An Improvement on Page Ranking Based on Visits of Links

Shweta Agarwal[1], Bharat Bhushan Agarwal[2]

[1] Moradabad Institute of Technology, Mahamaya Technical University
Ram Ganga Vihar, Phase-2 Moradabad-244001, Uttar Pradesh, INDIA

[2] IFTM University, School of Engineering and Technology
Lodhipur Rajput, Delhi Road (NH-24) Moradabad-244102, Uttar Pradesh, INDIA

Abstract: ***Web search can simply be considered as a process of user enters the query and search system returns a set of most relevant pages based on the query. But results returned are not mostly relevant to user's query and ranking of the pages are not efficient according to user requirement. In order to improve the precision of ranking of the web pages, after analyzing the page rank and its various versions, we proposed one more factor "the total time spent to read the web pages" to be included in our algorithm that signifies the importance of a web page for a user and thus helps in increasing the accuracy of web page ranking.***

Keywords: Inlink, Outlink, Page Rank, Visits of links**.**

1. Introduction

Day by day the information keeps piling on in this massive web structure. Hence, it becomes necessary to structure this diverse and dynamic unstructured storage of data. For the purpose mentioned it is important to understand and analyze the underlying data structure of web for effective and efficient information extraction with the increasing demand of users. So it has become necessary for the search engines to give most specific and user need satisfying results. There are lot of search engines but few like Google, Yahoo, etc. are famous because of their crawling and ranking methodology. Every day they solve and satisfy millions of queries. So, Ranking methodology becomes a very important aspect of web mining in all the three components of search engine (i.e. Crawler, Indexer, Ranking mechanism). Figure 1 shows the concept of Search engines, which are used to find information from the WWW. They download, index and store hundreds of millions of web pages. They answer thousands of queries every day. They act like content collector as they keep record of all the information available on WWW [1].

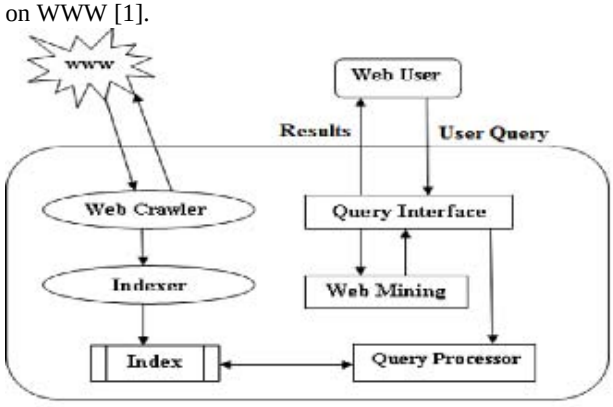

Figure 1: Search Engine Architecture

In web search, ranking algorithms play an important role in ranking web pages so that the user could get the good result which is more relevant to the user's query.

The structure of this paper is as follows. Section 2, we present the literature survey and related work of ranking algorithms based on link structure will be discuss in section 3. Then, section 4 defines the problem definition and to yield more accurate results, our improvement method for the computation of page rank will be discussed in section 5. In section 6, results are shown and finally, we conclude our work in section 7 and some future work will be discussed in section 8.

2. Literature Survey

Web mining is a data mining technique used to extract information from World Wide Web [2]. According to analysis targets, it is classified into three basic categories –

i) **Web Content Mining:** - It is the process of extracting useful information from the contents of web documents. This mining technique is used on the web documents and results page that are obtained from a search engine.
ii) **Web Structure Mining:** - It is the processes of discovering link structure of the hyperlinks in inter documents level from the web. It is used in many application areas.
iii) **Web Usage Mining:** - It is the process to discover interesting usage patterns from web data in order to understand and better serve the needs of web-based applications.

3. Previous Work

With the rising demand of information on web, search engines have to adopt various techniques to prioritize web pages. It is a great deal of work to rank pages such that it gives user most appropriate results according to its requirement. To make it happen various algorithms have been designed and introduced with different perspective. Some algorithms use link structure of web pages whereas other use content to define relevancy of web pages to user

queries. Here are some ranking algorithms discussed with their varying nature of web mining category, working, and input parameters etc.

