Improving Pagerank Calculation by using Content Weight

Rutusha Joshi[1], Vinit Kumar Gupta[2]

[1, 2] Hasmukh Goswami College of Engineering, Ahmedabad, India

Abstract: ***World is full of information. The World Wide Web serves as major source of getting such information. Web plays dynamic role because it contains vast data as collection of large number of WebPages and every second new pages are added, updated and deleted in web. Retrieving efficient, relevant and meaningful information from this large source of information is very challenging job. Every search engine applies an algorithm to large number of WebPages in search results which calculators rank of every WebPages and ensure that most efficient and relevant WebPages as per query made by user appear first in search results. In this paper We have analyzed few algorithms which uses link structure or web structure mining and few algorithms which uses web content mining for calculating the page rank value of WebPages and proposed one algorithm which uses both web structure mining as well as web content mining as hybrid for calculating the page rank value of WebPages. This gives better and efficient results as compare to other and overcome some limitations of web structure mining based algorithms.***

Keywords: Visit count, Page ranking algorithm, in links, out links, , weighted page rank, ratio rank, Enhanced RatioRank, Web Structure Mining, Web Content Mining

1. Introduction

With rapid and constant development in World Wide Web, Internet has become the world's most popular, useful and richest source of information. Search starts with World Wide Web. Search engine navigates the web by crawling. In crawling the crawler follow links from page to page. Owners of sites choose whether their sites are crawled or not. Then pages are sorted by their content and other factors and the index keep track of it all. As user search some keywords or query, algorithms get to work and retrieve large number of search results in the form of WebPages. There are millions of WebPages in search results which can be both relevant as well as irrelevant from user's query. It is impossible to check all the results. How to get the most efficient and relevant search results from these large set of search results is the main challenge in web. For this purpose and for efficient and relevant search results as needed by user, many page ranking algorithms are used by search engines which calculate page rank values of the WebPages

The main objective to use page ranking algorithms is to provide page rank values to every webpage in search results and to place most efficient and relevant search results in top of the search results list. There are two categories of page ranking algorithms. They are based either on web structure mining or web content mining. The page ranking algorithm which use web structure mining doesn't care about user's query, Only link structure of WebPages are considered in calculation of page rank value of WebPages. On the other side the page ranking algorithm which uses web content mining take user's query into account and doesn't care about link structure of WebPages for calculating page rank values of the WebPages. Algorithms which use link structure has mainly many challenges like emphasis on old pages, theme drift, page cheating.

2. Pagerank Algorithm

The very basic algorithm used by Google for calculating page rank value is Page Rank Algorithm. Page Rank Algorithm is invented by Sergery Brin and Larry Page one of the Co-founders of Google. Page Rank Algorithm uses web structure mining means link structure for calculating page rank value of any webpage. If any page has more in links pointing to it, the page rank value of that page is high. If a page that pointed by any important pages, the page rank value of that page is high. Following fig. shows one web graph in which page A has three in links from pages B, C and D.

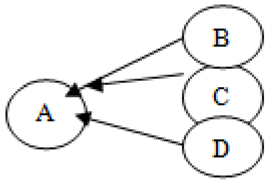

Figure 1: A Web structure

Following equation calculate page rank value of webpage A.

$$PR(A) = \frac{1-d}{n} + d\left(\frac{PR(B)}{O(B)} + \frac{PR(C)}{O(C)} + \frac{PR(D)}{O(D)}\right) \qquad [1]$$

Where, PR (A) is page rank of page A. d is dampening factor for accounting some portion of value to the page which has no in links. It is generally set to 0.85. O (B), O(C) and O (D) are out links of pages B, C and D respectively. Because of page rank uses link structure there are some issues in page rank as explained below:

More emphasis on old WebPages - In link structure based algorithms more in links to the page that means more important the page is and page rank value of that page is high[1]. As compared to new pages old pages have obviously more in links because they exist for long in web. That does not mean that page rank values of old pages are high as compared to new pages. So this is one issue in link structure based algorithms.

Theme Drift- Link structure based algorithms use only link structure means more links to page more important the page is and higher the value of page rank. So the results are independent of the keywords and the user's query this is called problem of theme drift.

Cheating of pages- Some site owners insert fake links to WebPages in their website to increase the page rank value of those pages that is called page cheating.

