PR(O_v) are 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.

Algorithm: how actually the proposed algorithm works is as follows:

Step 1: Take the link structure of the retrieved webpages corresponding to any user query.

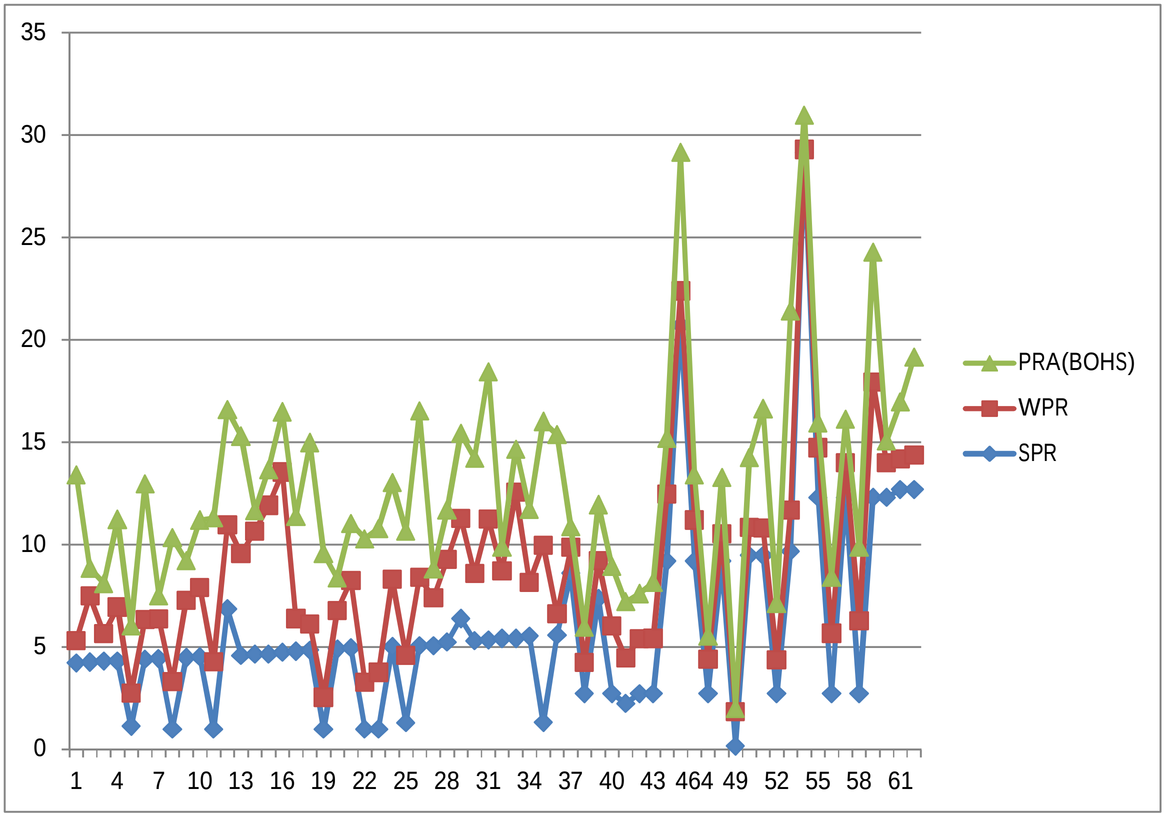
Step 2: Assign 1 as the initial ranking to all the webpages.

Step 3: Calculate the ranking of the webpages using the proposed webpage ranking formula in equation (3).

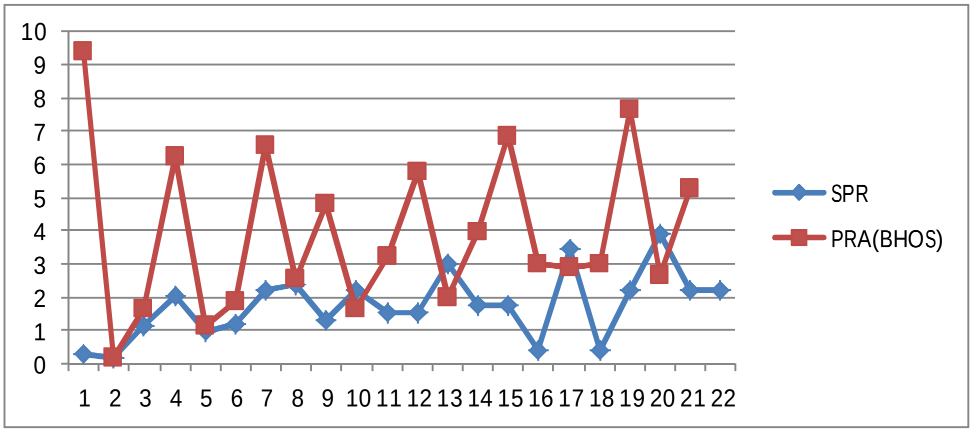
Step 4: Repeat the process iteratively until ranks of all the webpages are not stable means doesn't shows much of the change in the ranking during the iterative cycle of the computation cycle.

