Reading Time: A Method for Improving the Ranking Scores of Web Pages

Shweta Agarwal
Asst. Prof., CS&IT Deptt.
MIT, Moradabad, U.P. India

Bharat Bhushan Agarwal
Asst. Prof., CS&IT Deptt.
IFTM, Moradabad, U.P. India

ABSTRACT
In order to improve the ranking of web pages, we analyze the original PageRank and its improved versions; and then record the reading time and visited links of the page to check the degree of importance to the users. It means we include the reading time as a time factor as well to improve the precision of the ranking, which can be treated as the combination of usage and link structure in another way.

Keywords
Search Engine, Reading Time, Visits of links, PageRank, Outlink, Inlink.

1. INTRODUCTION
With the rapid development of world-wide web, Internet has become the world's richest and most dense source of information. The users face the problem in getting relevant and useful information from the large number of disorder information. Search engines become an important tool for users to retrieve information for a particular query. However, current search engines can't fully satisfy the user's need; and raises many new challenges for information retrieval. Web mining [1] came into being in this environment, and quickly became research focus in the field of the data mining and information retrieval. Currently, the most classic Web structure mining algorithm is PageRank algorithm[2] that Sergey Brin and Larry Page proposed at Stanford University[3], in order to verify the performance of the algorithm, they successfully applied it to the Google search engine prototype, and now Google has become the world's most well-known search engine. However, PageRank algorithm is simply to start from the perspective of hyperlinks-based structural analysis and completely ignoring the other factors, thus it is difficult to achieve the better results in some time. Therefore, interest is growing in personalized search engines[4] that need to infer user search preferences. The typical sequence of user behaviour in a web search session is highlighted below –

1. Submit a search query by typing in a few keywords.
2. Wait until the webpage rank list is returned by the search engine.
3. Scan the title and/or summary of each document, which is usually provided in the returned search result page.
4. Click on the links to the documents that the user is interested in, which might be several.
5. Wait until the desired page(s) are loaded.
6. Browse/read the loaded page(s).
7. After looking through all the opened pages, the user may click on more links in the webpage rank list to request more web pages or submit a new query using other keywords if the initial search results do not serve his search interest well.

In this paper, we focus on step 6, which can provide visiting time that refers to the time a user will spend on reading a document, which we suppose can reflect the usefulness of the information in the document as conceived by the user. We propose an algorithm to compute user-oriented webpage ranks according to personal visiting time. This new algorithm behaves differently from conventional search engines which always return the same webpage rank for the same query submitted at different times or by different users despite that user's interest in the page might vary or change.

2. BACKGROUND
In this section, we'll introduce the background of our research. A short review of PageRank algorithm[5] is in section 2.1.

2.1 Page Rank Algorithm
Page Rank was developed at Stanford University by Larry Page[7] and Sergey Brin[6] in 1996. It is one of the method that Google (one of the famous search engine) uses to determine the importance or relevance of a web page. It's one of the major factor is that it is used to determine which pages appear in search results. This algorithm is the most commonly used algorithm to determine the appearance of web page in search result ranking.

The basic idea of Page Rank is that a page is considered important if many other important pages link to it. So in this concept, the rank of a page is given by the rank of those pages which link to it. Hence, the Page Rank of a document is always determined recursively by the Page Rank of other documents. To rank web pages with their popularity, this algorithm uses number of pages that points to it, also known as indegree algorithm (since it ranks web pages according to their indegree). So, page rank provides a better approach that can compute the importance of web page by simply counting the number of pages that are linking to it. These type of links are called as backlinks. If this type of link comes from an important page then this link has higher weightage than those which are coming from non-important pages. The link from one page to another is considered as a vote. Not only the total number of votes that a page receives is important but the

relevancy and popularity of pages that casts the vote is also important.

2.2 Working

Working of the Page Rank algorithm depends upon link structure of the web pages. The Page Rank algorithm 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. The Page Rank gave importance to the back link in deciding the rank score. In Page Rank, the rank score of a page, p, is equally divided among its outgoing links. The values assigned to the outgoing links of page p are in turn used to calculate the ranks of the pages to which page p is pointing.

