Survey on Various Ranking Algorithms

Manpreet Kaur Gill[#1], Nidhi Bhatla[*2]

[#]*M.Tech, Research Scholar,*
Department of Computer Science and Engineering,
RIMT College of Engineering & Technology, Mandi Gobindgarh, Fatehgarh Sahib, Punjab, India
[1]manpreetgill49@gmail.com

[*]*Assistant Professor,*
Department of Computer Science and Engineering,
RIMT College of Engineering & Technology, Mandi Gobindgarh, Fatehgarh Sahib, Punjab, India
[2]engineernidhi@yahoo.com

***Abstract*— World Wide Web contains distributed heterogeneous information in which number of data and hyperlinks are there. In today's world it becomes more important for authors to have their papers and articles well ranked in these search engines in order to meet the requirements of the users. There is a knowledge about the ranking algorithms is essential to authors in respect to the uniqueness of the results retrieved by the user query. Web page ranking algorithms play an important role in ranking web pages so that the user could retrieve the page which is most relevant to the user's query. In this paper, different Page Ranking algorithms like PageRank (PR), WPR (Weighted PageRank), HITS (Hyperlink- Induced Topic Search) and WPCR (Weighted Page Content Rank) algorithms are discussed. There are ranking algorithms such as PageRank and Weighted PageRank in which Web Structure Mining technique is commonly used. On the other hand, Weighted Page Content Rank based on both web content mining and web structure mining that shows the relevancy of the pages to a given query is better as compared to the PageRank and Weighted PageRank algorithms. The paper focuses on their strengths, weaknesses, variations, and carefully analyses these algorithms to find out the further scope of research in web page ranking algorithm.**

***Keywords*— Data mining, Web mining, Text mining, Weighted Page content rank, Weighted Page rank, HITS, Page Rank.**

I. INTRODUCTION

Data mining is the process of analyzing the data from different data resources and summarizing them into useful information. To manage the data in a multi-dimensional database system, data mining consists of these steps: Extract, Transform, and Load transaction data onto the data warehouse system, and then Store. Data mining considered as synonyms for Knowledge Discovery in Database (KDD). Data mining is actually the part of the knowledge discovery (KDD) process. Data mining is a process which can be applied to a number of applications like weather forecasting, fraud detection, electric load prediction etc. Some data mining techniques are [1]:

- Statistics
- Association
- Classification
- Clustering
- Prediction
- Sequential patterns
- Decision tree.

A. Text Mining

Text mining is the process of discovery of text from the text documents or interesting knowledge. It is a challenging task to help the users in finding what the user' actually want from the number of text documents. It is quite difficult to deal with the text which is in unstructured form. The purpose of the text mining is to finding "nuggets" of interesting information from the natural language text [2]. To answer the complicated questions and to do the web searches with intelligence is the main aim of the text mining tools. Text mining uses the automated methods for achieving the unusually knowledge which is available in text documents. Techniques of text mining are [3]:

- Natural Language Processing (NLP)
- Information Extraction (IE).

B. Web Mining

Web mining comes under the application of data mining. It refers to overall process of finding the useful and previously unknown information from the web services and documents.

The steps as shown in Fig. 1 are involved in the web mining process [4]:

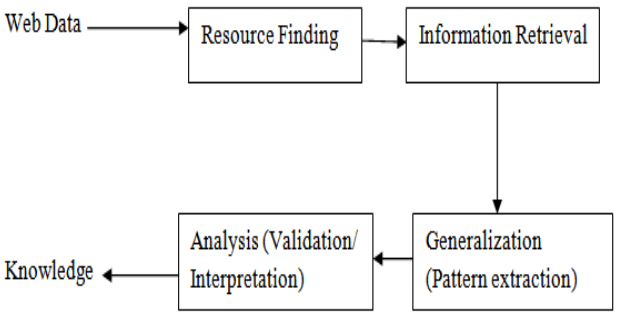

Fig. 1 Web mining process

1) Resource Finding: It is the function for retrieving the relevant web documents.

