Improvement in Weighted Page Rank Algorithm using Efficiency and Precision

Er. Manika Dutta[1]
Department of Computer Science, Himachal Pradesh University, Shimla, India[1]
E-mail: 14manika@gmail.com

Dr. Kishori Lal Bansal[2]
Professor , Department of Computer Science, Himachal Pradesh University, Shimla, India[2]
E-mail: kishorilalbansal@yahoo.com

Abstract: Weighted Page Rank (WPR) algorithm is an extension to the standard Page Rank algorithm of Google. WPR assigns larger rank values to more important pages considering both inlinks and outgoing links of the web pages and assigns weight to both of them. WPR resolves the core problem of rank sink present in Page Rank algorithm. In this paper we have proposed the improvement in existing WPR algorithm using its two parameters efficiency and precision. Both these parameters discuss the performance of the proposed Weighted Page Rank algorithm using wampserver 2.4, MATLAB R2013A The Math Work Inc. and My Sql database 5.0. New improved WPR hold efficiency above average and also significant improved relevancy. As a result, improved relevancy results in higher precision.

Keywords: ***Weighted Page Rank algorithm, Page Rank algorithm, Efficiency, Precision, Relevancy.***

I. INTRODUCTION

Although Weighted Page Rank algorithm takes into account content, backlinks and forward links it completely ignores the relevancy and comprises average efficiency. Weighted Page Rank algorithm is an extension to the standard page rank algorithm of Google[1]. Discovered in 1988 by Larry Page and Sergery Brin , Page rank is a link analysis algorithm representing the numerical value which denotes the importance of a web page by counting the number of backlinks. Hence importance of a web page becomes directly proportional to the number of web pages linked to it. It is only associated with the individual web page and not the entire website[2]. Page rank algorithm computes the principal eigen vector of the matrix whose elements describes the hyperlinks of the web graph using the Power method[3]. It is non keyword specific and link structure based thus evaluates approximately 25 billion web pages present on the world wide web to assign a rank score. To every user query submitted to Google it combines the precomputed Page rank score with text matching score and after that assigns a rank. Page rank uniformly divides the page rank score equally among all its outlinks. This algorithm states that if a page has some important incoming links to it then its outgoing links to other pages also become important. As a result Page Rank takes the backlinks into account and propagates the ranking through links,i.e. a page has a high rank if the sum of the ranks of its backlinks is high. Figure 1 shows an example of backlinks. Here page A is a backlink of pages B and C, while both pages B and C together act as backlinks to page D.

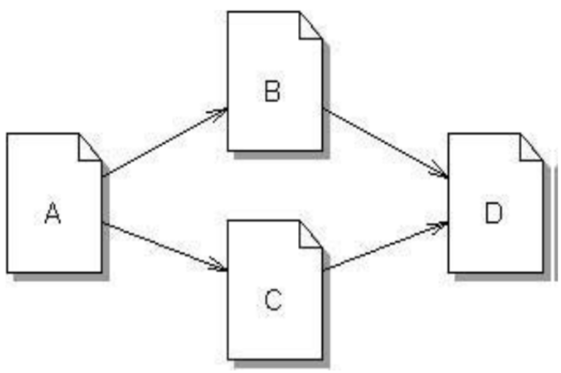

Figure 1 An example of backlinks[4].

Mathematically PageRank is defined as:

$$PR(u) = c \sum_{v \in B(u)} PR(v) / N_v \quad (1)$$

In Equation (1) u represents a web page. B(u) is the set of pages that point to u . PR(u) and PR(v) are rank score of page u and v respectively. N_v represents the number of outgoing links of page v. c is a factor used for normalization. In Figure 2, c =1.0 is used to simplify the calculation and it also shows how Page Rank uniformly distributes its rank score among all its following links.

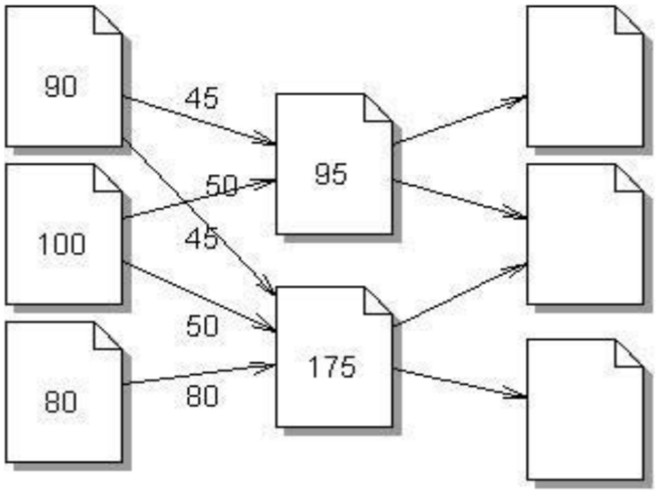

Figure 2 Distribution of Page Ranks[2].

1.1 Problem of Rank Sink: In Page Rank algorithm , the rank score of a page is evenly divided among all 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. The rank score of pages of a website could be calculated iteratively starting from any web page. Within a website, two or more pages might connect to each other to from a loop. If these pages did not refer to but are referred to by other web pages outside the loop, they would accumulate rank but will never distribute any rank. This scenario is called *rank sink* [4].

To solve the rank sink problem, user activities are observed. A phenomenon founds out that not all users follow the existing links . For example, after viewing page a, some users may not decide to follow the existing links but decide to go to page b , which is not directly linked to a. For this purpose the users just type the URL of page b into the URL text field and jump to page b directly. In this case the rank of page b should be affected by page a even though these two pages are not directly connected. Therefore rank sink gets abolished now.

To eliminate the problem of rank sink, later on Page Rank was modified as per the random surfer model. Equation(2) shows the modified Page rank together with accumulation of damping factor.

$$PR(u) = (1-d) + d \sum_{v \in B(u)} PR(v)/N_v \quad (2)$$

d – is a damping factor usually set to 0.85 which is defined as the probability of users following the direct links.
(1-d) – is the page rank distribution from non directly linked web pages.

1.2 Merits of Page rank

- As it is a query independent algorithm i.e it precomputes the rank score hence it takes very less time.
- As this algorithm computes the rank score at indexing time and not at query time so it is more feasible.
- It returns important pages as rank is calculated on the basis of popularity of a page.
- It is less susceptible to link spam because for calculating rank value of a page, it considers the entire web graph rather than a small subset.

1.3 Demerits of Page rank

- It favors older pages, because for a new page even a good one will not have many links unless it is part of an existing web site or a loop of web pages.
- Relevancy of the returned pages to user query is very less as content of web page is not considered.