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 to it. If backlink comes from an important page, then that backlink is given a higher weighting than those backlinks comes from non-important pages. In a simple way, link from one page to another page may be considered as a vote. However, not only the number of votes a page receives is considered important, but the importance or the relevance of the ones that cast these votes as well. Thus, the modified version is given in eq (1)

$$PR(P) = (1-d) + d\left(\frac{PR(T1)}{C(T1)} + \ldots\ldots\frac{PR(Tn)}{C(Tn)}\right)\ldots\ldots\ldots(1)$$

where, d is a damping factor which can be set between 0 and 1, we usually 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)).

The Page Rank forms a probability distribution over the web pages so the sum of Page Ranks of all web pages will be one. The Page Rank of a page can be calculated without knowing the final value of Page Rank of other pages. Page Rank of a page depends on the number of pages pointing to a page.

2.3 Example illustrating the working of PageRank

To illustrate the working of original PageRank algorithm, let's take an example of hyperlinked structure (see Figure1).

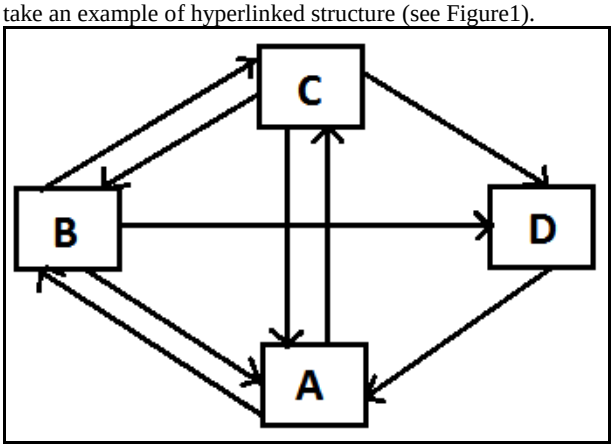

Figure 1: A simple web graph

We take a small web graph consisting of four pages A,B,C and D, where page A links to page B and C, page B links to page A,C and D, page C links to page A,B and D. Also, page D links to page A.

2.4 Calculation of PageRank Algorithm

Assume the initial page rank of all the pages as 1 and page rank for pages A, B, C and D can be calculated by using eq (1) as follows:

PR(A) = (1-d) + d(PR(B)/C(B) + PR(C)/C(C) + PR(D)/C(D))

= (1-0.85) + 0.85(1/3 + 1/3 + 1/1)

= 1.566667

PR(B) = (1-d) + d(PR(A)/C(A) + PR(C)/C(C))

= (1-0.85) + 0.85(1.566667/2 + 1/3)

= 1.0991667

PR(C) = (1-d) + d(PR(A)/C(A) + PR(B)/C(B))

= (1-0.85) + 0.85(1.566667/2 + 1.0991667/3)

= 1.127264

PR(D) = (1-d) + d(PR(C)/C(C) + PR(B)/C(B))

= (1-0.85) + 0.85(1.127264/3 + 1.0991667/3)

= 0.780822

In the same way, further iterations will be calculated until two consecutive iterations have the same page rank value.

2.5 Advantages of PageRank algorithm

The advantages of PageRank algorithm are as follows:

1. This algorithm is robust against span since its not easy for a webpage owner to add inlinks to his/her page from other important pages.
2. PageRank is a global measure and is query independent.
3. PageRank algorithm is more feasible in today's scenario since it performs computations at crawl time rather than query time.

2.6 Drawbacks of the existing algorithm

The Page Rank algorithm is being used by Google. It has some limitations which are given below:

1. The major drawback of Page Rank is that it favors the older pages, because a new page, even a very good one will not have many links unless it is a part of an existing site.
2. It is purely based on the concept of backlinks.
3. It is query-independent and cannot by itself distinguish between pages that are authoritative in general and pages that are authoritative on the query topic.
4. As people know the secrets to obtaining a higher Page Rank, the data can be manipulated. For example, Google Bombs, Link Farming etc.