3.1 Page Ranking Algorithm

Page Rank was developed at Stanford University by Larry Page and Sergey Brin in 1996. It is based on the concepts that if a page contains important links towards it then the links of this page towards the other page are also to be considered as important pages. It considers the back link in deciding the rank score. If the addition of all the ranks of the back links is large then the page it is provided has large rank. A simplified version of Page Rank is given below:

$$PR(u) = \sum_{v \varepsilon B(u)} PR(v) / N(v) \quad (1)$$

where, u represents a web page, B (u) is the set of pages that point to u, PR (u) and PR (v) are rank scores of page u and v respectively, N(v) indicates the number of outgoing links of page v.

In Page Rank, the rank of page p, is evenly divided among its outgoing links. Later Page Rank was modified observing that not all users follow the direct links on WWW. Therefore, it provides a more advanced way to compute the importance or relevance of a web page than simply counting the number of pages that are linking it. If a backlink comes from an important page, then that backlink is given a higher weighting than those backlinks comes from non-important pages. Thus, the modified version is given as-

$$PR(P)=(1-d)+d(PR(T1)/C(T1)+.....+PR(Tn)/C(Tn)) \quad (2)$$

where, d is a damping factor which set its value to 0.85. d can be thought of as the probability of users following the links and could regard (1 − d) as the page rank distribution from non-directly linked pages. We assume several pages T1…..Tn which point to it i.e., are links. PR(T1) is the incoming link to page A and C(T1) is the outgoing link from page T1 (such as PR(T1)).

3.2 Weighted Page Ranking Algorithm

Weighted Page Rank algorithm was proposed by Wenpu Xing and Ali Ghorbani. This algorithm takes into account the importance of both the inlinks and outlinks of the pages and distributes rank scores based on the popularity of the pages[3].

The popularity from the number of inlinks and outlinks can be calculated as $W_{in}(v,u)$ and $W_{out}(v,u)$ respectively. $W_{in}(v,u)$ is the weight of link(v, u) which is calculated based on the number of inlinks of page u and the number of inlinks of all reference pages of page v.

$$W_{in}(v,u) = Iu / \sum_{p \varepsilon R(v)} Ip \quad (3)$$

where, Iu and Ip represents number of inlinks of page u and page p respectively, R (v) denotes the reference page list of page v.

$W_{out}(v,u)$ given in eq. (4) is the weight of link(v, u) which can be calculated based on the number of outlinks of page u and the number of outlinks of all reference pages of page v.

$$W_{out}(v,u) = Ou / \sum_{p \varepsilon R(v)} Op \quad (4)$$

where, Ou and Op represents the number of outlinks of page u and p, respectively. R(v) represents the reference page list of page v.

Considering the importance of pages, the original PageRank formula is modified below -

$$PR(u) = (1-d) + d / \sum_{v \varepsilon B(u)} PR(v) \; Win(v,u) \; Wout(v,u) \quad (5)$$

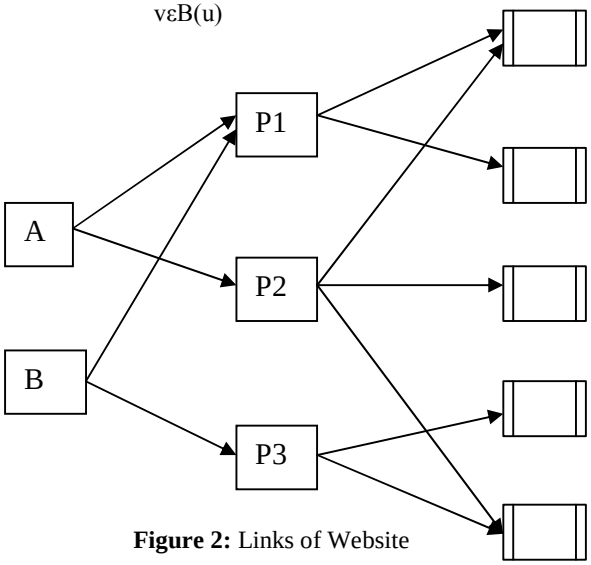

Figure 2: Links of Website

3.3 Page Rank Based on VOL

We have seen that original Page Rank algorithm, the rank score of a page p, is equally divided among its outgoing links or we can say for a page, an inbound links brings rank value from base page, p(rank value of page p divided by number of links on that page). Here, an improved Page Rank algorithm has been proposed by Gyanendra in which more rank value is assigned to the outgoing links which is most visited by users [4]. In this manner a page rank value is calculate based on visits of inbound links. The modified version based on VOL is given as-

$$PR(u) = (1-d) + d \sum_{v \varepsilon B(u)} Lu \; (PR(v))/ TL(v) \quad (6)$$

where, Lu is the total number of link's visits which is pointing page u from v, TL(v) denotes total number of visits of all links present on v.

