Improve an Enhanced-Ratio Rank Algorithm based on Reading Time

Jinal V. Patel[1], Rimi Gupta[2]

[1]M.E. Student, Dept. of Computer Engineering, Sardar Vallabhbhai Patel Institute of Technology. Vasad, Gujarat, India

[2]Assistant Professor, Dept. of Computer Engineering, Sardar Vallabhbhai Patel Institute of Technology, Vasad, Gujarat, India

Abstract: ***Web mining is an active research area in the present scenario. Web mining is defined as the application of the data mining techniques on the World Wide Web for finding the hidden information. It can be classified into three categories: web content mining, web structure mining, and web usage mining. Page ranking is based on the web structure mining. Page rank is extensively used for rank the webpages using various page ranking algorithms like page ranking algorithm, weighted page ranking algorithm, enhanced-ratio rank algorithm, reading time based page ranking algorithm, page level keyword based algorithm, etc. By Comparison between all these algorithms, conclude that enhanced-ratio rank is better because it use inlinks, outlinks and also visit of links. And my proposed algorithm is adding an average reading time in enhanced ratio rank algorithm for the getting a better results and to reduce the limitation of existing algorithm. In this paper we add the results of enhanced-ratio rank and page ranking algorithm based on reading time and proposed algorithm that is improve an Enhanced-Ratio Rank Algorithm Based on Reading Time. Results shown that proposed algorithm gives the better result because it assign the rank based on the inlinks, outlinks, visit of links of the webpages and use the average reading time of the webpages. Here we use the different webpage on the Book Information and the average reading time of different users.***

Keywords: Web Mining, Page Rank, Weighted Page Rank, Enhanced-Ratio Rank, Reading Time Based Page Rank

1. Introduction

The Internet is the collection of large number of data that serves millions of users worldwide. It is a huge store of distributed documents. Internet is increasing day by day so there is a challenge for website owner to provide proper and relevant information to the internet user. But all of this information is not relevant to user. There are various challenges associated with the ranking of web pages such that some web pages are made only for navigation purpose and some pages of the web do not possess the quality of self-descriptiveness. For ranking of web pages, several algorithms are proposed including Page rank algorithm, Weighted Rank algorithm, Enhance-Ratio Rank algorithm. This Algorithm is Useful for Rank the pages. This first three algorithm is based on the links which is related to the webpages. In Page Rank Algorithm Backlinks are used as a input parameter, In Weighted Rank algorithm Backlinks and Forward links are consider as a input parameter, In Enhanced-Ratio rank algorithm Inlinks, Outlinks and VOI is use as a input parameter.

A. Web Mining

Web miningis the use of data mining techniques to automatically discover and extract information from Web data - including Web documents, hyperlinks between documents, usage logs of web sites, etc.It is the information service center for news, e-commerce, and advertisement, government, education, financial management, education, etc. [1]

We have developed Web mining framework for evaluating ecommerce web sites.

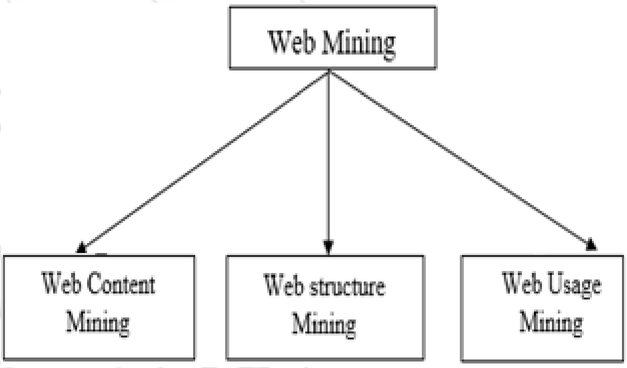

Figure 1: Web mining framework

B. Classification of Web Mining[3]

1) Web content mining
2) Web structure mining
3) Web usage mining

1) Web Content Mining:Web Content Mining is the process of extracting useful information from the contents of Web documents. Web content mining is used to extract the text, image, or other information and knowledge component of the web content. For example, which sites sell cars? Which pages are in Chinese? Which pages introduce the music, or introduce news? Search engines, intelligent agents, and some recommend use content mining to help the user in the vast network of space to find the necessary content.
2) Web Structure Mining:Web Structure Mining is a procedure of concentrating data from linkages of website pages. It can be regarded as the process of discovering structure information from the Web. It is used to extract the network topology information, that is, the link between pages of information. Mine knowledge from the WWW organization and links. For example, which pages

are linked to other pages? Which pages point to other page?

