# Improved Time Aware Incremental PageRanking Using Personalized Link Structure Analysis

[1]Shail K Dinkar, [2]Hemant Kumar

[1]Department of Computer Application,Govind Ballabh Pant Engineering College,Pauri Garhwal-246194,India

[2]Department of Mechanical Engineering, Govind Ballabh Pant Engineering College,Pauri Garhwal-246194,India

## Abstract

This paper devises time-aware and Link Analysis based improvement pageranking technique. The time-aware techniques exploit temporal information, present in networks like the World Wide Web, to produce rankings reflecting authority with regard to a temporal interest. The link prediction analysis techniques produce rankings based on the relative change with regard to a temporal interest. The time-aware methods extend PageRank and are defined incrementally. The link prediction based methods are defined independently, one extending PageRank and the other based on a comparison of precomputed page rankings. The method proposed in this paper suggests an improvement in time-aware PageRanking method by giving information of temporal interest of a particular user on a specific web page.So,this method is useful for finding time spent by a user on a particular web page which suggests user's temporal behavior and link prediction for a particular webpage by that user.

## Keywords

PageRank, Time Aware Incremental PageRank, Temporal Interest,Link Structure, Webgraph

## 1.Introduction

The two predominant paradigms for finding information on the Web are navigation and search[1] . Most Web users typically use a Web browser to navigate a Web site. They start with the home page or a Web page found through a search engine or linked from another Web site, and then follow the hyperlinks they think relevant in the starting page and the subsequent pages, until they have found the desired information in one or more pages.

The importance of web pages can be computed by PageRank method.This methos is used by the Google Web search engine.Two views[9] have been used for interpreting the PageRank method and its score (i) stochastic i.e. random surfer: The steady-state distribution of a Markov chain is termed as PageRank values, and (ii) algebraic: the Page Rank values form the eigenvalue 1 to the corresponding eigenvector of the Web link matrix. The Interaction Information Retrieval ($I^2R$) method[9] is a non-classical information retrieval paradigm, which represents a connectionist approach based on dynamic systems. In the present paper, a different interpretation of PageRank is proposed, namely, a dynamic systems viewpoint, by showing that the PageRank method can be formally interpreted as a particular case of the Interaction information Retrieval method.

On the other hand, contents of Web pages, extended anchor texts, hyperlinks between Web pages, and Web usage data of a Web site are rich sources of data for mining knowledge about the Web site and its users. The knowledge can be used to assist users to navigate the Web site and search for desired information more effectively and efficiently. By viewing Web pages as nodes and hyperlinks between them as directed edges between nodes, we can construct a link structure of a Web site. Contents of the Web pages and extended anchor texts are properties of these nodes and hyperlinks. Hyperlinks convey conceptual relationships between Web pages. User traversals on hyperlinks can be extracted from Web usage data and used as properties of the hyperlinks. The constructed *link structure* therefore contains information about Website contents, hyperlinks, and user behavior.

The analysis of the link structure shows about the user behavior and the amount of time spent by a user on a specific web page. This temporal interest shown by a user which changes according to the time spent by a user affects the importance of a web page[5]. There is importance of structural navigation in the new generation of Web browsers[2]. Current algorithms for ranking Web pages[3,4] such as PageRank and HITS consider only information about hyperlinks.Another approach which is based on personalized link structure analysis and time spent by a user on a specified web page is also introduced[5].This approach used time constraint in improved version of original PageRank algorithm. In this paper an improved approach of Time Aware incremental pageRank is proposed which provides the temporal information of a particular user including time spent by that user on a specific webpage.

