Augmented Page Ranking Algorithm

Rekha Singhal
Department of Computer Science
Swami Keshvanand Institute of Technology, Management & Gramothan , Jaipur, Rajasthan, India

Saurabh Ranjan Srivastava
Department of Computer Science
Swami Keshvanand Institute of Technology, Management & Gramothan , Jaipur, Rajasthan, India

Abstract- As the world wide web has millions of web pages and consists a large collection of information available online and a plenty web pages which are recently added and removed due to the dynamic nature of web. Due to this huge collection, user faces difficulties to find correct and useful information. Search engine helps to retrieve the results from this huge collection. When user puts a query to search engine, it provides a number of results in the form of a list of relevant web pages. But the ranking of these web pages should be proper. If improper ranking of web pages will be there then user will have to explore whole list to discover the appropriate web page(s). Although many ranking algorithms like HITS algorithm, Page Ranking algorithm, Weighted Page Rank algorithm and their extension are developed to achieve accurate results, but none of these ranking algorithms provide a page ranking with high accuracy. In this paper, an Augmented Page Rank algorithm (APR), an extension to standard Page Rank algorithm is introduced. APR depends upon the web structure. IPR takes the importance of inlinks and outlinks and distribute the page ranking on the basis of the page ranking of all the web pages which are linked to it.

Keywords – World Wide Web, page rank, inlink, outlink, weighted page rank, HITS algorithm.

I. Introduction

1.1 Introduction

In this competitive world and with the wide utilization of the internet in e-learning and e-commerce, to find the need of users are the primary goals of the website , so that they can provide useful information. Therefore, to analyze users' behavior is becoming increasingly essential. So, to discover the content of the Web, the users' past behavior, and the users' need for web pages in future, web mining is used. It consists of three things, Web Structure Mining (WSM), Web content Mining (WCM), and Web Usage Mining (WUM) [3, 4, 5]. This paper will deal with web structure mining. Web structure mining is basically used to discover the structure of all links which are hyperlinked between web pages. In web mining, outlinks and inlinks provide useful, valuable and important information. And this is fact, a web page with high popularity is referred by another important web page the facts that a web- page with high popularity is generally referred to by pages. That's why, web structure mining is considered as an important as well as appropriate approach in the area of web mining. But because of the fast growth in web content, users are lost by this wealthy hyperlinked structure of web. And the primary goals of web owners' are to provide related and useful information to fulfill the users' need. So when queries are put by users to a search engine, a large number of web pages are returned for users in the form of result lists. These lists of results have so many irrelevant and relevant web pages with respect to users' query. The users are assisted to navigate the result list by various applied web page ranking algorithms. These ranking algorithms are used by search engine to arrange the results to be shown to those users. So that users can find the valuable result at first. Many page ranking algorithms were developed, such as HITS, PageRank, SIMRANK etc. Almost all the page ranking algorithms developed are either context or link specific. In this paper, a page ranking algorithm is proposed, called Improved PageRank Algorithm that will also focus on web structure mining. This algorithm is based on ranking of the web page which links to it and ranking of those web pages which are inlinked by that page which links to it. APR algorithm is an extension to the standard PageRank algorithm. The major reason of this proposed algorithm is to find more relevant web pages for user on top of the search results. That's why, to display the most relevant and valuable web pages on the top of the result list according to the user's need and behavior, this concept is useful. Because it will reduce the searching time at a very large scale. More details of this proposed algorithm will be given later in paper.

The rest things of the paper are as follows. Section II will give a review of related work in detailed manner. Section III summarizes the entire done task. Section IV will present various components required in the implementation and

http://dx.doi.org/10.21172/1.72.579

evaluation of IPR. Section V will provide a comparison details and Section VI will provide the conclusions and future scope of the present work.

1.2 *Related Work*

1. *PageRank Algorithm*

This algorithm is developed by L. Page and S. Brin[1]. As we know World Wide Web consists of a linked structure and this algorithm uses this hyperlinked structure to provide the ranking to all web pages. According to this algorithm, so many important inlinks are to any web page, the importance of this web page is increased. It means ranking is done on the basis of inlinks by PageRank algorithm. As the numbers of inlinks are increased, web page has higher ranking value than others. For example:

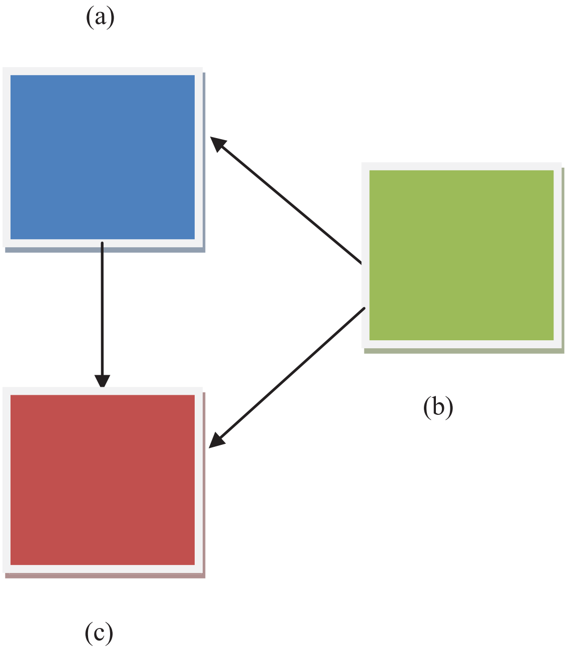

Fig 1 :- Web Graph of three pages

PageRank of a web page can be calculated as:

$$PR\ (u) = (1 - d) * d \sum_{V = B(u)} PR(v) / N_v$$

Here,
d is a dampening factor.
u is a webpage to be found the rank,
v represents the inlinked webpage to web page u,
PR(v) and PR(u) are page ranking of web pages v and u respectively.
B(u) provides the webpages which have inlinks to page u
N is the number of outlinks from page v

2. *HITS Algorithm*

This algorithm takes web pages from a rich hyperlinked web graph for a query and categorizes into two categories, Authorities and Hubs. These categories are based on the hyperlinks. In it, initially a score is given to all web pages that are retrieved for the user's query and then hub and authorities are calculated and then the page rank of a web page is determined on the base of authorities and hubs. And hub and authority can be calculated as follows:

$$A_p = \sum_{q \in B(p)} H_q$$

$$H_p = \sum_{q\in F(p)} A_q$$

3. WEIGHTED PAGE RANK ALGORITHM

This ranking algorithm is based on inlinks and outlinks. It provides a greater number to more important web pages rather than distributing the rank score evenly among its all outbounds. Inlinks and outlinks provide the weight to the link. Inlinks and outlinks are calculated as:

$$W^{\text{in}}_{(v,u)} = \sum_{p\in R(v)} I\text{u} / I\text{p}$$

Here
Iu and Ip are the number of inlinks of page u and page p, respectively.

R (v) shows the reference page list of page v.

$$W^{\text{out}}_{(\ v,u)} = \sum_{p\in R(v)} O\text{u} / \ O\text{p}$$

Here
Ou and Op are the number of outlinks of page u and page p, respectively.

R (v) denotes the reference page list of page v.

Weighted page rank can be calculated as:

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

4. Page Ranking Algorithm based on Number of visit of links

It is an extension to standard page rank algorithm. In it one more parameter is added to PageRank. That parameter is number of visit of links. In this manner the score of a web page is calculated on the basis of the visits of inlinks which is as follows:

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

5. Weighted Page Rank(VOL)

It is an extension to Weighted Page Rank algorithm. In it, with WPR the browsing behavior of user is also considered. This behavior is majored by number of visits of all links. WPR (VOL) is calculated by:

$$WPR_{\text{vol}}\ (u) = (1 - d) + d \sum_{v\epsilon B(u)} \frac{L\text{u}\ WPR_{\text{vol}}\ (v)^{\text{in}}_{(v,u)}}{TL(v)}$$

6. Improved Page Ranking Algorithm

It is an extension to standard page rank algorithm. It takes the mean value of PageRank of all pages and calculate the rank of a web page by dividing its page rank by mean value. This is the formula to calculate Improved Page Rank.

PR(A) = .15 + .85 (PR(T1)/C(T1) + PR(T2)/C(T2) + ……. + PR(Tn)/C(Tn))

mean value of = Summation of page ranks of all web pages / number of web pages
Norm PR (A) = PR (A) / mean value

7. RATIO RANK

In this algorithm, on the basis of popularity the weights of outlinks and inlinks are taken in a particular defined ratio. Popularity defines the no. of outlinks and inlinks. The rank score can be calculated as follows:

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

1.3 *Summarization Of Various Web Page Ranking Algorithms*

Algorithms	Techniques/ Methods	Limitations
Page Rank Algorithm[1]	It is based on the Web graph of web pages that considers inlinks for web page ranking	Ranking is done at the time of indexing
HITS Algorithm[2]	For the ranking of web pages two scores, Hubs and Authority score are used	Theme drift
Weighted Page Rank Algorithm[6]	It is an extension to standard PageRank Algorithm that is totally based on inlinks and outlinks.	Theme Drift
Page Ranking Algorithm based on Number of visit of links[7]	Extended PageRank algorithm with the visit of links	Theme drift
Weighted Page Rank(VOL)[9]	Extended Weighted PageRank algorithm with the visit of links	Theme drift
Improved Page Ranking Algorithm[8]	It replaces the web page ranking by mean value of all web pages ranks	Theme drift and low relevancy
Ratio Rank[10]	In it, a static computation of web pages is done by using inbounds and outbounds with number of visit of links	Theme drift

II. PROPOSED ALGORITHM

As we know the standard PageRank algorithm provides more value to more important web pages. According to it, PageRank of a web page is calculated by its all inlinked web pages' PageRank that are evenly divided by outbounds of these inlinked web pages. But some outbounds do not have relevant content, still those web pages are having equal amount of distribution of that inlinked web page's PageRank. Finally the web page which is more relevant gets less score than it should have. To improve the score of important web pages in graph, an Augmented Page Rank algorithm is proposed. In it, page rank score of a web page is calculated by the page score of its inlinked web page divided by the sum of page scores of all outbounds of this linked web page.
The improved version of standard PageRank algorithm is shown as follows:

$$APR(u) = (1\text{-}d)+d \sum_{v \in B(u)} \frac{APR(v)}{\sum_{y \in F(v)} APR(y)}$$

Notations are:

1. d is a dampening factor,
2. u represents a web page,
3. B (u) is the set of pages that point to u,
4. F (v) is the set of pages that are pointed from v,
5. APR(u), APR (v) and APR (y) are rank scores of page u, v and y respectively

Various steps of this proposed algorithm are as follows:

1. Search an appropriate website:- Search a website that has rich hyperlinks.
2. Create a web graph: - Build a web graph from this website.
3. Apply the proposed formula:- To calculate the rank score of a web page by this formula initially assume page rank of all web pages to be any value. Eg. 1. Then calculate page rank of all pages by this formula
4. Repetition of step 3: Step 3 will be repeated until the values are to be stable.

III. Experiment and Result

In result analysis part, some results, got from the proposed work are described. Here a hyperlinked graph is taken, shown in fig. 1.

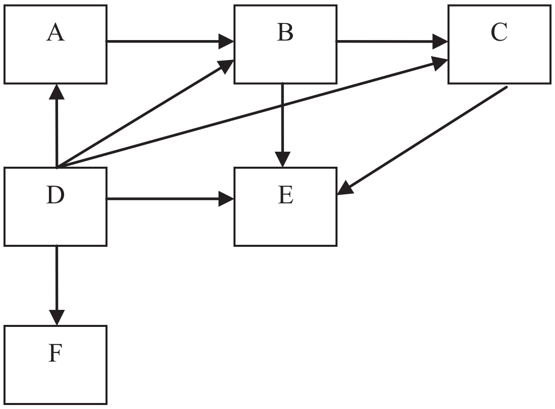

FIG. 1 Hyperlinked web graph

Here we have calculated Rank score of each web page based on standard page rank algorithm and proposed algorithm. After this calculation, different rank scores of the web pages are found.

According to graph page F should have lowest score but according to PageRank both web pages A and F have same rank score. Another issue that web page C should have more rank score than web page B. But if we calculate these scores according to our proposed algorithm, it is giving desired results as shown in fig. 2.

Web Page	SPR	APR
A	.1755	.1777143789
B	.324675	.334803322
C	.313486875	.6005
D	.15	.15
E	.5799507188	.8792
F	.1755	.1926

FIG. 2 Rank score of SPR and APR for different web pages

By the use of fig. 2, we have given two graphical representations of proposed algorithm in fig. 3 and fig. 4. In this we can observe variations between SPR and APR.