This type of information can be extract using web structure mining. This is also used to analyse the link Structure of the web.

This can be of two types,

- Hyperlink Structure:

A Hyperlink is a structural unit that connects a Web page to different location, either within the same Web page or to a different Web page. It can be *Intra-DocumentHyperlink* that connects to different location in the same page and hyperlink that connects two different pages is called an *Inter-Document Hyperlink.*

- Document Structure:

The content within a Web page can also be organized in a tree-structured format, based on the various HTML and XML tags within the page.

3) Web Usage Mining:Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Web usage mining is used to extract about the customer how to use the browser and use the page links.

For example, which pages are the client accesses? How long spent on each page? What next click on? What are the entry and exit routes?

2. Existing Algorithms

1) Enhanced-Ratio Rank: Enhancing Impact of Inlinks and Outlinks.[6]

The Enhanced- Ratio Rank is the extension to the Weighted Page Ranking Algorithm in which more weights are given to the Inlinks and Outlinks on the basis of the popularity of the links. In this algorithm the page is considered to be more important if the Inlinks of that webpage is visited by the user more than any other webpage and many other good pages out linked by it, means in overall it may be said that all the features which are considered will come together to rank the webpage. The Enhanced-Ratio Rank both the Inlinks and the Outlinks are being considered for computing the page ranking, and the number of times the user visit the particular link.

The $W^{in}_{(v,u)}$and $W^{out}_{(v,u)}$are used to record the popularity of the Inlinks and Outlinks based on the Inlinks and Outlinks of that link.

The mathematical equations of the weights are given as follows:

$W^{in}_{(v,u)}$ is the weight of *link*(*v*, *u*) calculated based onthe number of inlinks of page *u* and the number of inlinksof all reference pages of page *v*.

$$\boldsymbol{W^{in}_{(v,u)} = {I_u}\Big/{\sum_{p=R(v)} I_p}}$$

Where,

- *Iu* and *Ip* represent the number of inlinks of page *u* and page *p*, respectively.
- *R*(*v*) denotes the reference pagelist of page *v*.

$W^{out}_{(v,u)}$is the weight of *link*(*v*, *u*) calculated based onthe number of outlinks of page *u* and the number of outlinksof all reference pages of page *v*.

$$\boldsymbol{W^{out}_{(v,u)} = {O_u}\Big/{\sum_{p=R(v)} O_p}}$$

Where,

- *Ou* and *Op* represent the number of outlinks of page *u* and page *p*, respectively.
- *R*(*v*) denotes the reference page list of page *v*.

Mathematical Equation for Enhanced-Ratio Rank Algorithm

$$\textbf{PR (u) =(1-d) + d*}\sum_{v\epsilon B(u)} \frac{[\left(V_u * 0.7 * W^{in}_{(v,u)} + 0.3 * W^{out}_{(v,u)}\right).\textbf{PR(v)}]}{TL(v)}$$

Where,

- PR (u) and PR (v) are ranking of the webpages u and v respectively,
- $W^{in}_{(v,u)}$ and $W^{out}_{(v,u)}$are used to record the popularity of the Inlinks and the Outlinks,
- 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) are the pages which points to webpage u.

In algorithm Inlinks of any webpage votes more to increase the rank of the webpage, less of that by Outlinks, that's why more weightage is given to the Inlinks than to the Outlinks and the ratio as 70/30 is defined on the basis of the better results than all other ratios.

Steps for the Enhanced- Ratio Rank Algorithm:

1) Take the link structure of the retrieved webpages from the crawler.
2) Obtain the web graph from the link structure of the retrieved webpages.
3) Assign as the initial ranking to all the webpages.
4) Calculate the weights of Inlinks and Outlinks.
5) Apply the proposed algorithm (Ratio Rank).
6) Repeat the process iteratively until ranks of all webpages are stable means same in two consecutive iteration.