## 2.PageRank

The PageRank algorithm was intuitively justified to model a random surfer in which a user clicks on links at random and the rank of a page signifies the probability of a user arriving at that page. A user can arrive at a page either by clicking on the links or by randomly jumping to a page. The algorithm includes a parameter d, which represents the probability of a user continuing to click on links and (1–d) as the probability that the user jumps to a random page. PageRank of a page is determined using the random surfer model described

above. The PageRank of a page A can be computed[6,7,8] as follows:

$$PR(A) = \frac{1-d}{N} + d * \sum_{foreveryj \in S} PR(j) / L(j) \qquad (1)$$

where,
PR(j) PageRank of page j.

S Set of nodes that have a link to page A.

L(j) No. of out going links of page j.

d Damping factor that is set to a value between 0 and 1. It is usually set to 0.85 for the web graph.

1-d probability that the user jumps to a random page.

N is the number of nodes in the graph.

In PageRank,the rank value of a page A is evenly distributed among its outgoing links.These values are then used to calculate the PageRank of the pages to which page A is pointing.Within a website,there is a possibility that two or more pages may be connected with each other to form a loop.These pages can be referred by other web pages but do not distribute the PageRank values.This is called a sink problem[3].

The PageRank of all the pages can be computed using an iterative algorithm.Each page is given an initial value and the PageRank of all the pages is calculated in several iterations[10] as shown in example

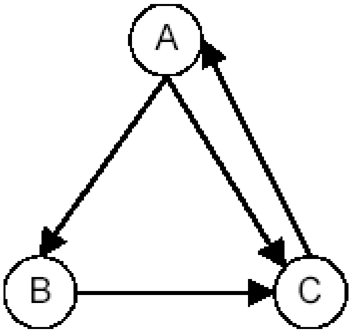


Figure 2.1 : Link Structure for PageRank Calculation

In Figure 2.1, a link structure consisting of three pages *A* , *B* and *C* , where page *A* is linked to pages *B* and *C* page *B* is linked to page *C* ,and page *C* is linked to page *A* . Suppose the damping factor *d* is 0.5. We get the following linear equations for the PageRank calculation

PR(A )= 0.5/3+ 0.5 PR(C )
PR(B ) =0.5/3 +0.5 (PR (A )/2)
PR(C )= 0.5/3+ 0.5 (PR (A )/2+PR (B ))
We can solve these equations to get the PageRanks of the pages as follows

PR(A )= 14/39=0.39
PR (B)=10/39=0.26
PR(C)=15/39=0.38

## 3. Proposed Approach

An improvement has been proposed[5]in pagerank algorithm given by Brin and Page[3].This improvement is based upon the calculation of time spent by a user on a specified web page.It suggested that pagerank of a page is dependent on time because any modification in a page content may change the no. of total links(incoming and outgoing) on a page which leads to change in the time spent by a user.

In this paper, another extension to the improved PageRank method [5] is proposed. The proposed approach is based on personalization of rankings in such a way that for every user (or a group of users) a ranking is computed with respect to the user's interests. This temporal interest can be predicted by link structure analysis.This analysis is capable of showing the time spent by a user on a specified page. In a practical implementation, this user's interest could be derived from a collection of bookmarks or from the user's surfing history.

## 4. Improved Incremental PageRank Algorithm

### 4.1 Methodolgy

Every link pointing to a Web page is associated with a timestamp which corresponds to the time when the referencing Web page was last modified.In this case, personalized link structure analysis is made to find out the users' behavior and then time spent by a user on a particular page by assigning a time stamp to each page so that the age of page can be found. This age of page will be calculated in two ways:

(i) Time spent by a user on a particular page will give the age of that page.

(ii) When page is modified, the no. of links may also change.

In this case, the time spent by a user will be taken after the modification.

Since the number of outgoing links and incoming links may change due to the the change in temporal interest of a user, it will definitely lead to the changes in PageRank of a web page. So, it is necessary to find the PageRank of a page per unit time. During a search for a topic, the relevant pages are retrieved on the basis of their PageRank values. The existing algorithm can be modified having time constraint[5].