2) Information Selection and Pre-processing: In this, the selection and preprocessing of information is automated.

3) Generalization: In this, general patterns are automatically discovered at single website as well as across multiple websites. In generalization machine learning and data mining techniques are used to find the patterns.

4) Analysis: In this, the validation and interpretation is done for the pattern mining.

There are three categories of web mining as shown in Fig. 2 which are: Web Usage Mining, Web Content Mining, and Web Structure Mining [4] [5].

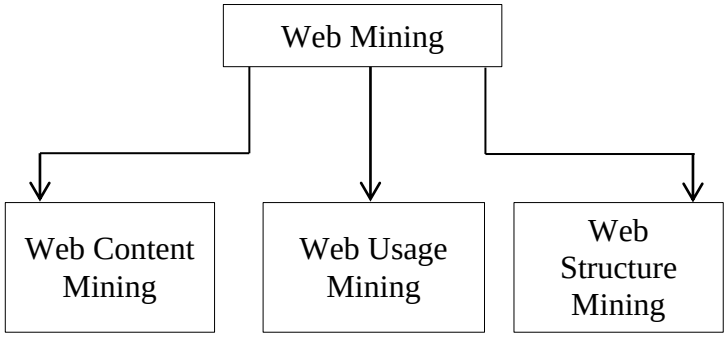

Fig.2 Web Mining Categories

1) Web Usage Mining (WUM): Web usage mining is used to find usage patterns from Web data in order to understand Web based applications. It consists of three phases which are pattern analysis, pattern discovery, and preprocessing. In this, servers, proxies, and client applications can easily capture data about Web usage.

2) Web Content Mining (WCM): Web Content Mining is the process of extracting meaningful information according to the contents of Web documents. Content data consists of facts like images, texts, audio, video, HTML documents, and data in tables that are included in a web page. The individual pages which are mined in web content mining is the primary Web resources can be used to grouping, categorizing, analyzing, and retrieving documents.

3) Web Structure Mining (WSM): The aim of web structure mining is to create a structural summary about the web page and website. In this, web structure mining is used to extract the patterns in the web from hyperlinks. To connect the web page to another location hyperlink is used which is a structural component.

II. Web Search Engine

Search engines provide the way to the users to find specific information on the vast World Wide Web. The term web search engine is used in relation to Web. It consists of three elements which consist of important information for search engine are the user search, presentation and ranking of results, discovery & the database.

In this, the components of search engine architecture are as under as shown in Fig.3 [6]:

1) Crawlers: A crawler is a computer program which is used to create entries in the search engine index by visiting the websites to read their content and other information. They are mainly used to visit the sites that are either new websites or updated websites. Crawlers gained this name because at a time they crawls a page as well as the links to other pages through website until all pages have been read.

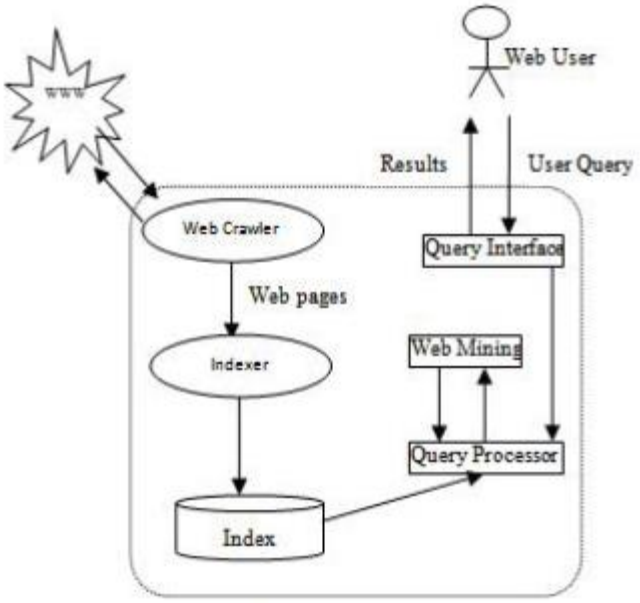

Fig.3 Working Architecture of a Search Engine

2) Indexer: The pages collected by the crawler are processed and indexed by the indexer. It saves the URL of all pages to which each word belongs after the extraction of keywords.