3. READING TIME BASED PAGE RANKING ALGORITHM

The importance of pages is different to users even though their link structure is same because their content is different. On the basis of interest of user, the importance of a page is determined. If the page is interested by the user, the reading time of that page will also be larger than the pages don't accord with user's interest. It means the content of that page is most probably the user want to search. Hence to improve the

precision of ranking score of pages, we proposed an algorithm which is an improvement on the existing page rank algorithm. Its main aim is to rank those pages at higher position that are most likely by the user. This can be done by including the reading time as a time factor into the computation of the ranking algorithm. Time factor of a page shows how much a user like a page.

3.1 To Calculate Reading Time of a Web Page

We calculate the reading time of a web page based on client side script. When a user clicks on a webpage, the script will be loaded on the client side from web server and starts counting time the user spends on a web page. Also, script will monitor the cursor movement event of mouse as well as key press event of keyboard to occur. When an event occurs, it will start counting the time in seconds and when the web page is closed, script will send a message to the web server with information about the reading time of current web page and hyperlink.

On server side, a database of log file will be used to record the web page id, hyperlinks of that page, hit count of hyperlinks which is incremented every time a hit occurs on hyperlink and visiting time of that page which starts calculated between when a user opens a web page and until closes it.

The database or log files will accessed by crawler at the time of crawling. This hit count and time information will be stored in search engine's database which is used to calculate the rank value of different web pages or document.

It calculates rank value of a web page based on the user's visits on incoming links of that page and total time that a user spend in reading a document. The ordering of pages in this way increases the relevancy of pages and thereof provides the user with quality search results. As a result, user may find the desired content in the top few pages, thus search space can be reduced to a large scale.

3.2 Working of Proposed Algorithm

The working of this proposed algorithm not only considers link structure but also includes the user's focus on a particular page. The improved version of Page Ranking algorithm is given in equation 2.

$$PR[u] = \left(\frac{1-d}{N}\right) + \left\{\left(\frac{d\sum_{v\varepsilon B[u]} Lu\,(PR[v])}{TL[v]}\right)\right\} RT[u] \quad \ldots(2)$$

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 v respectively,
Lu is the total of visits of link which is pointing page u from page 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,
N is the total number of web pages.

3.3 Example illustrating the working of Reading time based PageRank Algorithm

To illustrate the working of proposed PageRank algorithm, let us take an example of hyperlinked structure shown in figure 2.

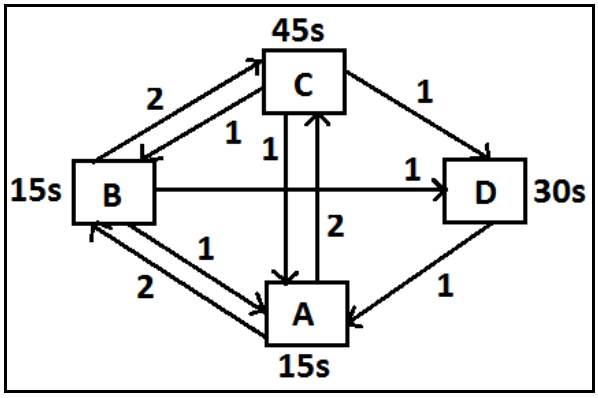

Figure 2: A web graph with reading time

We regard a small web graph consisting of four pages A,B,C and D, where page A links to page B and C, page B links to page A,C and D, page C links to page A,B and D. Also, page D links to page A and each link has its corresponding visits. And each web page has its reading time in seconds.