Hence,the pagerank of a page i per unit time can be calculated as follows[5]:

Do

[Start of loop]

$$PR(i)=\frac{1-d}{N}+d*\sum foreveryj\in S\,(PR(j)/L(j))/T(i)) \qquad (2)$$

increment the value of i by 1

[end of loop while i=n]

Where T(i) is the time spent on a page i by a user.

### 4.2 Improved Algorithm

Time Aware incremental pagerank algorithm can be improved by analyzing the structure of a link to find out the variation of time spent by a user . It may happen that a user visits a particular page on a **day X**(say pagerank value is x) for **duration t1** and the same user visits the same page on **day Y** for **duration t2**.Proposed algorithm proposes that due to the change in visiting time of this particular page will cause the change in its pagerank value(say y).

The proposed algorithm will calculate the pagerank of a web page due to the change in visiting time of the same user on same day which will help to predict the link structure and user behaviour for a particular web page. A new parameter to calculate the visiting time is induced in the proposed algorithm. The algorithm is as follows:

find T(i)

increment the value by 1

[end of loop while i=n]

Now the PageRank of page i per unit time can be calculated as follows

Do

[Start of loop]

$$PR(i)=\frac{1-d}{N}+d*\sum foreveryj\in S\,(PR(j)/L(j))/T(i))+D \qquad (3)$$

increment the value of i by 1

[end of loop while i=n]

Where

$$D=\frac{\left|T(i)_{previous}-T(i)_{current}\right|}{H*M*S} \qquad (4)$$

and

H=24 hours.

M=60 minutes.

S=60 seconds.

Here, the parameter D is used to find the time spent by a user on a particular page in a day. Let us suppose that a user X spends time T1 on a page A at any point of time on a particular date. After some time, it may possible that the same user X again visits the same page A for a time T2.

In a time aware incremental pageranking, the total time spent by a user will have an impact the ranking value of a page. This impact factor can be evaluated by use of parameter D.The inclusion of this parameter will be able to analyze the personalized link structure which in turn will reflect the temporal behavior of a user.

## 5. Experimental Analysis

### 5.1 Simulation Parameter

Simulation of the proposed algorithm is done by developing a tool in ASP.NET in front end and SQL Server in back end. The simulation parameters are shown in table 1.

**Table 1**. Simulation Parameter

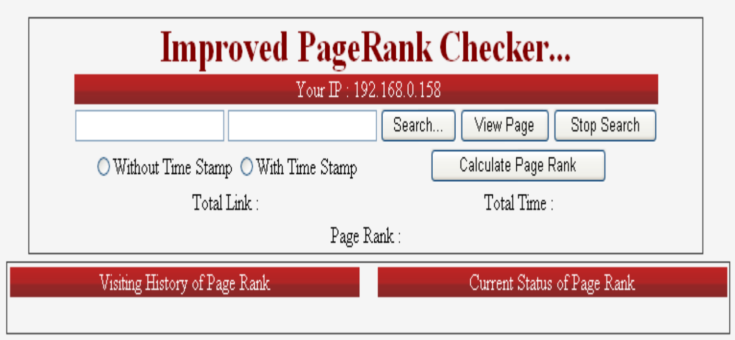
| Sr. No. | Parameter | Meaning |
|---|---|---|
| 1. | PR(i) | PageRank of page i |
| 2. | D | Damping Factor(0.85) |
| 3. | N | No. of nodes(pages) in web graph |
| 4. | S | Set of pages having link to page i |
| 5. | PR(j) | PageRank of page j having link to page i |
| 6. | L(j) | No. of out link from page i |
| 7. | T(i) | Time spent on page i |
| 8. | D | Difference between previous and current visiting time for a da |
| 9. | H | Hours in a day |
| 10. | M | Minutes per hour |
| 11. | S | Seconds per minute |

### 5.2 Evaluation of Time Aware Improved Algorithm

The implementation of this algorithm gives not only Time aware incremental pageranking but also reflects the temporal behavior of a user through personalized link structure analysis.