3) Query processor: The query engine receives the query which is given by the user and when it matches the query keywords with index, it provides the URL's of the pages to the users.

4) WWW: The World Wide Web is the users and resources using the HTTP protocol over the internet. The World Wide Web (WWW) is a system of interlinked documents accessed through the Internet. Web may contain video, images, audio, text, and other multimedia applications which navigates between them using hyperlinks.

III. Link Analysis Algorithms

In the development of web search Link analysis the analysis of hyperlinks and the graph structure of the Web have been helpful which is one of the factors considered by web search engines in computing a composite rank for a web page on any given user query. The directed graph configuration is

known as web graph. There are several algorithms based on link analysis. The important algorithms are Hypertext Induced Topic Search (HITS), Page Rank (PR), Weighted Page Rank (WPR), and Weighted Page Content Rank (WPCR).

A. HITS Algorithm

HITS stands for Hypertext Induced Topic Search. It is a link based algorithm which is used to rank pages that are retrieved from the web according to the given user query based on their textual information. When the user retrieved the required pages then HITS algorithm started ignoring textual information and starts focusing only on the web structure.

In [13], the algorithm is used to rank the relevant pages and treat all the links equally for the distribution of rank scores. In this, HITS rank the pages by analysing their in-links and out-links. The web pages that points to the hyperlinks are known as hubs but the hyperlinks that points to the web pages are known as authorities. Let ap and hp represent the authority and hub scores of page p, respectively. B (p) and I (p) denote the set of referrer and reference pages of page p, respectively. The scores of hubs and authorities are calculated as follows:

$$ap = \sum_{q \varepsilon B(p)} hq$$

$$hq = \sum^{n}_{q \varepsilon I(p)} aq$$

In [14], HITS provides the root set by taking top n pages then incorporates the root set with the webpages linked to it and webpages linked from it to form the base set. The webpages and links included in the base set are used to form sub graph. In this paper, they gave the pseudo code for the HITS algorithm which is used to normalised the hub and authority values after each step by dividing each authority value by the square root of the sum of the squares of all authority values, and dividing each hub value by the square root of the sum of the squares of all hub values.

In [15], when a user gave a query then HITS first creates the neighbourhood graph for the user query. This neighbourhood graph contained top 200 pages retrieved from the web search engine. It also contained the nearly top 200 pages linked by the retrieved 200 web pages and web pages that linked to these 200 top pages. Then hub and authority values are calculated.

1) Advantages: The following are the advantages of the algorithm are as under [14]:

- HITS algorithm is a sensitive to user query as compared to page rank. Authority and hub values are obtained on the basis of the retrieved important pages.
- HITS algorithm produces web graph from the pages that are retrieved to a user query.
- HITS algorithm calculates the hub and authority values accurately.

2) Disadvantages: There are number of advantages exists of this algorithm, but beside of the advantages there are some of the disadvantages of the algorithm which are shown as under [14]:

- The major drawback is that HITS is a query dependent algorithm because it takes more time for query evaluation. When user creates a new web page then user links his/her page with another web page because user assumes that he links the relevant page with his/her page. But sometimes it's irrelevant authority page which is also a major drawback of HITS.
- To obtain relevant pages, HITS works only on traditional search engine due to which it's not feasible with today's search engine.