3.4 Calculation of Proposed Algorithm

Assume the initial page rank of all the pages as 1 and page rank for pages A, B, C and D can be calculated by using eq (2) as follows:

PR(A) = ((1-d)/N) + d*RT(A)*((PR(B) * LA/TL(B)) + (PR(C) * LA /TL(C)) + (PR(D) * LA /TL(D)))
= (1-0.85)/4 + 0.85*0.25*(1*1/4 + 1*1/3 + 1*1/1)
= 0.373958

PR(B) = ((1-d)/N) + d*RT(B)*(PR(A)*LB/TL(A) + PR(C)*LB/TL(C))
= (1-0.85)/4 + 0.85*0.25*(0.373958*2/4 + 1*1/3)
= 0.148066

PR(C) = ((1-d)/N) + d*RT(C)* (PR(A)*LC/TL(A) + PR(B)*LC /TL(B))
= (1-0.85)/4 + 0.85*0.75*(0.373958 * 2/4 + 0.148066 * 2/4)
= 0.0203895

PR(D) = ((1-d)/N) + d*RT(D)* (PR(B)* LD /TL(B) + PR(C)*LD/TL(C))
= (1-0.85)/4 + 0.85*0.5*(0.148066 * 1/4 + 0.0203895 * 1/3)
= 0.082117

In the same way, further iterations will be calculated until two consecutive iterations have the same page rank value.

3.5 Advantages

Some advantages of the proposed algorithm are as follows:-

1. In this algorithm, a user can not intentionally increase the rank of a web page by visiting it for more time or multiple times, because the rank also depends upon probability of visits of inlinked pages.

2. The rank of any page by using the page rank algorithm will be same either it is submitted by different users at different time despite the user's interest in the page may vary or change because it is totally dependent on web link structure of the web graph. While the ordering of pages using visting time is more target-oriented.

3. As visiting time method uses link structure of pages and their browsing behavior based interest, the top returned pages in the result list are supposed to be highly relevant to the user information needs.

4. Updating the visited time value in a database will avoid the large value set as well as less complicate the calculation done in computing the page rank.

4. EXPERIMENTAL RESULTS

Experiments are performed on four hyperlink web pages that are pageRank, Backlink, Seo and Keywords, their page rank values have gone through various iterations until two consecutive iterations have the same page rank value.
Table1 shows various page rank values by applying PageRank algorithm and the page rank values of four web pages by applying our proposed Reading Time based page ranking algorithm is shown in table2. We combine these two tables and shows the final page rank values of the two algorithm in table3.

We have also implemented a search engine and displays the search result based on the query entered and the ranking algorithm used. When the user enters the word "search" to be searched, the results returned by the search engine as per PageRank algorithm is shown in figure 3. In such results, user needs to traverse the result pages for finding the relevant one according to their need. The results produced by the proposed Reading Time based Page Ranking algorithm is shown in figure 4, which is much efficient as it takes user browsing behavior in consideration.

The total number of iterations recorded for the simulation of traditional PageRank and the proposed reading time based page ranking algorithm to reach a convergence value is tabulated in table 4. Since the number of iterations for calculating the page rank values in the proposed reading time based page ranking algorithm are reduced.
Hence, the time complexity of the proposed algorithm is less as compared to the conventional pageRank algorithm.

Table 1. Page rank values for different web pages as per Page Rank Algorithm

Iterations	pageRank	Backlink	Seo	Keywords
1	1	1	1	1
2	1.566667	1.099167	1.127264	0.780822
3	1.444521	1.083313	1.070860	0.760349
4	1.406646	1.051235	1.045674	0.744124
5	1.376630	1.031342	1.027281	0.733277
6	1.356562	1.017602	1.014859	0.725864
7	1.342848	1.008254	1.006382	0.721814
8	1.333505	1.001881	1.000606	0.717372
9	1.327137	0.997538	0.996669	0.715026
10	1.322797	0.994578	0.993986	0.713427
11	1.319839	0.992561	0.992157	0.712337
12	1.317823	0.991186	0.990911	0.711594
13	1.316449	0.990249	0.990061	0.711088
14	1.315513	0.989610	0.989483	0.710743
…….	……..	……..	……..	……..
…….	……..	……..	……..	……..
20	1.313709	0.988380	0.988368	0.710079
21	1.313645	0.988337	0.988328	0.710055
22	1.313602	0.988307	0.988301	0.710039
23	1.313572	0.988828	0.988283	0.710028
24	1.313552	0.988273	0.988270	0.710021
25	1.313538	0.988264	0.988262	0.710015
26	1.313529	0.988257	0.988256	0.710012
27	1.313522	0.988253	0.988252	0.710010
28	1.313518	0.988250	0.988249	0.710008
29	1.313515	0.988248	0.988247	0.710007
30	1.313513	0.988246	0.988246	0.710006
31	1.313511	0.988245	0.988245	0.710006
32	1.313511	0.988245	0.988245	0.710005
33	1.313510	0.988244	0.988244	0.710005
34	1.313509	0.988244	0.988244	0.710005
35	1.313509	0.988244	0.988244	0.710005