4. Problem Definition

With the tremendous growth and increasing demand of

information on web it has become quite necessary to satisfy the user's demand upto the level of his/her expectation. User always expects to get the most relevant results, which, with such complex structure and varying queries becomes hard to provide for a Search Engine. Hence different Ranking algorithms like Page Rank (PR), WPR (Weighted Page Rank), Page Rank based on VOL algorithms are used in different Search Engines to deal with such problems but fails to focus on the user query preference, therefore finding the content of the Web and retrieving the user's interests and needs from their behavior is a crucial factor.

5. Approach

To display the more target oriented pages at the top of the search list, we propose a new approach which focuses on the user query preference, where consideration is done on the most useful or important pages. To determine the useful pages, we take time spent on reading a document by a user as an essential factor which decides the importance of a page. Reading time is the time spends by a user in reading a page, which we suppose reflects the usefulness of information in the page as conceived by the user.

This proposed approach will compute the web page rank according to visits of links of inbound links as well as personal attention given to the web page. This algorithm behaves completely different from traditional page ranking algorithm which always return the same web page rank for the same query submitted by different users at different time despite the user's interest in the page may vary or change.

$$PR(u)=((1-d)/N)+\{(d \sum_{v \varepsilon B[u]} L_u(PR(v))/TL(v))\}RT(u) \quad (7)$$

where,
d is the dampening factor,
u and v represents the web pages,
B[u] is the set of pages that points to page u,
PR(u) and PR(v) are the page ranks of page u and page v respectively,
L_u is the total of visits of link which is pointing page u from v,
TL(v) represents total number of visits of all links present on v.
RT(u) is the maximum of the time that user's take to read a page u

5.1 How to Calculate the Reading time of a web page

To calculate the reading time of a web page, a client side script is used. Whenever a web page is accessed the script will be loaded on the client side from web server. Script will monitor the click, keyboard as well as cursor moving event to occur. When an event occur and if that event will happen over hyperlink then it will send a message to web server with information of current web page and hyperlink and at the same time it starts counting the reading time of a web page and sends the measured reading time record and user identification number to the server side.

On the server side, the database of log file will be used to record the web page id, hyperlinks of that page, hit count of hyperlinks and reading time of the link page. This file will accessed by crawler at the time of crawling. This crawled information will be stored in search engine's database which is used to calculate the rank value of different web pages.

To avoid the large value set of reading time as well as to less complicate our calculation, the value of reading time which will be sent to the server will be compared with the last updated time value (if exists), if the new time value will be larger than already existing then the existing one will get replaced with the new value in the log file.

6. Results And Discussion

To simulate it, let us take the example of hyperlink structure that consists of four pages A,B,C and D with maximum time spent by user on a page is mentioned in seconds along with number of visits on each link as shown in figure below -

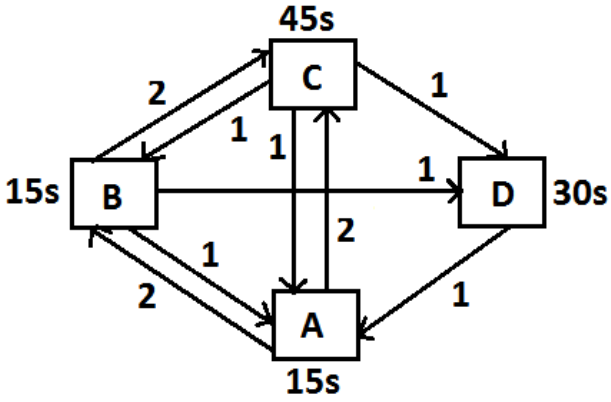

Figure 3: Web link structure with time factor

Let us assume the initial page rank of all pages as 1 and the value of damping factor d is 0.85. The page rank for pages A, B, C and D can be calculated by the following equations:

PR(A) = ((1-d)/N) + d*RT(A)*((PR(B) * LA/TL(B)) + (PR(C) * LA /TL(C)) + (PR(D) * LA /TL(D)))

PR(B) = ((1-d)/N) + d*RT(B)*((PR(A)*LB/TL(A)) + (PR(C)*LB/TL(C)))

PR(C) = ((1-d)/N) + d*RT(C)*((PR(A)*LC/TL(A)) + (PR(B)*LC /TL(B)))

PR(D) = ((1-d)/N) + d*RT(D)*((PR(B)* LD /TL(B) + PR(C)*LD/TL(C))

Now, by using the above formulas, the page rank will be calculated as mention in Table 1.

Table 1: Iterative calculation for the proposed algorithm

Iterations	*A*	*B*	*C*	*D*
1	1	1	1	1
2	0.373958	0.148066	0.0203895	0.082117
3	0.077259	0.060151	0.081299	0.055408
4	0.058229	0.049445	0.071821	0.052928
5	0.056461	0.048586	0.070984	0.052718
6	0.056312	0.048511	0.070912	0.052700
7	0.056299	0.048505	0.070906	0.052699
8	0.056298	0.048504	0.070906	0.052699
9	0.056298	0.048504	0.070906	0.052699

According to the results shown in Table 1, the pages will be display in order-

Page C > Page A > Page D > Page B

Here, page C will get higher rank that means preference is given to that page which is found useful for a user or on which user spends most of the time.

By applying various algorithms on the web graph shown in Figure 3, we get various variations in page rank which is shown in figure below:

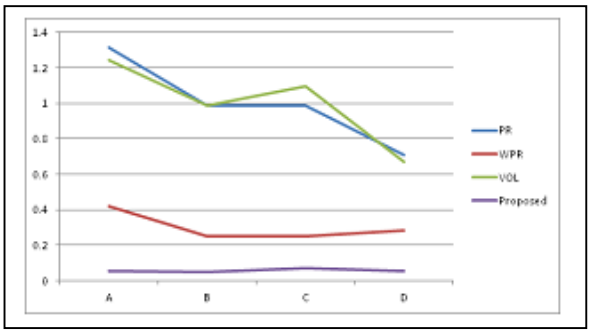

Figure 4: Variations of various page ranking algorithms

7. Conclusion

Web Mining is the data mining technique that automatically discovers or extracts the information from web document. Page Rank, Weighted Page Rank and Visits of Links based Page Rank are used in web structure mining to rank the relevant pages. In this paper, we focused that these algorithms may not get the required relevant document easily. To solve this problem, we utilize the time factor to increase the accuracy of web page ranking. The results of this algorithm are very satisfactory and are in agreement with the applied theory for developing the algorithm.

8. Future Work

In future, we are planning to carry out performance analysis of our proposed algorithm and working on finding required relevant and important pages more easily and fastly. We can also include the concept of captcha on click of each web page url, which avoids the use of automated machines as well as robots which may help the web page to increase its rank by keep browsing the page through cursor.

REFERENCES

[1] Neelam Duhan, A. K. Sharma, Komal Kumar Bhatia, "Page Ranking Algorithms: A Survey" Advance Computing Conference, 2009. IACC 2009 IEEE International.

[2] Tamanna Bhatia," Link Analysis Algorithms For Web Mining ", IJCST Vol. 2, Issue 2, June 2011.

[3] Wenpu Xing and Ali Ghorbani, Weighted PageRank Algorithm, Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR '04), IEEE, 2004

[4] http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=6075206&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D6075206

Author Profile

Shweta Agarwal completed the B.Tech degree in Computer Science and Engineering from College of Engineering and Technology Moradabad in 2008. During 2008-2010, she worked in SeagalSOFT, Noida as a Web Developer and then she joined MIT as an Assistant Professor in February 2010. Now, she is working with the same institute.

Bharat Bhushan Agarwal completed his B.Tech degree in Computer Science and Engineering from Moradabad Institute of Technology, Moradabad in 2003. Post, he did M.Tech from Uttar Pradesh Technical University Noida in year 2008. During his M.Tech he joined CET Moradabad as a Lecturer and then joined IFTM University Moradabad in 2012 as an Assistant Professor. He is also pursuing his Phd from Teerthankar Mahaveer University Moradabad.