International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Improvement in PageRank Algorithm by Efficient Calculation of PageRank Algorithm

Yash Joshi[1], Hiteishi Diwanji[2]

[1,2] L. D. College of Engineering, Ahmedabad, India

Abstract: ***World Wide Web is the major source of information and collection of millions of web pages. The information on the world wide web is growing rapidly day by day and millions of pages are added, deleted and updated every day in the web. the world is full of questions and the web is serving as the major source for gaining the useful information that user wants. Users searches for topics based on their interest. The big issue that user faces during accessing the web is to retrieve the relevant result based on user query from the millions of web pages available in search results. In the modern era, it is striving to have a search engine which can give better and relevant results based on the query that user have searched. In this paper the pagerank algorithm is analyzed and improved by collecting user feedback and using web structure mining and web content mining through the parameters weights of clicks, weights of clicks time, pagerank of page and content weight. The result of existing algorithm is compared with proposed algorithm which shows that proposed algorithm gives better results based on user query than existing algorithm.***

Keywords: page rank, user feedback, Clicks Weights, Weights of Clicks Time, Content Weight

1. Introduction

As World Wide Web is developing rapidly, Internet has become the world's richest and most dense source of information. Users face the problem which is how to get relevant and useful information from the large number of disorder information. In today's era there is lots of information contained by World Wide Web and people use various available search engines to search relevant data of their need. Based on interest user can query data and get relevant information and based on the searched results any one can judge that the search engine is powerful or not.

Google is more powerful search engine in today's era and gives relevant information based on user interest. When user searches for a particular topic and enter the query then search engine gives results in sorting manner. Google uses pagerank. Algorithm which gives relevant results based on user query, and gives results in sorting manner and at the top of the searched results the information is more relevant to the user query.

Pagerank is the hyperlink based algorithm based on web structure mining uses inlinks and outlinks of web pages to calculate the pagerank value of a particular page. And based on the pagerank values the searched results are sorted and displayed to users.

Pagerank algorithm simply ranks the web pages and gives the relevant result to user based on the query. Search engines must gives relevant results based on user query so user can get information of their need and satisfied with results.

Web Mining [3] is used to extract useful information from users past behaviour and it includes Web Structure Mining, Web Content Mining and Web Usage Mining. And is It uses both Web Structure Mining [4] and Web Content Mining [5][6][7]. Web structure mining is used to generate the structural summary of the website and web pages. Web structure mining is used to extract patterns from hyperlinks from the web. Another kind of web structure mining is mining document structure. It uses tree like structure to describe HTML and XML documents.

Web Content Mining is the process of extracting the useful information from different contents of web documents. Web content mining is related to text mining and data mining but different from both of them. It is related to data mining because many data mining techniques can be applied to the web content mining.

2. Page Rank Algorithm

Page Rank Algorithm [1] is developed by S. Brin and L. Page the cofounder of Google's pagerank algorithm. Page rank is nothing but numeric value of a particular page which shows how important the page is. The more value of the pagerank the more important the page is and hence in sorting of results that will be at the top of the searched results. Pagerank algorithm is hyperlink based algorithm which calculates the pagerank value of any page based on inlinks and outlinks of a particular page. If there are more links pointing to a particular page the more important page is and it has higher pagerank value than other pages and if there is important hyperlink of the page pointing to the another page then that page is also important and hence pagerank value of that page is also high.

As per [2] the working of Pagerank algorithm. Web consists of hyperlinked structure and page ranking algorithm uses that structure to rank the web pages. If there are more important inlinks to a particular page then that page is also more important.

The basic expression of page rank algorithm is given as [2]:

$$PR(a) = (1 - d) + d\left[\frac{PR(t1)}{C(t1)} + \ldots + \frac{PR(tn)}{C(tn)}\right]$$

Where,

PR (a) is the PageRank value of page a.