B. Page Rank Algorithm

It is the most commonly used algorithm for ranking. The working of this algorithm is depends upon the link structure of the web pages. In this if there are important links towards a page then the links towards the other pages from it is also considered as important. In this back link is used to provide the rank score and if the rank of the back links is large then the rank given is large rank of particular page [16]. It is used by the Google Internet search engine which is named after Larry Page. It assigns the page rank to every web page while operating on web which means it is query independent [14]. Many factors determine the ranking of Google search results but PageRank continues to provide the basis for all of Google's web search tools. In this algorithm, web structure mining is used to get the relevant pages on the top by providing rank to the different web pages. It is not only the algorithm used by Google to order search engine results, but it is the first algorithm used by company and it is best known [17].

In [14] [16], simplified version of page rank is given as:-

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)},$$

where the PageRank value for a web page u is dependent on the PageRank values for each web page v out of the set Bu (this set contains all pages linking to web page u), divided by the number L(v) of links from page v. In [13], among the outgoing links the rank score of a page is divided. Values given to the outgoing links of page are used to calculate the ranks of the pages to which the given page is pointing.

In [12], the page rank algorithm was implemented in order to rank the web pages. But in this number of iterations for calculating page rank is more which increases the time complexity. Then to reduce the number of iterations they proposed the enhanced version of algorithm proposed page rank algorithm which also reduces the time complexity.

In [14], the running time of the page rank algorithm are taken three factors into account which are number of iterations (i), number of web pages (n) and number of outgoing edges of each web page (O_n). The PageRank algorithm provides ranked pages in the sorting order to users based on the given query.

1) Advantages: Some of the advantages of the algorithm are given under as [14] [12]:

- In page rank algorithm, the query time cost is less as compared to the HITS algorithm.
- It is more efficient and feasible than HITS as it performs computations during crawl time instead of query time.
- In this, the values are assigns to every document independent of query i.e. it is a query independent algorithm.
- It is Content independent Algorithm.
- Page Rank is depends upon the linking structure of the web Page.

2) Disadvantages: There are the following disadvantages of the algorithm are [14]:

- Spider trap is a group of pages if there is no link exists within the group or to the outside group then spider trap problem occurs.
- When infinite link cycles occur in pages in a network then rank sink problem occurs as shown in Fig. 4.

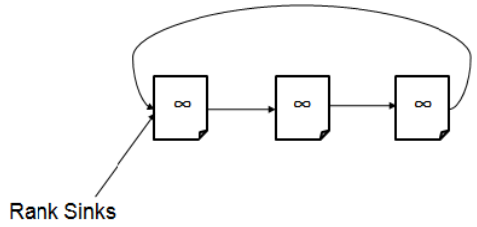

Fig. 4 Illustration of Rank Sink

- If there are circular references in the website then this algorithm reduces the page rank of front pages shown in fig. 5.

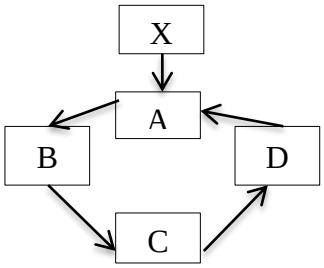

Fig. 5 Illustration of Circular References

- When a page links to another page which has no further outgoing links, such link is known as dangling link.
- Dead Ends: The pages with no outgoing links are known as dead ends.

C. Weighted Page Rank Algorithm

It is the extension of the original page rank algorithm In this larger rank values are assigned to more important pages instead of dividing the rank value of a page equally among its outgoing linked pages. The value given to the outgoing links is proportional to the importance of that page [13]. WPR performs better than the conventional Page Rank algorithm in terms of returning larger numbers of relevant pages to a given query. The popularity from the number of in links and out links is recorded as $W^{in}(v,u)$ and $W^{out}(v,u)$, respectively.

$W^{in}(v,u)$ is the weight of $link(v, u)$ calculated based on the number of in links of page u and the number of in links of all reference pages of page v.

$$W^{in}(v,u) = \frac{I_u}{\sum_{p \in R(v)} I_p}$$

where Iu and Ip represent the number of in links of page u and page p. The reference page list of page v is denoted by R(v).