- Presence of Dangling links. This occurs when a page contains a link such that the hypertext points to a page with no outgoing links.
- Web pages present in a network may get in infinite link cycles and result in the problem of rank sink.
- Dead Ends: Dead ends are pages with no outgoing links.
- Spider Traps: A group of pages is called a Spider Trap if there are no links from within a group to outside the group.

1.4 Weighted PageRank Algorithm

In 2004, Winpu Xing and Ali Ghorbani proposed Weighted Page Rank algorithm (WPR) algorithm which is an modification to the standard Page Rank algorithm. This Weighted Page rank algorithm resolves the problem of rank sink present in Page rank algorithm. This algorithm assumes that if a page is more popular ,more linkages other web pages tend to have to it or are linked by it. Equation (3) denotes that WPR assigns large rank values to more important pages instead of uniformly distributing the rank score among all the outgoing links [5].

$$\textbf{PageValue } \alpha \textbf{ Popularity} \quad (3)$$

While calculating the popularity of a web page both inlinks as well as outlinks are considered and weight is assigned to both of them which are denoted as W_{in} (v,u) and W_{out} (v,u) respectively.

Equation (4) gives formula for W_{in} (v,u) which is the weight of link (v, u) calculated based on the number of inlinks of page u and the number of inlinks of all reference pages of page v.

$$W_{in}(v,u) = I_u / \sum p€R(v)\ I_p \quad (4)$$

In equation (4) I_u and I_p represent the number of inlinks of page u and page p respectively. R(v) denotes the reference page list of page v.

Equation (5) gives formula for W_{out}(v,u) which is the weight of link (v,u) calculated based on the number of outlinks of page u and the number of outlinks of all reference page list of page v.

$$W_{out}(v, u) = O_u / \sum p€R(v)\ O_p \quad (5)$$

Also here O_u and O_p denotes the number of outlinks of page u and page p respectively. R(v) denotes the reference page list of page v. d is a damping factor which can be set between 0 and 1(usually d is taken 0.85). Hence equation (6) gives the mathematical and modified Page rank formula for Weighted Page Rank algorithm. In this equation (6) d and (1-d) gives the probability of users following the direct and indirect links respectively.

$$PR(u) = (1-d) + d \sum PR(v)\ W_{in}(v,u)\ W_{out}(v,u) \quad (6)$$

1.4.1 Merits of Weighted Page Rank

- High quality web pages are returned by the web pages as compared to the Page rank algorithm.
- It is more efficient than Page rank because rank value of a page is divided among it's outlink pages according to the importance of that page irrespective of present in a loop or not.

1.4.2 Demerits of Weighted Page Rank

- As this algorithm considers only link structure and not the content of the page , it returns less relevant pages to the search query.

Table 1 Comparison between Page rank and Weighted Page rank algorithm[6]

Algorithm	**Page Rank**	**Weighted Page Rank**
Main technique	Web structure mining	Web structure mining and web content mining
I/O parameters	Backlinks	Content, Backlinks and Forward links
Working	Computes page rank at the time of indexing of pages.	Weight of web pages is calculated on the basis of input and output links.
Efficiency	Very less	Average
Significance	High, backlinks are considered.	High, the pages are sorted according to relevance.
Drawbacks	Results come at the time of indexing and not at query time.	Relevancy is ignored.
Complexity	O(log n)	<O(log n)
Quality of result	Medium	Higher than Page rank

II. RELATED WORK

Usually end users find, extract, filter and evaluate the desired useful information by means of an automated tool called search engine while accessing the internet. Although page ranking algorithms could be either link or content based but search engines typically use link analysis algorithms to rank and find the quality web pages according to user needs[7]. Web mining categorizes user and pages by analyzing the user behavior, page content and order of the URL's.

Although Sequential modified Page Rank algorithm resolved the issues such as (Rank Sink, Dangling node, topic drift etc.) present in the basic page rank algorithm, but it still had issues regarding efficiency and relevancy. Authors explored numerous modified page ranking algorithms in various environments used by researchers such as parallel distributed etc.[8]. Word Sense Disambiguation resolved the problem of identifying the senses of word in textual context when word had multiple meanings. The proposed Dynamic Page Rank algorithm calculated the Reciprocal Rank for both the algorithms and presented comparative results [9]. Both Page Rank and Weighted Page Rank algorithm are query independent algorithms as they are based on the web structure, mines hyperlink of the web graph. Nidhi shalya et al. [10] proposed a new modified page rank algorithm based on both content and link structure, thereby reducing the search space.

Semantic web will be the next generation web hence ontology based ranking algorithms will be dominant in future[11]. Kaushal kumar et al. [12] observed that web mining lies at the core of Page Rank calculation. They also studied the variations of Page rank and Weighted page rank based on the number of visit of links. Page rank algorithm was applied to nodes in a linked database.i.e any database of documents with citations. Page rank asserts the importance of a page by assigning a particular rank to a page which in turn affects the rank of other web pages in the search results. A relationship was deduced to calculate the Page Rank of a web page as a function of the link distance from a web page whose Page Rank is known. Damping factor d=0.85 spreads uniformly as part of the rank[13]. Hema dubey et al. [14] proposed a new optimized page rank algorithm based on the normalization technique calculating the mean value of page ranks. Consequently number of iterations and time complexity reduced.

As Weighted Page rank algorithm doesn't provide the relevant results at the top of the retrieved list. Thus a new weighted page rank algorithm based on the content of the pages was proposed. This Weighted page content rank algorithm used both links and contents of the web graph and also provided relevant results at the top of the list[15]. Also Nagappan et al.[16] proposed an enhanced Agent based Weighted PageRank algorithm which improved the order of the pages in the retrieved result list by means of an agent. Hence users got relieved from the problem of finding irrelevant results at the top. While both Page rank and Weighted page rank calculates the page rank score at indexing time, HITS(Hyper Induced Topic Search) does so at query time. For that reason in future need arises for an algorithm which computes the rank score of web pages at both query as well as indexing time[17].

III. IMPLEMENTATION OF PROPOSED ALGORITHM

We have used wampsever 2.4, MATLAB R2013a The Math Work Inc. and a dataset of various web pages from different e-books for running the code of both existing Weighted page rank and proposed Weighted page rank algorithm in MATLAB using simulation. MATLAB software which stands for MATrix LABoratory was developed by LINPACK(linear system package) and ESIPACK(Eigen system package) projects for easy admission to matrix software. MATLAB is a modern programming language which takes array as a basic element and delivers high performance for technical processes e.g. research work. Its commands are easy to use for graphics that provides faster results of visualization of images. Wampserver is a windows web development environment. It allows the end users to create web applications with Apache, PHP and mySql database. It also comes with phpMy admin to easily manage the databases. Also MySql 5.0 is used. Figure 3 to Figure 9 displays the series of screenshots of all the outputs after running the code of proposed Weighted Page rank algorithm having efficiency above average and also significantly improved relevancy leading to higher precision.