$C(t1)$ is the number of hyperlinks point out from t1.

t1,t2, … , tn are pages pointing to page a.

d is the dampening factor and the value of which is lies between 0 to 1 and generally that will be taken as 0.85.

Dampening factor d is the probability that a user will follow the particular link and $(1 - d)$ is for non direct links to any web page. In the page ranking algorithm the ranking of web pages is done based on the importance of web pages. And the ranking of pages is done at the indexing time not at the retrieval time. If web page with no outlinks found then the user directly goes to chosen bookmark at random.

3. Improvement of Pagerank Algorithm

To get more relevant results based on user query, in this paper there are mainly three parameters used named Clicks weight, weight of Clicks Time and Content Weight.

1. Clicks Weight[2]
   If user clicks on any links then that will indicates that the content on that page meets user requirements. If there are more number of clicks on that page the more page is able to meet the user requirements. Number of click of a particular page for a certain time period are to be collected and clicks weight [2] based on the formula below will be calculated,
   $$S(a) = \alpha \ln(N + 1) + \alpha_0$$
   But when the new page arrives then it has no user clicks and hence some compensation must be given to that page so here α_0 is used to given compensation to that page. So α_0 is the Compensation coefficient and Value of α_0 is generally set to 0.3. α_0 reflects the importance attached to the new page. α is the attenuation coefficient which controls the Weight of Clicks and N is the number of clicks in particular time period. Here PR and S are in positive correlation.
2. Weight of Clicks Time[2]
   For in certain time period the content of a particular page is modified then weight of click time [2] is to be taken as 1 else that will be taken as $1 + \beta T$ as in equation given below.
   $$T(a) = \begin{cases} 1, T \leq 1 \\ 1 + \beta T, T > 1 \end{cases}$$
   In which $T = \begin{cases} T_{now} - T_{last}, T_{last} \neq NULL \\ T_{now} - T_{update}, T_{last} = NULL \end{cases}$

   Where,
   T_{update} and T_{last} are the modify time of particular page a.
   T indicates the time interval of particular page a. T is the difference between time of user real time search and the time when page was updated or created. If the clicks of page are 0 replace last click time as update or generate time. Here β is the attenuation coefficient and is generally set to 1/12. Here PR and T are in negative correlation. The content of any page must be updated within a certain time period to maintain the pagerank value or increase the pagerank value.
3. Content Weight[3]
   It is the weight of content of the web page with respect to query terms. This factor 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. Content weight [3] of a particular page is calculated based on user query. When user enters any query then the content weight of that particular web page based on the query entered by user will be calculated by the formula given below,
   $$W_c = X/Z$$

Where,
W_c is the Content Weight
X is Sum of Frequencies of Strings in S.
Z is Sum of Frequencies of all N Strings.
S is Set of the maximum possible strings occurring in page P.

4. Proposed Algorithm

Input : S(a) = clicks weight
T(a) = weight clicks time
W_c = Content Weight
PR(a) = PageRank of page a
Begin:
Step 1: Take the link structure of the retrieved webpages from the crawler.
Step 2: Obtain the webgraph from the link structure of the retrieved webpages.
Step 3: Using equation number (a), (b), Calculate the clicks weight and clicks time.
Step 4: Calculate the Content weight from the equation (c).
Step 5: Apply the proposed algorithm as in following equation:

$$PR(A) = PR(a) \times \frac{S(a)}{T(a)} \times W_c \ldots [A]$$

Where,
PR(a) old pagerank of page A .
PR(A) new pagerank of page A.
S(a) is the clicks weight.
T(a) is the weight clicks time.
Wc is content weight of the web pages with respect to the query terms.
Step 6: Iteratively repeat process until ranks of all web pages are stable means same in two consecutive iteration.
Here, we can see that the PR(a) and S(a) are in positive correlation with each other.
And PR(a) and T(a) are in negative correlation with each other.