Personalized link structure is based first to identify the user.A user can be identified by his IP address given or used over the internet.When a user visit a page A,this simulator displays the IP address of that user. If page A is visited already by the same IP Address, simulator gives the new page rank according to the current visiting time with date. If a page is first time visited then simulator will calculate the pagerank first time according to the visiting time. On the other hand, if page is already visited by the same user then the current time is added with the previously visiting time and pagerank is calculated according to the proposed algorithm shown above.

In addition to the implementation of Time aware incremental pageranking algorithm,the normal view of display screen shows the IP address of user.The display screen also includes the grid view of Visiting history of the pagerank and current status of pagerank(see figure 5.2.1).



Improved PageRank Checker...

Your IP : 192.168.0.158

Search...
View Page
Stop Search
○ Without Time Stamp ○ With Time Stamp
Calculate Page Rank

Total Link :

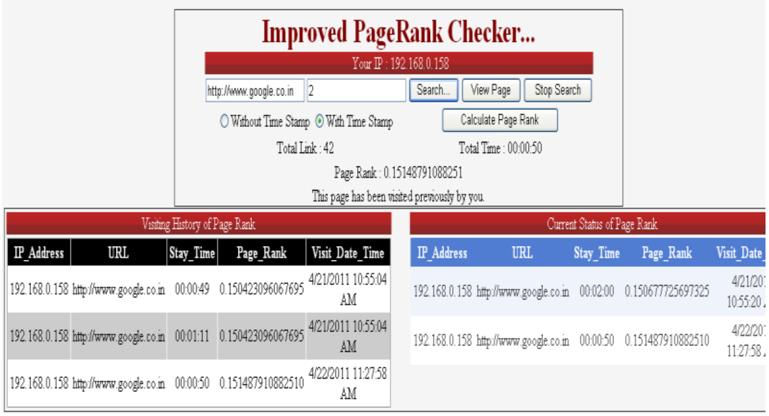
Total Time :

Page Rank :

Visiting History of Page Rank
Current Status of Page Rank

Figure 5.2.1: Normal View to Calculate PageRank with IP address

Calculating pagerank with time stamp gives pagerank on the basis of the no. of links visited per unit time(see figure 5.2.2).



Improved PageRank Checker...

Your IP : 192.168.0.158

http://www.google.co.in 2

Total Link : 42

Total Time : 00:00:50

Page Rank : 0.15148791088251

This page has been visited previously by you.

Visiting History of Page Rank

| IP_Address | URL | Stay_Time | Page_Rank | Visit_Date_Time |
|---|---|---|---|---|
| 192.168.0.158 | http://www.google.co.in | 00:00:49 | 0.150423096067695 | 4/21/2011 10:55:04 AM |
| 192.168.0.158 | http://www.google.co.in | 00:01:11 | 0.150423096067695 | 4/21/2011 10:55:04 AM |
| 192.168.0.158 | http://www.google.co.in | 00:00:50 | 0.151487910882510 | 4/22/2011 11:27:58 AM |

Current Status of Page Rank

| IP_Address | URL | Stay_Time | Page_Rank | Visit_Date_ |
|---|---|---|---|---|
| 192.168.0.158 | http://www.google.co.in | 00:02:00 | 0.150677725697325 | 4/21/20<br>10:55:20 |
| 192.168.0.158 | http://www.google.co.in | 00:00:50 | 0.151487910882510 | 4/22/20<br>11:27:58 |

Figure 5.2.2. Time Aware PageRanking with User's temporal interest history

So this algorithm is capable of:

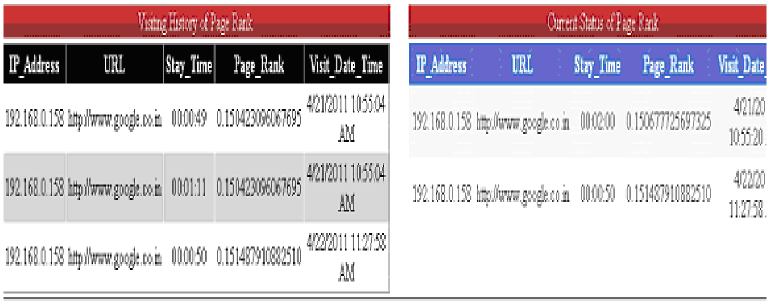
(i) To reflect the change in PageRank value of a page due to change in visiting time of a user.