3. Various Improvements In Pagerank Algorithm

We have analyzed various improvements in page rank algorithm as follows:

A. Ratio rank : Enhancing the impact of In links and Out links[7]

Ratio Rank is improved version of Page Rank Algorithm. Ratio Rank again uses link structure or web structure mining for calculating rank value of page. That uses weight of out links, weight of in links and number of visits of links by users as the parameters for calculating ratio rank of any page.

The equation from which ratio rank value is calculated is as follows:

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

\where, as per weighted page rank algorithm [4]

$$W^{in}_{(v,u)} = \frac{I_u}{\sum_{p=R(v)} I_p} [4] \qquad (a)$$

$$W^{out}_{(v,u)} = \frac{O_u}{\sum_{p=R(v)} O_p} [4] \qquad (b)$$

In which *RR (u)* and *RR (v)* are Ratio Rank values of pages u and v respectively. $W^{in}_{(v,u)}$ and $W^{out}_{(v,u)}$ are weight of in links and weight of out links. d is the dampening factor. V_u is the no. of visit counts of link which points from v to u. *TL(V)*is the total no. of visit of all links present on pages v, *B(u)* are the pages which points to webpage u, x is the ratio of weight of in links and y is the ratio of weight of out links. From empirical results, values of x and y is set between 0 and 1. Where value of X is set always higher than value of y because weight of in links is more important than weight of out links from the empirical results.

The results retrieved by Ratio Rank algorithm are more relevant than other algorithms because Ratio Rank uses three parameters which are visits of links by users, weight of out links and weight of in links in the equation of calculating Ratio Rank value. As Ratio Rank uses link structure, the problem of theme drift still exists in Ratio Rank.

B. Enhanced-Ratio Rank: Enhancing Impact of In links and Out links [8]

Enhanced-Ratio Rank also consider ratio of weight of the in links and weight of out links and visit counts of links by users for calculation of the rank value of particular page. It checks which ratio gives the best result i.e. which ratio of in links weight and out links weight helps to give better relevancy of the web pages. New Enhanced algorithm is given as follows:

$$RR(u) = (1-d) + d\sum_{v\in B(u)} \frac{\left(V_u * .7 * W^{in}_{(v,u)} + .3 * W^{out}_{(v,u)}\right)RR(v)}{TL(v)}$$

It uses same parameters as Ratio Rank equation. As in equation 70 percent of the weight of in links and the 30 percent of the weight of the out links is being used because this gives better result as compare to other ratios. By using all three parameters for computing the page rank value of WebPages and taking the best ratio of weight of in links and out links gives the better relevancy of web pages. But the problem of theme drift (some link may not give the search results about the query) still exists in this algorithm.

C. Weighted Page Content Rank for Ordering Web Search Result[9]

In Search results by page rank algorithm, some links are not according to the user's query because Page Rank is equally distributed to outgoing links and it is based on the number of in links and out links. Weighted Page Rank algorithm provides important information about a given query by using the structure of the web but some pages are irrelevant to given query, it still receives the highest rank because it has many in links and many out links. And there is a less determination of the relevancy of the pages to the given query. To overcome these limitations Weighted Page Content Rank (WPCR) is proposed which uses both web structure mining and web content mining for providing efficient web search results.

The Weighted Page Content Rank algorithm is as follows:
`

Step 1: Calculation of Relevance:
a) Find all meaningful word strings of query Q (say N)
b) Find whether the N strings are occurring in page P or not?
Z= Sum of frequencies of all N strings.
c) S= Set of the maximum possible strings occurring in P.
d) X= Sum of frequencies of strings in S.
e) Content Weight (CW) = X/Z (c)
f) C= No. of query terms in P
G) D= No. of all query terms of Q while ignoring stop words.
h) Probability Weight (PW) = C/D
Step 2: Calculation of rank value:
a) Find all back links of P i.e. reference page list of page P. (say it B).
b)

$$PR(P) = (1-d) + d\left[\sum_{v\in B}^{n} \mathrm{PR(V)}\, W^{in}_{(p,v)} W^{out}_{(p,v)}\right](CW + PW)$$

c) Output *PR (P)* i.e. the Rank score.
There are two new parameters used in above equation which are very important to understand:

Probability Weight: It is the probability of the query terms in the web page. This factor is the ratio of the query terms present in the webpage and the total number of terms in the fired query.