Table 2. Page rank values for different web pages as per Reading Time based Page Rank Algorithm

Iterations	pageRank	Backlink	Seo	Keywords
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

Table 3. Final values of PageRank and Reading Time based Page Rank Algorithm

Web Pages	PageRank Algorithm	Reading Time Based Page Ranking Algorithm
pageRank	1.313509	0.056298
Backlink	0.988244	0.048504
seo	0.988244	0.070906
keywords	0.710005	0.052699

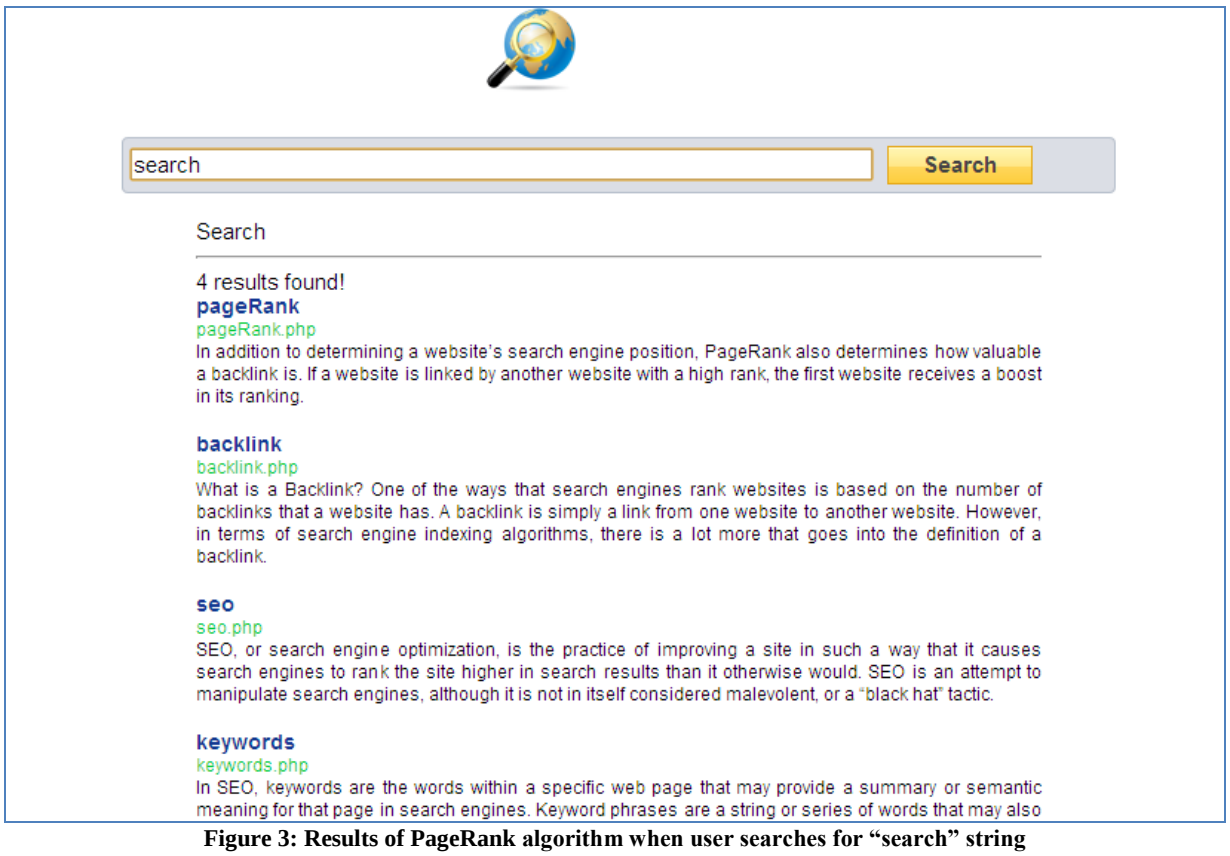

Figure 3: Results of PageRank algorithm when user searches for "search" string

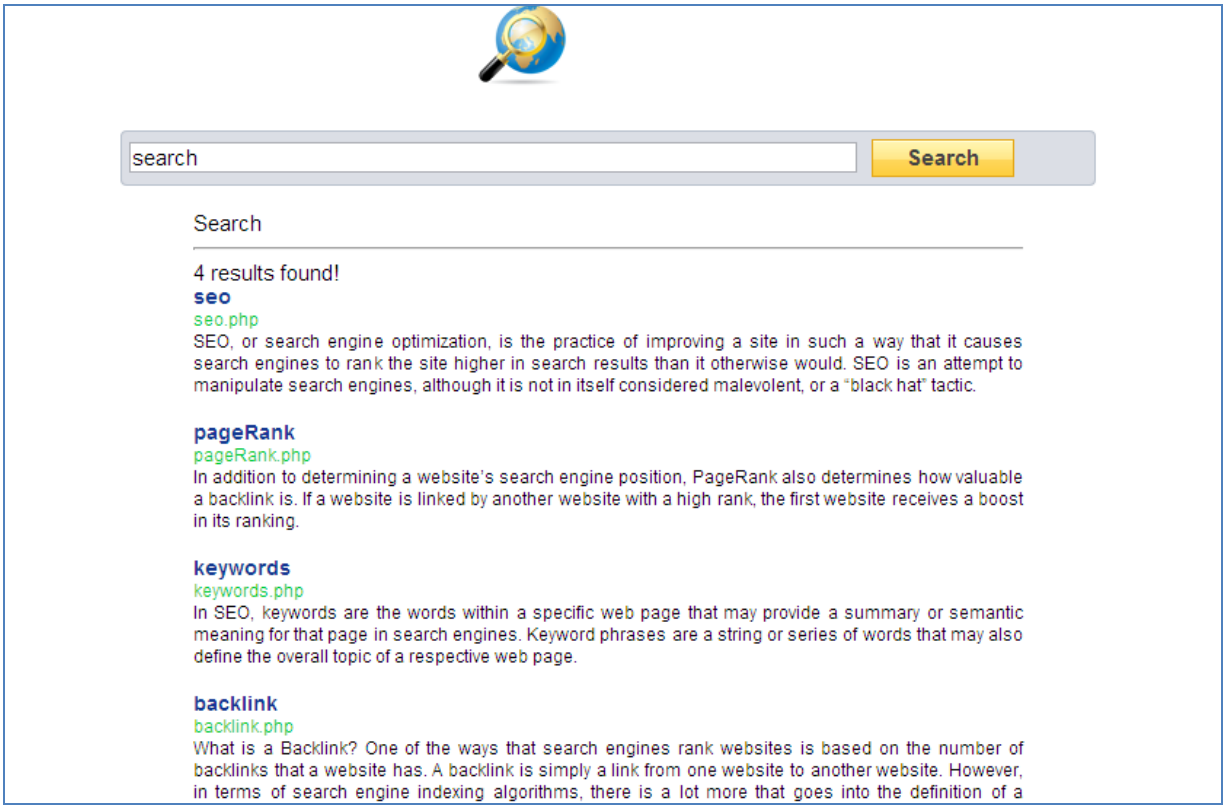

Figure 4: Result of proposed Reading Time based algorithm when user searches for "search" string

Table 3. Number of iterations performed by the conventional pageRank and proposed pageRank algorithm.