Advantages of Enhanced-Ratio Rank algorithm:

- The algorithm shows better accuracy in term of the relevancy of the pages returned because it uses the inlinks, outlinks and visit count of the links to rank the pages.
- This algorithm removes the limitation of the standard page ranking algorithm as if the page with no inlinks is reached then surfer jams while in the case of the Ratio Rank the rank will be calculated in terms of outlinks.
- It gives the ratio between inlinks and outlinks, according to ratio rank the web pages.

Drawbacks of Enhance-Ratio Rank Algorithm:

- It give the results based on the links.
- This algorithm does not give the results based on the user interest it gives the results according to the links of the pages.

3. Reading Time: A Method for Improving the Ranking Scores of Web Pages.[7]

Many Page Ranking algorithm gives the results based on the links of the web structure. 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.

The main aim of this algorithm 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. So, using reading time, calculate the rank of those pages which has a high Inlinks and the high reading time.

In this paper they 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. When the user close the web page, than script will send the message to the web server with the information about the reading time or the time which is spends by the user on that pages and the hyperlink.

On the server side, a database will used to store all the information about the pages including visit of links. Using this information calculate the rank of that page. And the result will the based on the Inlinks of that pages and the reading time which is spend by the user, so in general results is given based on the user interest of that pages.

The mathematical equation for calculating the rank of the pages using reading time is:

$$\mathbf{PR\ (u) = (1\text{-}d) + [\ d * \sum_{v \in B(u)} \frac{PR(v) * L_u}{TL(v)}] * RT(u)}$$

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 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.

Advantages Reading Time Page Ranking Algorithm:

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 visiting 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. Proposed Algorithm

In this paper a new page ranking algorithm is presented based on the links of the webpages and average reading time which is user spent on the webpages named as the improved anenhanced-ratio rank algorithm based on reading time. The algorithm take a weighted of the inlinks and outlinks and average reading time which is user spent on the webpages. Here we use the 70-30 ratio for the weight of the links, 70% weight is given to the inlinks because page rank algorithm gives the higher to those webpages which has more links as compare to the other webpages and 30% weight is given to the outlinks of webpages.

To, count the average reading time of the webpages we use the client side and server side script based on that we count the average reading time of the users. We use the click event and timeout for the purpose of that user can not intentionally increase the page rank by spending a more time on the that particular one webpages and rank that page at a higher position on the top of the list.

If no movement can be done than webpage can redirect to the homepage and the time which is spent on that webpages it can remove from the dataset.

The mathematical equation for calculating the rank of the pages using reading time is:

$$\text{PR(u)=} \; (1+d)+d*\sum_{v \epsilon B(u)} \frac{[\left(V_u * 0.7 * W^{in}_{(u,v)} + 0.3 * W^{out}_{(u,v)} \right).PR(v)]}{TL(u)} * \; T\ (u)$$

Where,

- d is the dampening factor,
- u and v represents the web pages,
- B[u] is the set of webpages,
- PR[u] and PR[v] are the page ranks of page u and v respectively,
- Lu is 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 Average reading time that user's take to read a page u,

A. Flowchart of Proposed System

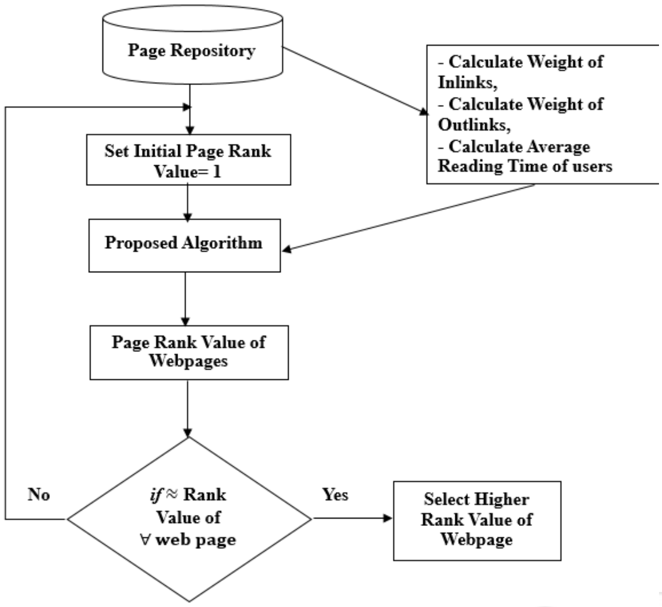
Page Repository
- Calculate Weight of Inlinks,
- Calculate Weight of Outlinks,
- Calculate Average Reading Time of users
Set Initial Page Rank Value= 1
Proposed Algorithm
Page Rank Value of Webpages
No
if ≈ Rank Value of ∀ web page
Yes
Select Higher Rank Value of Webpage