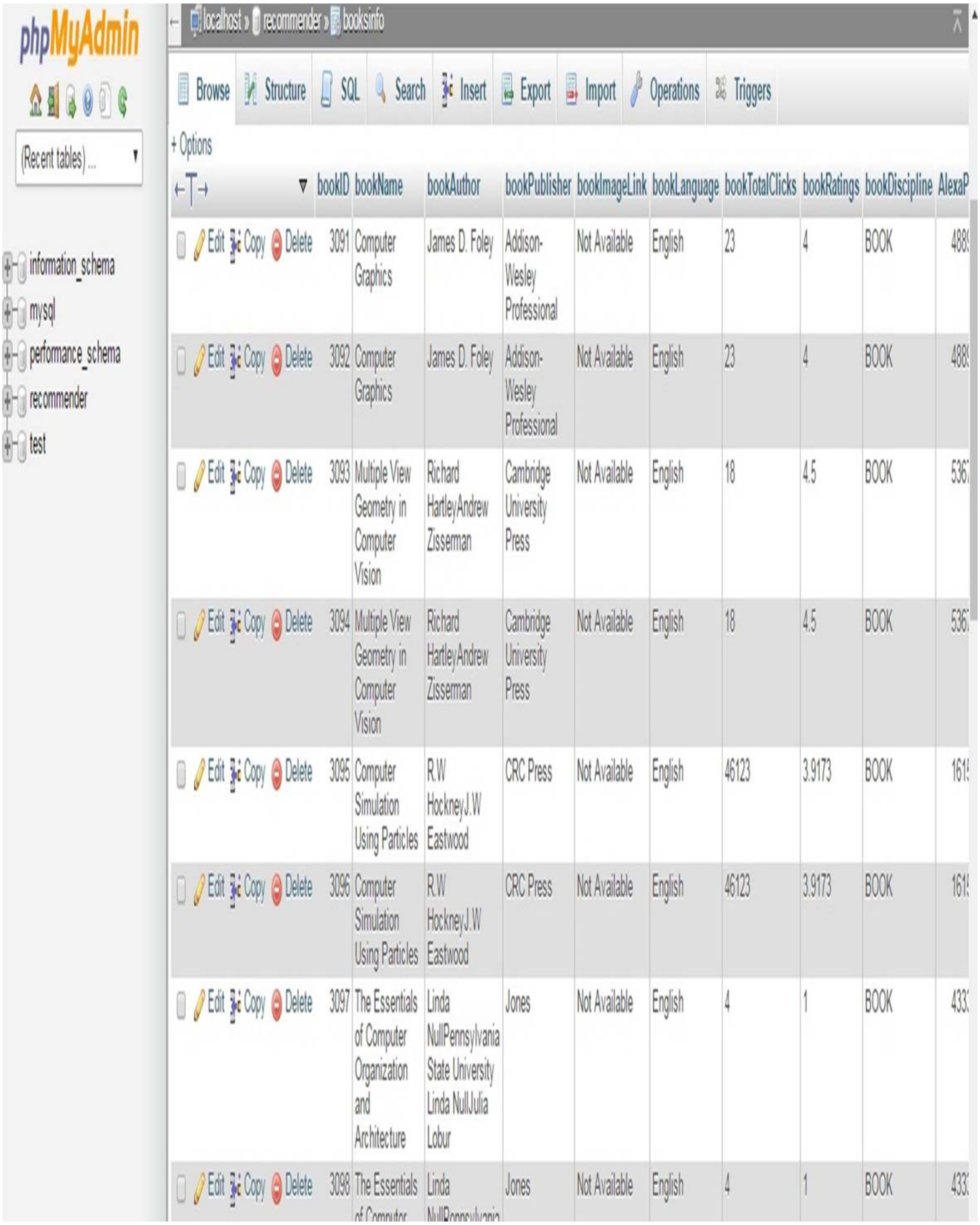

Figure 3 Screenshot of data set of web pages from various e-books.

Figure 3 shows the screenshot of the data set taken for the implementation of the proposed WPR which consists of web pages from different e-books.