No. of iterations for conventional PageRank Algorithm	No. of iterations for proposed Reading Time based PageRank Algorithm
35	**9**

5. COMPARISON BETWEEN PAGERANK AND READING TIME BASED PAGE RANK ALGORITHM

Table 4 below enlists the comparison of PageRank Algorithm and Reading Time based Page Ranking Algorithm.

Table 4. Comparison of PageRank and Reading Time Based Algorithm

Criteria	Page Rank	Reading Time Based PR Algorithm
Mining Technique Used	Web Structure Mining	Web Structure Mining, Web Usage Mining
Methodology Used	Rank is calculated on the basis of backlinks	Rank is evaluated on the basis of number of visits of inbound links as well as maximum time taken by a user to read a page.
Input Parameter	Backlinks	Server Log
Working Process	Those having more number of backlinks will get higher page rank.	More time user spend on a web page, more important the page is assumed to be.
Importance	Backlinks are considered	Consideration of the maximum time given to the pages.
Relevancy	Less	High due to the involvement of reading time

Quality of Results	Less	Higher than PR based on VOL
Advantages	It provide important information about given query by dividing rank value equally among its outlink pages.	Useful when two pages have the same link structure but different contents.
Limitations	It favours older pages, because a new page, even a very good one, will not have many links unless it is part of an existing site.	Periodic crawling of web servers.
Search Engine	Used in Google	Used in Research Model

6. CONCLUSION

In this paper, we have presented a modified page ranking algorithm which is more target oriented as compare to traditional page ranking algorithm. This modified algorithm calculates page rank based on user's interest, because of which the importance of a page is determined. It is not only considers the link structure, it includes user's focus on a particular page. But the main problem arises in this concept is the calculation of visiting time and visits of links, for which we have already mentioned the simple concept to monitor the visiting time and counting the visits. User usually spends a lot of time in surfing through the search results to find the pages of their interest. The paper presented an approach which is based on visiting time and link which increases the relevancy of pages and provides the user with quality search results that makes user to find the desired content in top few pages.

7. FUTURE DIRECTIONS

As part of future work, performance analysis of the proposed algorithm can be carried out on large database and working can also be done in finding required relevant and important pages more easily and fastly on large data set.

To avoid the limitation of proposed algorithm that a user can intentionally increase the reading time of a web page, the concept of time value on keyboard or cursor event can be included.

Also, the concept of captcha can be included on click of each web page url, which avoids the use of automated machines as well as robots that may help the web page to increase its rank by keep browsing the page through cursor.

8. ACKNOWLEDGEMENTS

Our thanks to the anonymous persons for their helpful comments and suggestions that have improved the quality of this paper.

9. REFERENCES

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

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

[3] Google PageRank – Algorithm available at http://pr.efactory.de/e-pagerank-algorithm.shtml.

[4] Dou, Z.; Song, R.; and Wen, J.-R. 2007. A large-scale evaluation and analysis of personalized search strategies. In WWW '07: Proc. of the 16th International Conference onWorld WideWeb, 581–590. New York, NY, USA: ACM.

[5] Brin S and Page L (1998), The Anatomy of a Large-Scale Hypertextual Web Search Engine_, Computer Networks and ISDN Systems, Vol. 30, Nos. 1-7, pp. 107-117.

[6] Brin, S., Motwani, R., Page, L., & Wingrad, T. (1998). What can you do with a web in your pocket? Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 21(2), 37 – 47. Retrieved November 4, 2002, from http://www.n3labs.com/pdf/brin98what.pdf.

[7] L. Page, S. Brin, R. Motwani, and T. Wino grad, "The PageRank Citation Ranking: Bringing order to the Web". Technical report, Stanford Digital Libraries SIDL-WP-1999-0120, 1999.