Figure 2: Flow diagram of propose work

B. Experimental Dataset

Table 1: Dataset

Sr. No.	Web Pages	url
1	Book1.aspx	http://localhost:1791/Book1.aspx
2	Book2.aspx	http://localhost:1791/Book2.aspx
3	Book3.aspx	http://localhost:1791/Book3.aspx
4	Book4.aspx	http://localhost:1791/Book4.aspx
5	Book5.aspx	http://localhost:1791/Book5.aspx
6	Book6.aspx	http://localhost:1791/Book6.aspx
7	Book7.aspx	http://localhost:1791/Book7.aspx
8	Book8.aspx	http://localhost:1791/Book8.aspx
9	Book9.aspx	http://localhost:1791/Book9.aspx
10	Book10.aspx	http://localhost:1791/Book10.aspx
11	Book11.aspx	http://localhost:1791/Book11.aspx
12	Book12.aspx	http://localhost:1791/Book12.aspx
13	Book13.aspx	http://localhost:1791/Book13.aspx
14	Book14.aspx	http://localhost:1791/Book14.aspx
15	Book15.aspx	http://localhost:1791/Book15.aspx

C. Implementation Results of Existing Algorithm

Table 2: Results of existing algorithm

Sr. No.	Webpages	PR Value Of Enhanced-ratio rank algorithm	PR Value of Reading time based page rank algorithm
1.	Book 1	0.171	0.6626
2.	Book 2	0.17	0.2133
3.	Book 3	0.16	0.7797
4.	Book 4	0.161	0.8269
5.	Book 5	0.159	0.1753
6.	Book 6	0.168	0.3051
7.	Book 7	0.166	0.7853
8.	Book 8	0.161	0.2376
9.	Book 9	0.166	0.5963
10.	Book 10	0.155	0.447
11.	Book 11	0.155	0.5905
12.	Book 12	0.166	0.8839
13.	Book 13	0.164	0.446
14.	Book 14	0.162	0.1212
15.	Book 15	0.159	0.2288

D. Graph for page rank value of existing algorithms:

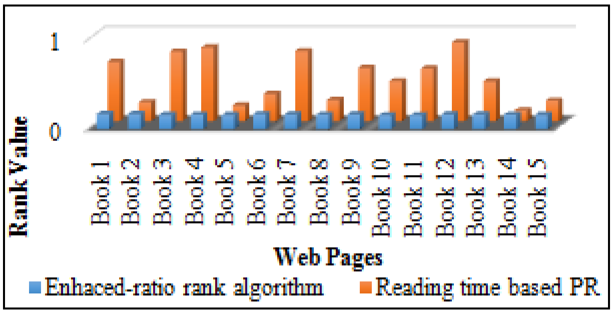

Figure 3: Graph of existing algorithm

In enhanced-ratio rank algorithm, webpage Book-1 has the higher page rank value as compare the other webpages, because this algorithm gives the results based on the 70% weight of inlinks, 30% weight value of outlinks of the webpages. And in reading time based page ranking algorithm, webpage Book-12 has the higher rank value as compare to the other webpages, because this algorithm gives the rank value based on the average reading tine of the users which is spent on the webpages and the visit of links of the webpages.

But the enhanced-ratio rank algorithm has limitation is that is give the results using only links of the webpages, so user can intentionally increase the links of the particular one webpage and gives the higher position on list of page rank. And reading time based page rank has limitation is that it use on inlinks and visit of links of the webpages it not use the out links of the webpages which is more important part in the gives the rank of webpages.