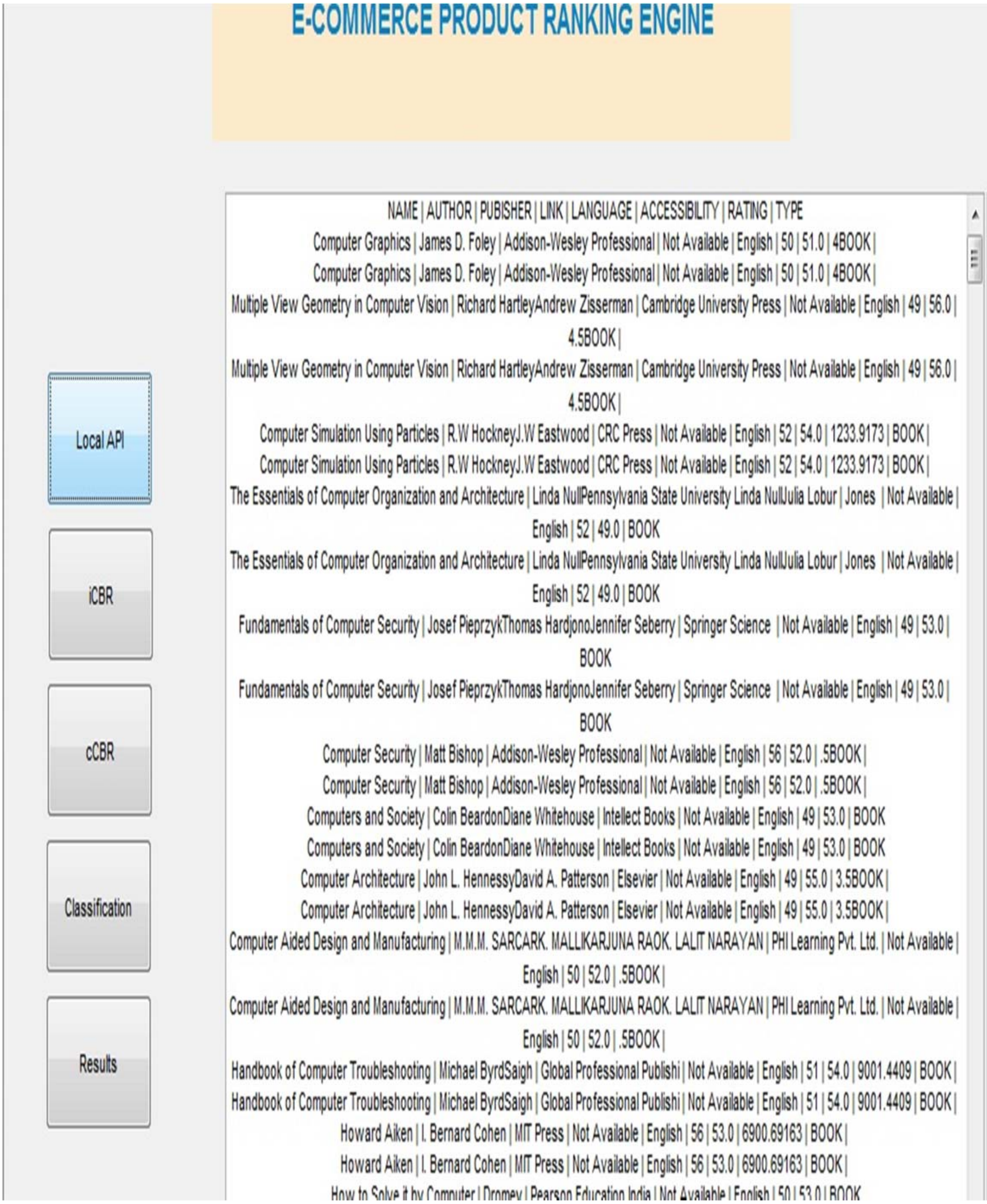

Figure 4 Screenshot of Local API for data sorting in WPR.

Figure 4 shows the screenshot of Local API(Application Programming Interface) of WPR which sorts the data according to name ,author and publisher of book, web link, book available in particular language, accessibility, rating and type.

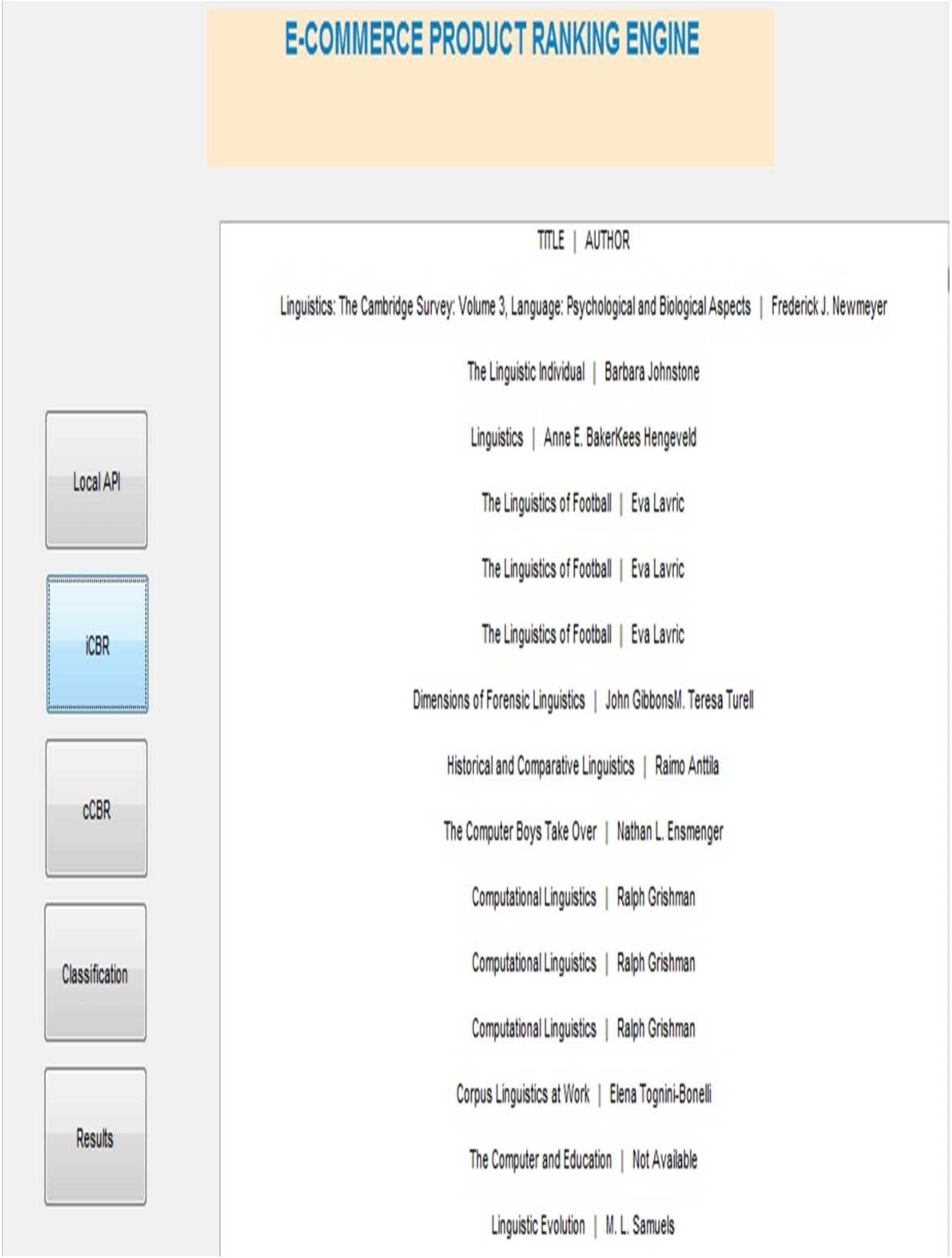

Figure 5 Screenshot of iCBR for data sorting in WPR.

Figure 5 depicts the screenshot of iCBR(Context Based Restructuring) in WPR which sorts the data into two categories namely book's title and author name.