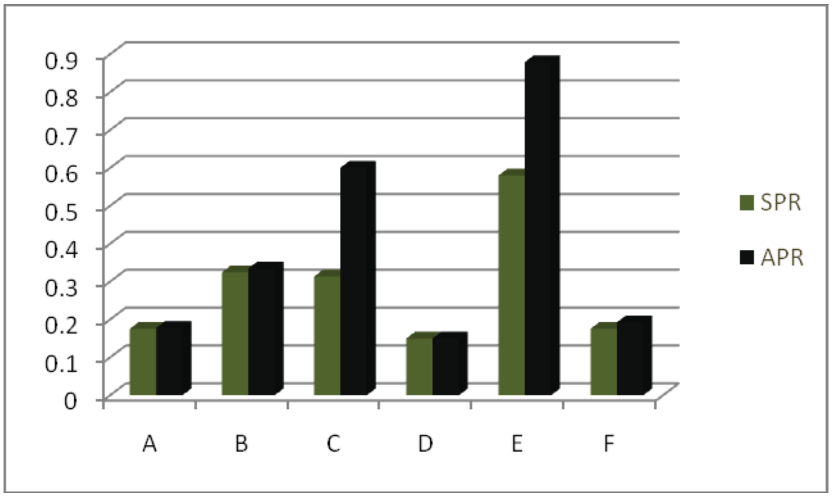

FIG. 3 shows variations between SPR and APR

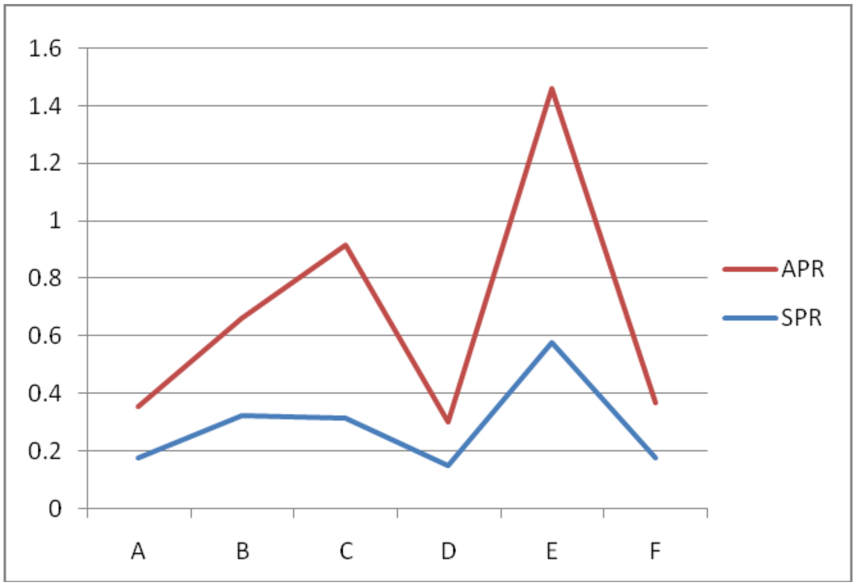

FIG. shows the flow of values of rank score (showing the difference between B and C as well as A and F)

IV.CONCLUSION

This paper presents various page ranking algorithms based on various techniques/ parameters, In paper certain page ranking algorithms, along with the proposed method. All presented ranking algorithms use hyperlinked structure. Some algorithms use inlinks and outlinks, some of others use number of visit of links by user (user behavior). The proposed page ranking algorithm uses the inlinks and outlinks to calculate the page ranking of all web pages which are linked in both forms. The proposed algorithm provides more relevancies in results than original PageRank algorithm. Because no other ranking algorithm calculates the ranking of a web page using the page rank of all web pages which are directly or indirectly linked to it. It will reduce the number of iterations for normalization also. The order of web pages according to this algorithm improves the relevancy of web pages to provide the quality search results to a user.

The future work in proposed algorithm includes the following: a) web page ranking algorithm can include a combination of this page rank and context of a web page on the basis of semantic score [11] (number of synonym words of a user query) and syntactic score [11] (number of exact words of a user query) to improve the results of web page ranking. b) Existed proposed algorithm can use the information of visit of links by the user having feedback from search engine about the web page that is chose by user from the overall result list. c) The problem of theme drift can be removed by some new techniques. d) By introducing some other parameters, some modifications cab be done 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] S. Madria, S. S. Bhowmick, W. K. Ng, and E.-P. Lim. Re- search issues in web data mining. In Proceedings of the Conference on Data Warehousing and Knowledge Discov- ery, pages 303–319, 1999.

[4] R. Kosala and H. Blockeel. Web mining research: A survey. ACM SIGKDD Explorations, 2(1):1–15, 2000.

[5] S. Pal, V. Talwar, and P. Mitra. Web mining in soft com- puting framework : Relevance, state of the art and future di- rections. IEEE Trans. Neural Networks, 13(5):1163–1177, 2002.

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

[7] 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.

[8] 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.

[9] 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.

[10] Ranveer Singh and Dilip Kumar Sharma, “RatioRank: Enhancing the Impact of Inlinks and Outlinks” 3rd IEEE International Advance Computing Conference (IACC), 2013

[11] M. Shalini Amin, S. Kabir and R. Kabir, “A Score based Web Page Ranking Algorithm, “International Journal of Computer Applications(0975 – 8887) Volume 110 – No. 12, January 2015