Content Weight: It is the weight of content of the web page with respect to query terms. This is the ratio of the sum of frequencies of highest possible query strings in order and sum of frequencies of all query strings in order. The maximum possible strings are selected in such a way that all such strings represent a different logical combination of words.

This algorithm uses both web structure mining and web content mining techniques for calculating page rank value. This algorithm is aimed to improve the order of the pages in the result list so that the user may get the efficient, relevant and important pages in top of the list.

D. An Effective Content based Web Page Ranking Approach [10]

In the Effective Content based Web Page Ranking Approach traditional Page Rank algorithm is analyzed. Page Rank algorithm has the limitation that is the rank score of a web page is divided evenly over its out linked WebPages and because of this, pages that are not relevant to the user query may get the higher rank value. To overcome this limitation, the new algorithm is produced which is query dependent and based on the web structure mining and web content mining.

New improved Algorithm:

The content based web page ranking algorithm is given as:

1) Initially, give PAGE RANK of all web pages to be 1.
2) Calculate page page ranks of all pages by following formula:

$$\mathrm{PR(u)} = (1 - \mathrm{d}) + \mathrm{d}\sum_{v\in B}^{n} \mathrm{PR(V)\,.WL(v,u).Wc}$$

Where
PR (u) and *PR (v)* are the Page rank value of page *u* and *v* respectively, *B (u)* is the reference page list of page u i.e. set of pages that point to u; *Wc* is the content weight [9] of the web pages with respect to the query terms.
3) Repeat step 2 until values of two consecutive iterations match.
This improved algorithm provides better result as per user's query than traditional page rank algorithm.

4. Proposed Algorithm

In this paper new enhanced page ranking algorithm is presented which exploits hybrid approach for calculating page rank value as it uses both web structure mining as well as web content mining. In this algorithm the importance and relevance of the WebPages is calculated by taking into account weight of in links, weight of out links and number of visit to the link by users and by taking new parameter content weight of the web pages with respect to the query termsWc.

Input: $W_{(v,u)}^{in}$=Weight of in links of the page.
$W_{(v,u)}^{out}$ =Weight of out links of the page.
$TL(v)$ =Total number of visits of all links present on v.
Wc = Content weight of the web pages with respect to the query terms.

Begin :

Step 1: Take the link structure of the retrieved WebPages from the crawler.

Step 2: Obtain the web graph from the link structure of the retrieved WebPages.

Step 3: Give initial page rank value to the all WebPages as one.

Step 4: Using equation number (a), (b), Calculate the weights of in links and out links and also calculate total no. of visits of all links by using client side script.

Step 5: Calculate the Content weight from the equation (c).

Step 6: Apply the proposed algorithm as in following equation

$$\mathrm{RR(u)} = (1 - \mathrm{d}) + \mathrm{d} * Wc \textstyle\sum_{v\in B(u)} \frac{(V_u *.7* W_{(v,u)}^{in} +.3*W_{(v,u)}^{out})RR(v)}{TL(v)}$$

Where,
RR (u) and RR (v) are ranking of the WebPages u and v respectively.
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) is the pages which points to webpage u.
$W_{(v,u)}^{in}$ is the weight of in links of connecting page *v* and *u*.
$W_{(v,u)}^{out}$ is the weight of out links of connecting page *v* and *u*.
Wc is content weight of the web pages with respect to the query terms.
Step 7 : Iteratively repeat process until ranks of all WebPages are stable i.e. same in two consecutive iteration.

This algorithm reduce the problem of theme drift which is present on every link structure based algorithms as it uses the new parameter content weight Wc from web content mining. Wc parameter takes user's query in to account and because of this the results retrieved are efficient and relevant as per user's query.

5. Experimental Analysis

For experimental purpose, we have taken four pages in our database which are crawled by crawler and the figure shows the web graph and links among the four web pages:

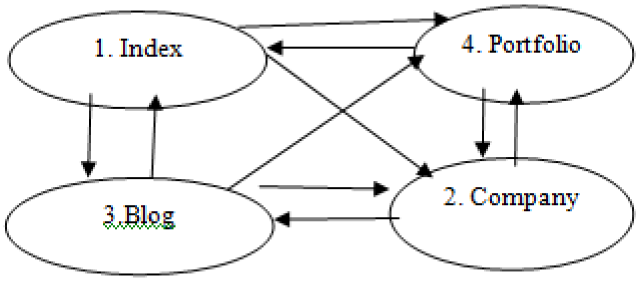

Figure 2: Sample Web Graph

We have applied page rank algorithm[1], enhanced ratio rank algorithm[2] and our proposed algorithm on this web graph and search one sentence. All four pages are retrieved in search results but according to algorithms the rank value of each page are different and hence the page which is top of the search results by page rank algorithm is different by enhanced ratio rank algorithm which is shown in below table:

Table 1: Comparison between the rank values by page rank, enhanced ratio rank and our proposed algorithm

Id	*WebPages*	*Pagerank Algorithm*	*Enhanced Ratio Rank Algorithm*	*Our Proposed algorithm*
1	Index	1.1239	0.1631	0.1680
2	Company	0.8758	0.2107	0.3586
3	Blog	1.1239	0.1579	0.1619
4	Portfolio	0.8758	0.2034	0.2155

The graph in the figure 3 shows the comparison of Pagerank Algorithm, Enhanced Ratio rank Algorithm and our Proposed Algorithm. As seen in graph, in our proposed algorithm rank values of pages are higher than enhanced ratio rank this is because of content weight parameter in our proposed algorithm. Content weight parameter compares user's query into account and compare it with contents in the WebPages. If it yields maximum comparison, content weight parameter value of that page will be higher and hence the rank value of that page will increase.

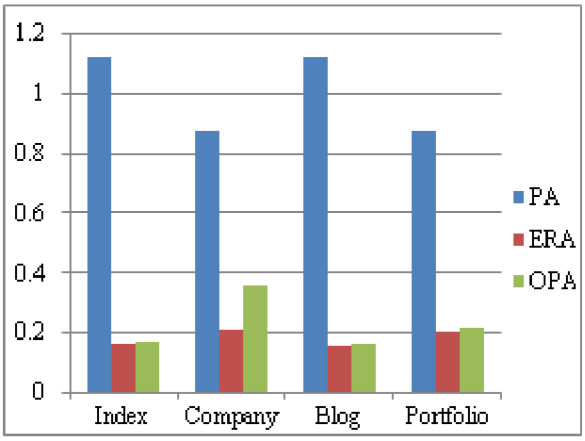

Figure 3: Ranking value comparison table

So we get efficient, relevant search results as per user's query.

6. Conclusion

In this paper we have analyzed various page ranking algorithms for getting efficient and relevant search results as per user's query. We have implemented basic page rank algorithm and one another algorithm that is enhanced ratio rank algorithm. Then we have understood that both the algorithms have the main challenge of theme drift. In our proposed algorithm we use hybrid approach as we take web structure mining and web content mining both for calculating page rank values of WebPages. After comparing all three algorithms we conclude that the Enhanced ratio rank provides the better results than the standard page ranking algorithm in terms of the better relevancy and ranking the non visited WebPages on the basis of the out link weights. By hybriding content weight parameter and enhanced ratio rank equation in our proposed equation provides more efficient and relevant search results as per user's query than page rank and enhanced ratio rank algorithm because when we will get maximum comparison strings in one webpage as per query by user the content weight parameter will increased and will increase the rank value of that page and we will get best relevant results as per user's query.

7. Acknowledgement

I would like to express my deep sense of gratitude to my guide, Asst Prof. Vinit Kumar Gupta for his valuable guidance and useful suggestions.

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] Wei Huang and Bin Li, "An Improved Method for the Computation of PageRank" *International Conference on Mechatronic Science, Electric Engineering and Computer* August 19-22, 2011, Jilin, China.

[6] Zhou Cailan and Chen Kai,Li Shasha," Improved PageRank Algorithm Based on Feedback of User Clicks" *IEEE* 2011.

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

[8] Ranveer Singh and Dilip Kumar Sharma," Enhanced-RATIORANK: Enhancing Impact of Inlinks and Outlinks" *IEEE Conference on Information and Communication Technologies* 2013.

[9] Pooja Sharma, Deepak Tyagi and Pawan Bhadana,"Weighted Page Content Rank For Ordering Web Search Result", *International Journal of Engineering Science & Technology* 2010, Vol. 2(12), PP. 7301-7310.

[10] Nidhi Shalya, Shashwat Shukla and Deepak Arora," An Effective Content Based Web Page Ranking Approach", *International Journal of Engineering Science and Technology (IJEST)* , Vol. 4 No.08 August 2012.