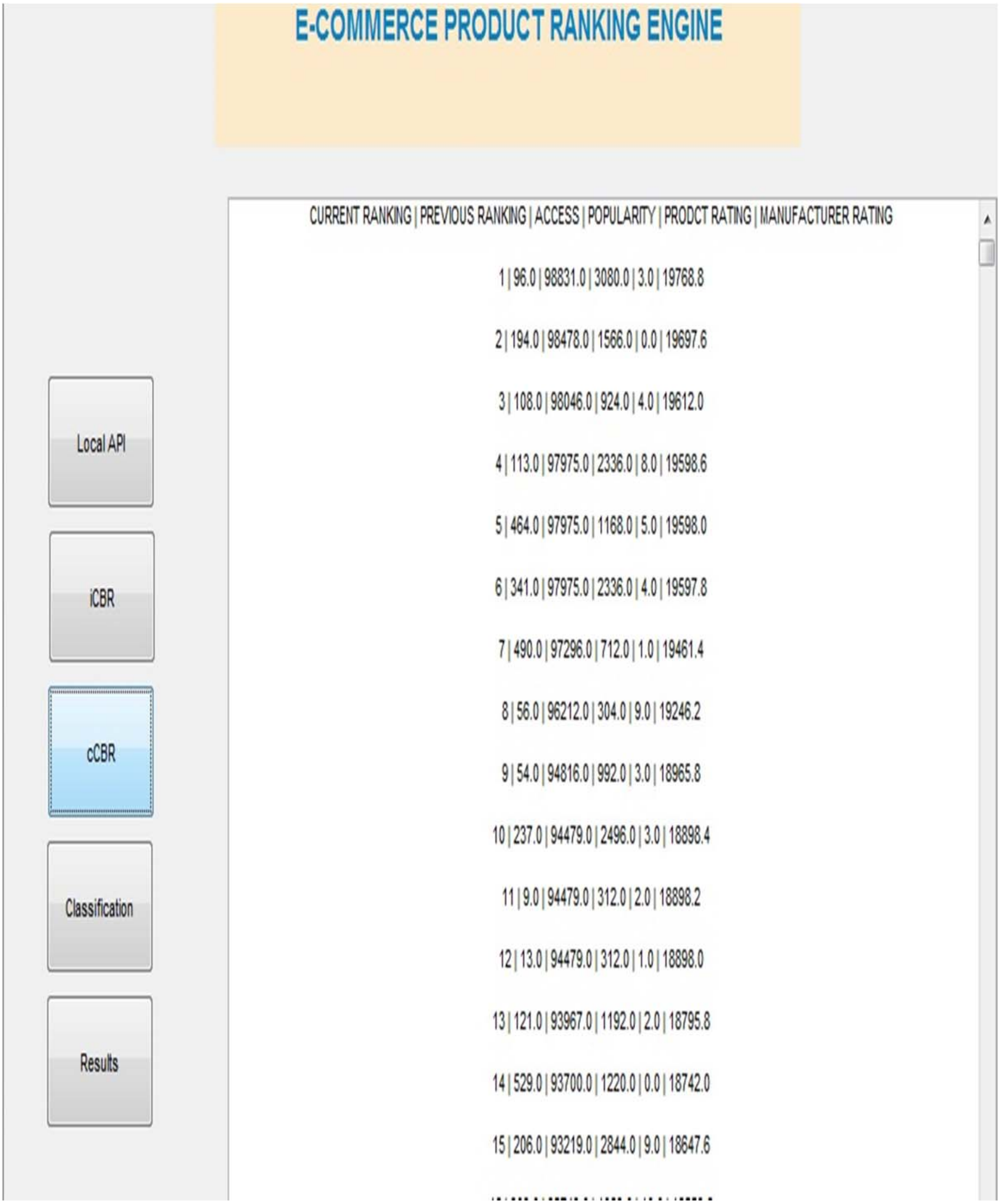

Figure 6 Screenshot of cCBR for data sorting in WPR.

Figure 6 shows the screenshot of cCBR(Context Based Restructuring) in WPR which sorts the data according to current and previous ranking of web pages, access, popularity, product rating and manufacturing rating.

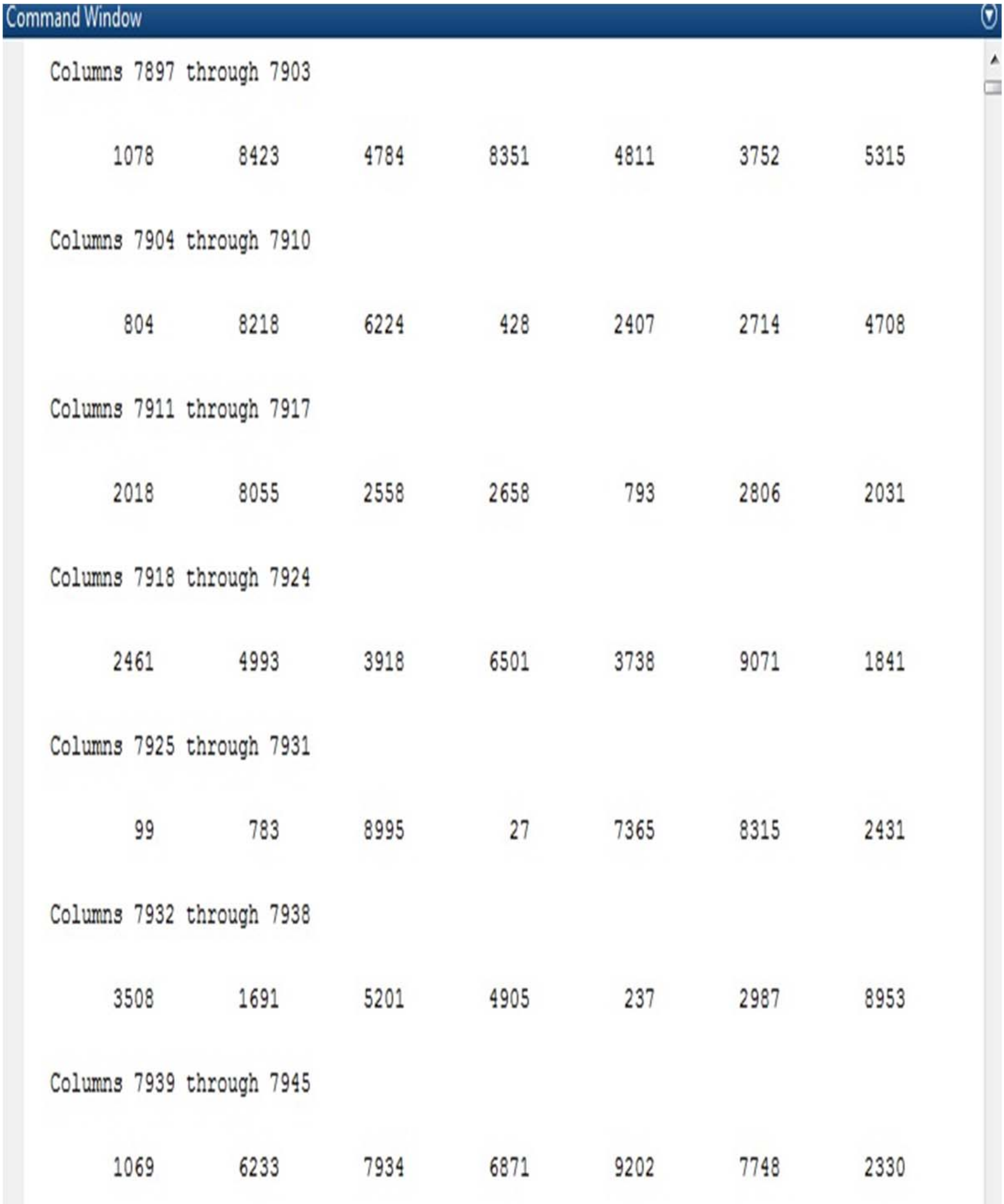

Figure 7 Screenshot of dialog box of command window in MATLAB for WPR showing sorted data across various columns.