(ii) It provides the history of a user reflecting how much time a user visits a particular link revealing his/her temporal interest.

## 6. Comparison With Improved Algorithm

Comparing with Time Aware Incremental PageRanking with Personalized Link structure analysis algorithm, suppose a user with IP address 192.168.0.158 visits a page www.google.co.in for 00:00:49 on dated 21-04-2011,then the pagerank calculated with improved algorithm comes to be 0.15042.After some time,the same page www.google.co.in is again visited by the same user with IP address 192.168.0.158 for 00:01:11.So the total time visited by the user becomes 00:02:00 which affects the PageRank value and it becomes 0.150677 which is different from the previous on(see table 2).

Table 2:Time Aware PageRanking with User's temporal interest history



Visiting History of Page Rank

| IP_Address | URL | Stay_Time | Page_Rank | Visit_Date_Time |
|---|---|---|---|---|
| 192.168.0.158 | http://www.google.co.in | 00:00:49 | 0.150423096067695 | 4/21/2011 10:55:04 AM |
| 192.168.0.158 | http://www.google.co.in | 00:01:11 | 0.150423096067695 | 4/21/2011 10:55:04 AM |
| 192.168.0.158 | http://www.google.co.in | 00:00:50 | 0.151487910882510 | 4/22/2011 11:27:58 AM |

Current Status of Page Rank

| IP_Address | URL | Stay_Time | Page_Rank | Visit_Date |
|---|---|---|---|---|
| 192.168.0.158 | http://www.google.co.in | 00:02:00 | 0.150677725697325 | 4/21/20<br>10:55:20 |
| 192.168.0.158 | http://www.google.co.in | 00:00:50 | 0.151487910882510 | 4/22/20<br>11:27:58 |

proposed algorithm with time constraint, the page rank value per unit time comes which provides more accurate and efficient result.

The PageRank values for www.google.co.in and www.ipu.ac.in is calculated using the tool based on proposed algorithm and result set for both pages are compared.

Further more, If page A is older than page B with same weight and content and the no. of links to page A and B are also same. Then the calculated page rank of B will be higher than page rank of A since the number of links earned by page B are more in less time than page A. Hence, this improved algorithm gives more accuracy in page rank calculation by including the constraint of age of page.

## 7. Conclusion and Future Work

The evaluation of result with the proposed algorithm can be justified taking time into consideration which shows more appropriate result than the previous one. The PageRank of a newly registered web page on world wide web can be found incrementally. As the age of this newly crawled page will be less than an older page, the number of incoming links pointing to this page will also be less. The number of incoming links to an older page will be more in spite of the importance of that page like content because it may possible that a newly crawled page may also be more or equally important than an older one. So analyzing the link structure for an individual user and finding the earned incoming link per unit time will provide more accurate and appropriate result

Furthemore,it may happen that a particular page is visited by a particular user frequenty for a period of time and after completion of that duration, the visiting time may be reduced. So the number of links per unit time will also be decreased thus affecting the PageRank value of that page. This algorithm will provide appropriate results in these situation also.

For finding users' temporal interest, link structure analysis is done in which an evaluation parameter is induced for link prediction. The temporal dimension of the users' interest at the different points of time, a links has different weights so that the approximation of these parameters is being induced in proposed algorithm. However, proposed algorithm has certain aspects to be done in future.