V. RESULT AND DISCUSSION

The proposed Algorithm " An Enhanced Page Ranking Algorithm Based on Weights and Third level Ranking of the Webpages" shows better results than Standard Page Ranking Algorithm [1] and the Weighted Page Ranking Algorithm[3] by ranking the non-visited webpage's using the out link weights. The proposed algorithm removes the problem of similarity ranking from the retrieved webpages, the graphs shows the results of proposed algorithm over Standard Page Ranking Algorithm [1] and the Weighted Page Ranking Algorithm[3] on some online links as www.rtu.ac.in, www.aryainstitutejpr.com, www.aryacollege.org.



Graph 1: Represents the similarity results of all three algorithms

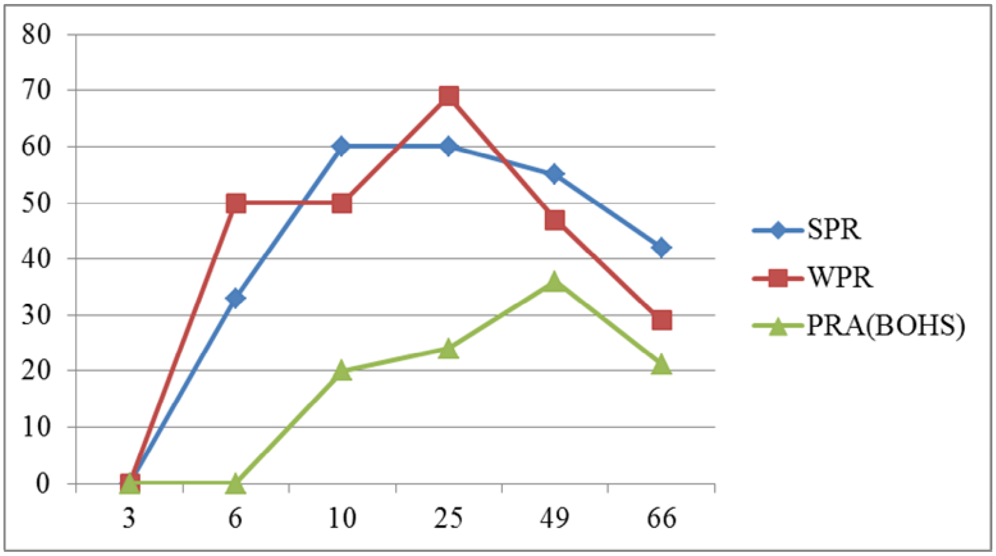


Graph 2: Represents the comparison results of SPR & PRA(BOHS) algorithm.

Table 1:

Table showing the comparison analysis of the SPR[4], WPR[8] and the proposed algorithm on three data sets.

No. of webpages	**Similarity %**		
	SPR[4]	**WPR[8]**	**PRA(BOHS)[14]**
3	0	0	0
6	33%	50%	0
10	60%	50%	20%
25	60%	69%	24%
49	55%	47%	36%
66	42%	29%	21.2%



Graph 3**:** Represents the performance of all three algorithms.

VI. CONCLUSION

In this paper provides a small review of various link based page ranking algorithm as SPR[1], WPR[3], WPRVOL[4], etc and also a new page ranking algorithm based on weights of in links and out links is also discussed in the paper which removes the similarity ranking problem. The proposed algorithm considers the third level ranking of the in linked web pages. As shown in result and experiment section the percentage of similar ranked webpages is less when the proposed algorithm is applied on some link. Some issues as ranking the dangling links, removing the similarity ranking are being resolved in proposed algorithm.

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.

[13] X. Zhang, H. Yu, C. Zhang, and X. Liu, " An Improved Weighted HITS Based on Smilarity and Popularity," Second International Multisymposium on Computer and Computational Science, IEEE, pp.477-480,2007.

[14] P. K. Sharma and Sanjay Tiwari, " A Noval Approach For Web Ranking Based On Weights of Links" International Journal on Recent Trends in Computing & Communication (IJRITCC) 2015, PP 5268-5272.