Figure 7 displays screenshot of dialog box of command window in MATLAB for WPR which classifies and sorts the data present in the matrix across various columns.

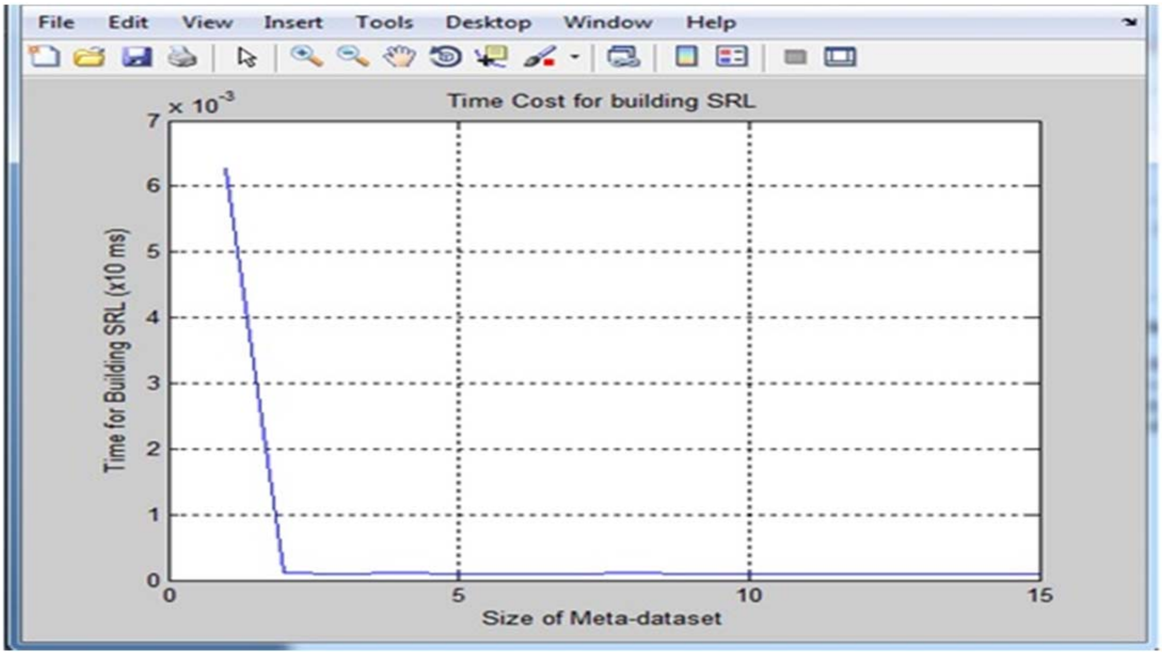

Figure 8 Time taken to build SRL.

Figure 8 shows the total time taken to build the SRL(Semantic Relevance Library), where time for building SRL is plotted on Y-axis and size of meta data set across X-axis. This will display the library having web pages to the user after applying the WPR algorithm.

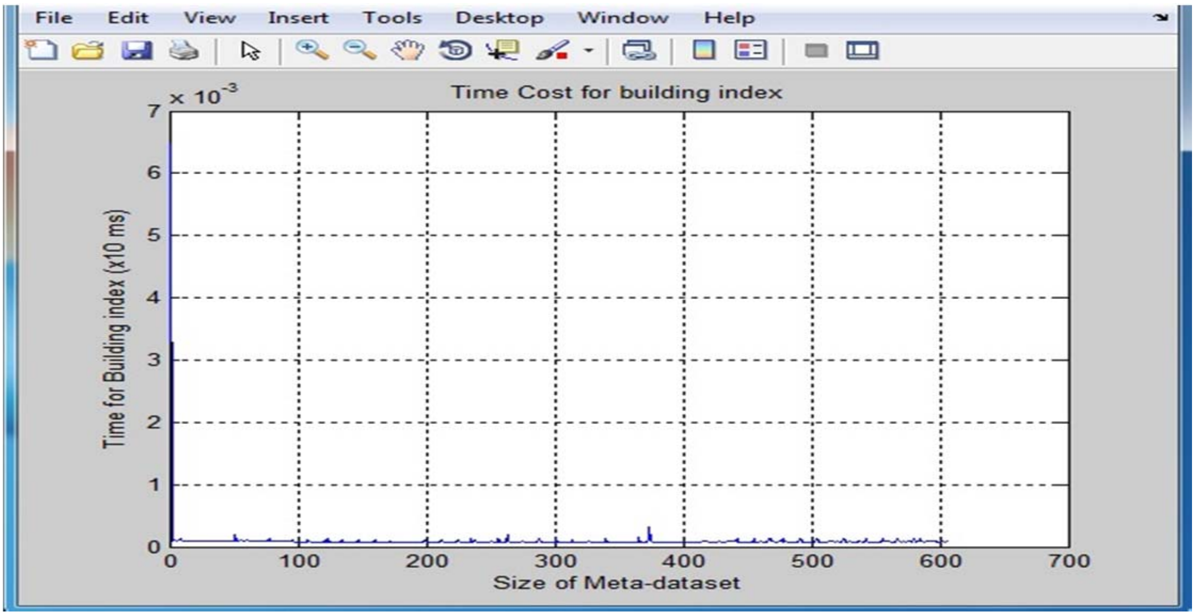

Figure 9 Time taken to build index.

Figure 9 shows the time taken to build the index of the matrix. It means that the data will be sorted according to page ranking for the entire data set. First WPR has been applied and after that data will be sorted according to page ranking.It shows time for building index acoss Y-axis and size of meta- dataset along X-axis.

IV. RESULTS

This section shows the comparative results of two parameters precision and efficiency respectively. Results have been obtained while performing of Weighted Page Rank algorithm on the dataset of web pages. Various iterations has been performed to check the consistency of the model.

4.1 Precision: It means how well the website priority tool (WPT) is working. Website priority tool allows comparison of websites using dropdown box and search box to specify string of specific product. The dropdown box adds as many URL's(Uniform Resource Locator) of the website and after comparison, WPT tool assigns priority to each candidate website based on the calculation of content priority module, time spent priority module, recommendation module and neural priority module[18]. Hence precision is used to measure the consistency of the results for each and every time the system runs. More the relevancy of the fetched web pages higher will be the consistency of the system. Higher consistency of the results implies that the website priority tool is working accurately. As a result, higher accuracy of website priority tool leads to higher precision. Relevancy is calculated by measuring the distance of the data. Data has been stored in array/ matrix form. Distance will be calculated for each row by comparing it with all other rows. For each row, lesser the distance between rows more relevant will be data and vice versa. Precision values of the proposed system have been obtained by applying multiple testing rounds(iterations) approximately 25 on the data set. Table 2 shows the values for the proposed WPR along with website priority tool and Google.