- The parameter to predict the temporal interest and calculating the time is taken for only 24 hours.
- Suppose a user visits a web page at 11:59:00 PM in night for the first time, then pagerank of that page will be calculated for the first time. As the date changes after 12:00:00 PM in night, the previous pagerank will not be considered and it will be calculated again.

## References

[1] Olston, C. and Chi, E. H. (2003) "ScentTrails: Integrating Browsing and Searching on the Web". *ACM Transactions on Computer-Human Interaction (TOCHI)*, Vol. 10, No. 3, pp. 177-197.
[2] Nielsen, J. (2000) „*Designing Web Usability*". New Riders Publishing, Indianapolis,Indiana, USA.
[3] L. Page,S. Brin,R. Motwani,and T. Winograd, "The PageRank citation ranking Bringing order to the web", Technical Report, Satnford Digital Library technologies SIDL-WP-1999-0120,1999.

[4] Kleinberg, J. M. (1998) "Authoritative Sources in a Hyperlinked Environment". In Proc. of Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, ACM Press, New York, pp. 668-677.
[5] Dinkar,Shail K.,Purwar,Ravindra Kumar(2011), "Time Aware Incremental PageRanking Based on Personalized Link Structure Analysis", International Journal of Research and Reviews in Computer Science,Vol. 2,No. 2,April 2011,pp. 383-388.
[6] Kim, S.J., & Lee, S.H. (2002). "An improved computation of the pagerank algorithm". In F. Crestani, M. Girolamo, & C.J. van Rijsbergen (Eds.), Proceedings of the European Colloquium on Information Retrieval (LNCS 2291. Pp. 73-85). London: Springer.
[7] Padmanabhan D, Desikan P, Srivastava J et al. (2005), "Web Intelligence", 2005 Proceedings_, The 2005 IEEE/WIC/ACM International Conference on Web Intelligence, Copiane University, France, September 19-22.
[8] Padmanabhan D, Desikan P, Srivastava J et al. (2005), "ICER: A Weighted Inter Cluster Edge Ranking for Clustered Graphs", Web

Intelligence 2005 Proceedings, The 2005 IEEE/WIC/ACM International Conference on Web Intelligence, Copiane University, France, September 19-22.

[9]Dominich, Sandor and Skrop,Adrienn(2005), “PageRank and interaction Information Retrieval”, Journal of The American Society For Information Science and Technology,56(1),pp. 63-69.

[10] Zhu, J., Hong, J. and Hughes, J. G., “PageCluster Mining Conceptual Link Hierarchies from Web Log Files for Adaptive Web Site Navigation” . ACM Transactions on Internet Technology (ACM TOIT), in press, 26 pages, 2003.

**First Author:** Mr. Shail K Dinkar is working as Assistant Professor at Department of Computer Application in Govind Ballabh Pant Engineering college,Pauri Garhwal,india.He has obtained MTech(Information Technology) from Guru Gobind Singh Indraprastha University,Delhi, Master of Computer Application from U P Tech University,Lucknow,India and Bachelor of Science from C.C.S. University,Meerut,India.He has more than nine years of teaching experience in the field of Computer Application and Information Technology. His research area includes Distributed Algorithm and Computation in Web Mining, Wireless and Adhoc Network and Image Processing. He has been actively involved in improving algorithm to enhance the rank value of a page. He has various papers presented and published in National/International conferences and journal of repute.

**Second Author:** Mr. Hemant Kumar is working as Assistant Professor at Department of Mechanical Engineering in Govind Ballabh Pant Engineering college,Pauri Garhwal,india. He has obtained BTech from Sant Longowal Institute of Engineering and Technology,Sangrur,Punja,India.He is pursuing MTech(Mech. Engg.) from NITTTR,Chandigarh,India.He has more than eleven years of teaching experience in the field of Mechnical Engineering. His research area includes Distributed Algorithm and Computation in Web Mining, Robotics and Welding Technology. He has various papers presented and published in National/International conferences and journal of repute.