5. Experimental Analysis

In the Experimental Analysis We have taken 5 web pages as shown in below figure named Index, Blog, Contact, Company and Portfolio as our Test Bed and based on the parameters inlinks, outlinks, weights of inlinks, weight clicks time and content weight we can calculates the Improved PageRank value of each web page based on the given equation [A]. The table given below shows the 5 web pages of test bed and useful parameters of each page is to be

calculated here and the results of Improved PageRank Algorithm are compared with the basic PageRank Algorithm which shows that the search result is improved to some extent. As shown in the below figure is the test bed which consists 5 web pages and inlinks and outlinks of each page are as shown in the figure, based on the inlinks and outlinks the PageRank value of each page is calculated and based on that the Improved PageRank value is calculated.

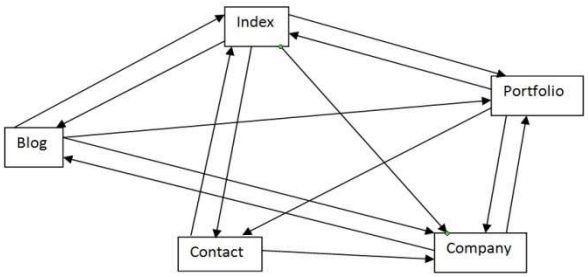

Figure 1: Test Bed of 5 Pages

Page_Name	*PageRank*	*No.of Clicks*	*S(i)*
Index	0.6359	7	1.6238
Blog	0.5380	16	1.8499
Contact	0.4065	6	1.5837
Company	0.5101	25	1.9774
Portfolio	0.6113	11	1.7454

T	*T(i)*	*Improved PR*	*Content Weight*	*Final Rank*
1	1	1.0325	2.3333	2.4093
1	1	0.9952	1	0.9952
1	1	0.6338	1	0.6438
1	1	1.0086	2	2.0173
1	1	1.0670	1	1.0670

The table shows the Comparison between pagerank and improved pagerank algorithm. Here Blog page has higher pagerank then Company page, but blog page has 16 clicks and content weight is 1 and company page has 25 clicks and content weight is 2 so that the improved pagerank of Company page is be higher than the Blog page.

6. Conclusion

Here in this paper, we have analyzed Improved PageRank Algorithm and PageRank Algorithm. The Parameters Clicks weight, weight of Clicks Time and content weight are to be calculated and are used to calculate the Improved PageRank. The Experimental Analysis done on the test bed which consists of 5 pages and inlinks and outlinks of each pages are used to calculate the PageRank algorithm, And the Comparison between PageRank and Improved PageRank values shows that the search result of Improved PageRank Algorithm is Improved from basic PageRank Algorithm.

7. Acknowledgement

I am deeply indebted and would like to express my gratitude to my guide Prof. Hiteishi M. Diwanji (Associate Professor, L. D. College of Engineering), for her great efforts and instructive comments in my work and give me 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] Zhou Cailan, Chen Kai,Li Shasha "Improved PageRank Algorithm based on feedback of user clicks" IEEE 2011
[3] Pooja Sharma, Deepak Tyagi, Pawan Bhadana "Weighted Page Content Rank for Ordering Web Search Result" International Journal of Engineering Science and Technology Vol. 2(12), 2010, 7301-7310
[4] S. Chakrabarti, B.E. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg "Mining the Web's link structure". Computer, 32(8):60-67, 1999
[5] R. Kosala and H. Blockeel. "Web mining research": A survey. ACM SIGKDD Explorations, 2(1):1-15, 2000.
[6] Raymond Kosala, Hendrik Blockeel, "Web Mining Research": A survey, ACM SIGKDD Explorations Newsletter, June 2000, Volume 2 Issue 1.
[7] Wang Jicheng, Huang Yuan, Wu Gangshan, Zhang Fuyan. Web mining: knowledge discovery on the Web Systems, man, and Cybernetics, 1999. IEEE SMC '99 Conference Proceedings. 1999 IEEE International Conference.