Table 2 The precision based result evaluation

Iteration	**Website Priority Tool**	**Google**	**Proposed**
1	1.5	1.9	2.75
2	2.65	2.3	2.8
3	2.48	2.1	2.85

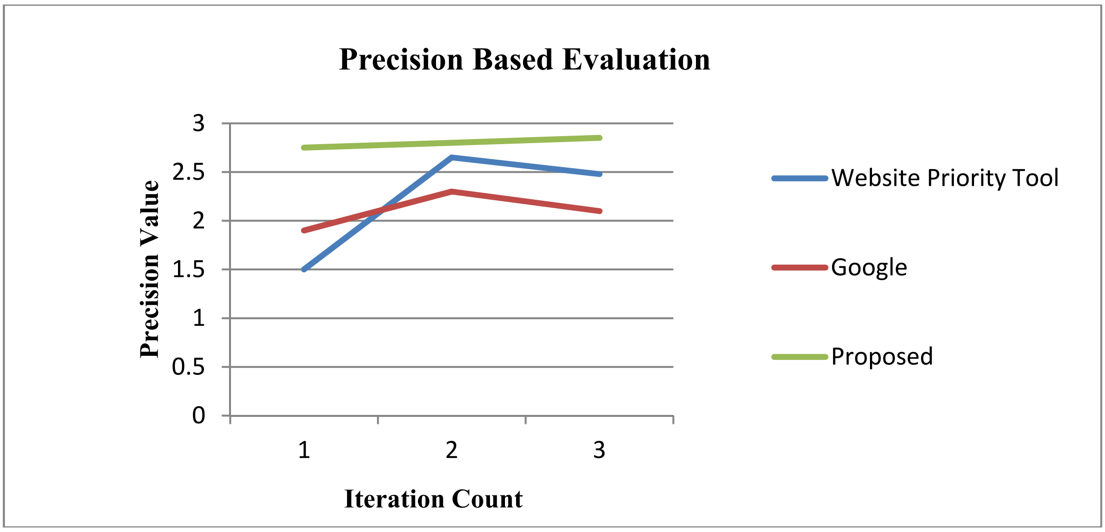

Figure 10 Precision based evaluation of the proposed model.

Figure 10 graphically shows the precision values for website priority tool, Google and the proposed WPR.The line graph here clearly shows that the proposed Weighted page rank algorithm has high precision values for all the iterations. The graphical design of both the models Page Rank and improved Weighted Page Rank showcase the comparative analysis which has been evaluated on the basis of the precision of the simulated results. The improvement has been recorded which justifies the improved performance of the proposed WPR model than the existing model.

4.2 Efficiency

Here the efficiency refers to the average efficiency. Normally efficiency is calculated by summing the entropy value, standard deviation and range.

4.2.1 How to calculate efficiency: WPR algorithm works upon the dataset/entities of data written in user readable (English) language. It is converted into binary values to apply the algorithm. All data has been stored into array/ matrix to apply WPR algorithm in sequential order. Proposed WPR algorithm gives results for each index of the matrix. After getting the individual result for each index i.e average will be calculated for entire matrix each testing round. Approximately 25 testing rounds were performed to obtain the average efficiency values for the proposed WPR algorithm. Table 3 contains values of average efficiency for both existing and proposed WPR. These values were obtained from each testing round by calculating the average.

Table 3 Efficiency Comparison table of existing and proposed WPR

Sr. No	Proposed WPR	Existing WPR
1	24.92	22.85
2	24.89	23
3	24.95	22.93

Table 4 shows five testing rounds(TR) taken for calculation of efficiency. All the values were aggregated for each testing round separately. Lastly sum total of all the aggregated values were taken which gave one value/entry of average efficiency of proposed WPR as listed in table 3. For example, from table 4, first value in table 3 was calculated as:

Efficiency: 4.9838+ 4.9825+ 4.9346+ 4.9773+ 4.9964= 24.87

Table 4 Various testing rounds for calculation of average efficiency

Sr. No.	TR1	TR2	TR3	TR4	TR5
1	0.9968	0.9957	0.9867	0.9958	0.9988
2	0.996	0.9961	0.9868	0.9957	0.999
3	0.9963	0.9962	0.9862	0.995	0.9993
4	0.9974	0.9969	0.98681	0.9953	0.9995
5	0.9973	0.9976	0.9881	0.9955	0.9998
Aggregated Value	4.9838	4.9825	4.9346	4.9773	4.9964

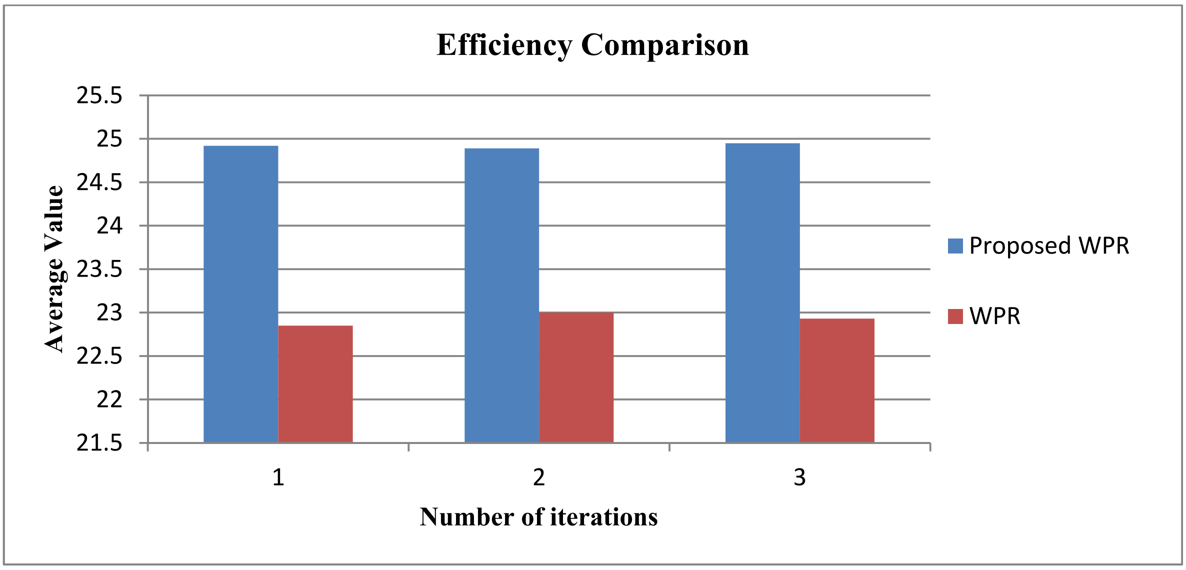

Figure 11 Comparison between proposed and existing WPR model for efficiency.