This propose algorithm is overcome the limitation of the both existing algorithm by adding the average reading time of the different user on the enhanced-ratio rank algorithm.

E. PageRank Value for Proposed algorithm

Table 2: PageRank Value of Proposed Algorithm

Sr. No.	Webpages	Page Rank Value
1.	Book 1	1.384
2.	Book 2	2.578
3.	Book 3	2.479
4.	Book 4	5.582
5.	Book 5	1.641
6.	Book 6	2.296
7.	Book 7	9.505
8.	Book 8	7.373
9.	Book 9	7.677
10.	Book 10	2.401
11.	Book 11	8.349
12.	Book 12	3.346
13.	Book 13	1.845
14.	Book 14	1.091
15.	Book 15	7.026

F. Graph for the page rank value of Propose algorithm

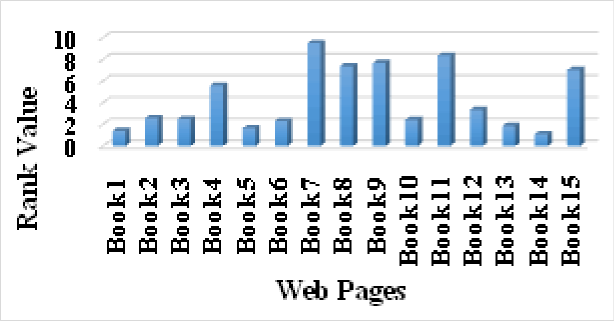

Figure 4: Graph for the page rank value of propose algorithm

5. Conclusion

By analysis some of the existing page rank algorithm has some limitation, which include enhanced ratio rank algorithm is gives the rank based on the LINKS, not gives the rank based on user interest. The reading time based algorithm is gives the rank based on the times and inlinks but not use outlinks of webpages. So, our proposed work is including reading time in enhanced ratio-rank algorithm for getting the best results than the other existing algorithms and overcome issue of both existing algorithm. Propose algorithm is gives the better results as compare to the existing algorithm and reduce the limitation of the both existing algorithm.

6. Acknowledgement

The authors would like to thank HOD, Prof. Bijal J. Talati for her continuous support, encouragement.

References

[1] Sergey Brin, Lawrence Page, "*The anatomy of a large-scale hyper textual Web search engine*", Computer Networks and ISDN Systems-1998.

[2] Wenpu Xing, Ali Ghorbani, "*Weighted PageRank Algorithm*", IEEE- 2004.

[3] Neelam Duhan, A. K. Sharma, Komal Kumar Bhatia,"*Page Ranking Algorithms: A Survey*" IACC-2009.

[4] Gyanendra Kumar1, Neelam Duhan2, A. K. Sharma,"*Page Ranking Based on Number of Visits of Links of Web Page*", ICCCT-2011.

[5] Neelam Tyagi, Simple Sharma, "*Weighted Page Rank Algorithm Based on Number of Visits of Links of Web Page*" IJSCE)-July 2012.

[6] Ranveer Singh, Dilip Kumar Sharma, "*Enhanced-RATIORANK: Enhancing Impact of Inlinks and Outlinks*", IEEE-2013.

[7] Shweta Agarwal, Bharat Bhushan Agarwal, "*Reading Time: A Method for Improving the Ranking Scores of Web Pages*", International Journal of Computer Applications, August 2013.

[8] Sanjay, Dharmender Kumar, "*A Review Paper on Page Ranking Algorithms*", IJARCET,June 2015.

[9] Dilip Kumar Sharma, A. K. Sharma, "*A Comparative Analysis of Web Page Ranking Algorithms*", IJCSE-2010.

[10] Shweta Agarwal, Bharat Bhushan Agarwal, "*Comparative Analysis of Various Web Page Ranking Algorithms*", International Journal of Computer Science & Information Technology, 2013.

[11] Ms.M.Sangeetha, Dr.K.Suresh Joseph, "*Page Ranking Algorithms used in Web Mining*" ICICES2014.

[12] Shikha Goel, Suita Yadav, "*Search Engine Evaluation Based on Page LevelKeywords*", IEEE.