Figure 11 shows the bar graph showing the comparative analysis between average efficiency values of existing WPR and proposed WPR. Normal efficiency has been calculated using three parameters (Entropy, Standard Deviation and Relevancy). Total 25 iterations has been performed on the defined data set to obtain the results for average efficiency using these three sub-parameters. Figure 11 shows comparison for three iterations showing that for every iteration proposed WPR has higher bar of average value than existing WPR. Hence the proposed system is more accurate in assigning the ranks to each web page using both backlinks and inlinks. Proposed system covers all the pages more efficiently to calculate the rank by scanning the content as well as user behavior.

V. CONCLUSION AND FUTURE WORK

Web mining lies at the core of Page Rank calculation. Page Rank is a link analysis algorithm which is the heart of Google. It represents the importance of a web page by counting the number of backlinks. It is non keyword specific and link structure based and hence signifies that the importance of a web page is directly proportional to the number of web pages linked to it. It is only associated with the individual web page and not the entire website. Page rank uniformly divides the page rank score equally among all its outlinks. Therefore on the whole a page has a high rank if the sum of the ranks of its backlinks is high. Weighted Page Rank is an extension to the standard Page Rank algorithm which resolves the main problem of rank sink present in Page rank. It takes into account both forward as well as backlinks by assigning weight to both of them. Weighted Page rank algorithm assigns page rank score in a non uniform fashion and gives utmost preference to more popular web pages. Although Weighted Page Rank algorithm resolves the problem of rank sink, it still possess average efficiency and lesser relevancy to Page Rank's very less efficiency and completely ignored relevancy. More relevancy implies more consistency of results which in turn signifies much accurately working of website priority tool. Higher accuracy of website priority tool leads to higher precision. Hence standard existing Weighted Page Rank algorithm has been improved significantly by considering two parameters namely efficiency and precision. Efficiency of web pages is the sum total of three sub parameters entropy, standard deviation and relevancy. Precision based evaluation of Page rank and improved Weighted Page Rank has been done with reference to a standard website priority tool. Average efficiency values of existing Weighted Page Rank and proposed Weighted Page Rank has been compared with the help of a bar graph. Average efficiency has been calculated by performing approximately 25 iterations(testing rounds) on the data set of web pages taken from different e-books. Performance of the proposed improved WPR algorithm has been discussed with the help of wampserver 2.4, MATLAB R2013A The Math Work Inc. and My Sql database 5.0.

In future an algorithm could be made that works on the merits of both Page rank and Weighted Page rank which will be based on both link structure as well as content of the web pages. Hence the new hybrid algorithm made will compute the rank score of the web pages at both query as well as indexing time. Assuming this proposition the hybrid algorithm will fetch most relevant web pages from the search process with an excellent efficiency.

REFERENCES

[1] Manika Dutta and K. L Bansal, "A Review Paper on Various Search Engines(Google, Yahoo, Altavista, Ask and Bing)" , International Journal on Recent and Innovation Trends in Computing and Communication, Vol. 4, Issue 8, August 2016.

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

[3] Sepandar Kamvar, Taher Haveliwala and Gene Golub, "Adaptive methods for the computation of PageRank" , Linear algebra and its Applications 386, ELSEVIER, pages 51-65, 2004.

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

[5] Dilip Kumar Sharma and A.K Sharma, "A Comparative Analysis of Web Page Ranking Algorithms", International Journal on Computer Science and Engineering , Vol.02, Issue 08, 2010.

[6] Arun Kumar Singh, Avinav Pathak and Dheeraj Sharma, " A Survey on Enhancing the Efficiency of Various Web Structure Mining Algorithms" , International Journal of Computer Applications Technology and Research, Vol. 2, Issue 6, pages 771-774.

[7] Alexandrous Ntoulas, Junghoo Cho and Christopher Olston, " What's New on the Web? The Evolution of the Web from a Search Engine Perspective" , WWW2004, May 17–22 , 2004, New York, New York, USA. ACM 1-58113-844-X/04/0005.

[8] Atul Kumar Srivastva, Mitali Srivastva, Rakhi Garg and P.K Mishra, "Comparative Study of Web Page Ranking Algorithms", International Journal of Emerging Technologies in Computational and Applied Sciences, 7(1), Dec 2013-Feb 2014, pp.26-32.

[9] Rekha Jain, Sulochna Nathawat and G.N Purohit, "Enhanced Retrieval of Web Pages using Improved Page Rank Algorithm", International Journal on Natural Language Computing, Vol. 2, No. 2, April 2013.

[10] Nidhi Shalya, Shashwat Shukla and Deepak Arora, "An Effective Content Based Web Page Ranking Approach", International Journal of Engineering Science and Technology, Vol.4, No.08, August 2012.

[11] Satish Muppidi and Venkata Naveen Koraganji, "Survey of Contemporary Ranking Algorithms", International Journal of Applied Engineering Research, Vol.11, No.1(2016).

[12] Kaushal Kumar, Abhaya and Fungayi Donewell Mukoko, "PageRank algorithms and its variations", IOSR Journal of Computer Engineering(IOSR-JCE), Vol.14, Issue 1, pages 38-45, sep-oct 2013.

[13] Aditya Murgai and Naresh Sharma, “Page Rank Algorithm Expressed in Terms of Link Distance and a Modified Procedure of Page Rank Calculation”, 3rd International Conference on Computer Modeling and Simulation, 2011.
[14] Hema Dubey and B.N Roy, “An Improved Page Rank Algorithm based on Optimized Normalization Technique”, International Journal of Computer Science and Information Technologies, Vol.2(5), 2011, pp.2183-2188.
[15] Rashmi Rani and Vinod Jain, “Weighted Page Rank using the Rank Improvement”, International Journal of Scientific and Research Publications , Vol.3, Issue 7, June 2013.
[16] Nagappan V.K and Dr. P. Elango, “Agent Based Weighted Page Ranking Algorithms for Web Content Information Retrieval”, International Conference on Computing and Communication Technologies (ICCCT’15) , 978-1-4799-7623-2/15 Copyright 2015 IEEE.
[17] Sanjay and Dharmender Kumar, “A Review Paper on Page Ranking Algorithms”, International Journal of Advanced Research in Computer Engineering & Technology, Vol.4, Issue 6, June 2015.
[18] Neha Verma, Dheeraj Malhotra, Monica Malhotra and Jatinder Singh, “ E-commerce website ranking using semantic web mining and neural computing” , International Conference on Advanced Computing Technologies and Applications(ICACTA-2015), Procedia Computer Science 45, pages 42-51, 2015.