$W^{out}(v,u)$ is the weight of link(v, u) calculated based on the number of out links of page u and the number of out links of all reference pages of page v [14].

$$W^{out}(v,u) = \frac{O_u}{\sum_{p \in R(v)} O_p}$$

where Op and Ou represent the number of out links of page p and page u, respectively. The reference page list of page v is denoted by R(v).To calculate the rank in WPR algorithm:

$WPR(u) = 1-d + d \sum_{v \in B(u)} PR(v)\ Win(v,u)\ Wout(v,u)$ [14]

In [13], the proposed system represents the evaluation of weighted page algorithm in which the work carried out as:

- In this, website with number of hyperlinks is necessarily founded because this algorithm relies on web structure.
- After this a required web map is generated for the selected website.
- Then a set of relevant pages with the given user query is retrieved using the search engine. The relevant set of pages is called the root set.
- The pages are directly point to or are pointed to by the pages in the root set creates a base set.
- Then this algorithm is applied on the base set and after improving the rank, the algorithm is evaluated by comparing their results.

In [15], weighted page rank algorithm with k-means algorithm is used in order to reduce the execution time. In this algorithm the user get the relevant and important pages easily as it employs web structure mining. In the proposed work, firstly create a database at first end in which the number of records of sites which we want to cluster with records are included. Next the software has been used for carried out the mode of operation. Then in this the pre-processing of records after loading them in main database occurred. This pre-

processing is carried out with clustering techniques which cluster the database in sets of datasets which are easy for processing. Then implication takes place in which proposed algorithms are implemented i.e. K-means with weighted page algorithm. Data patterns are analysed which is efficient in terms of time taken with regard to previous research's.

In [16], the work was proposed in which weighted page algorithm with K-means algorithm is applied to reduce the execution time and to improve the performance. In this, the above mentioned algorithms are implemented step by step then on the basis of comparison result sets are undertaken of previous research with proposed method. The proposed work was done on the basis of the websites in which rank given to the sites by applying this algorithm. In the graph shown in fig. 6, we compare the results taken from the previous work done and where (1,2,3……,10) are the ID's of the entered web links and accordingly the execution time (microseconds) for calculating ranks are shown:

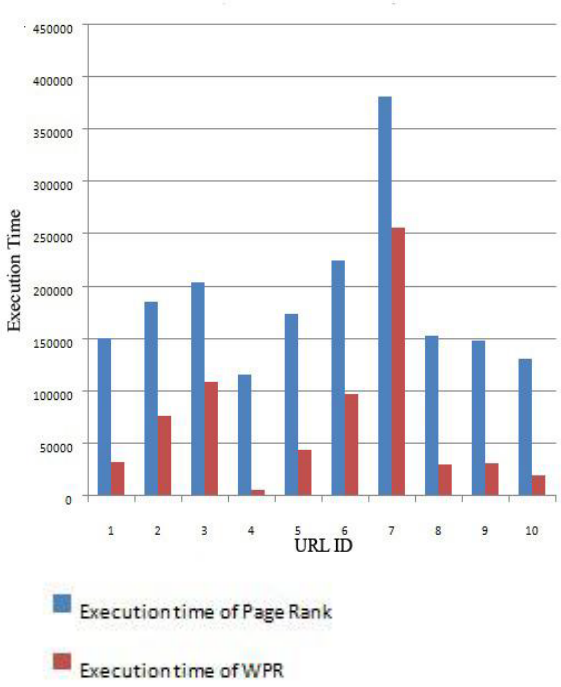

Fig. 6 Graph of results comparison

From the above graph, they conclude that the execution time for the weighted page rank (proposed work) for calculating the rank of the page is less than the execution time for the page rank (previous work).

1) Advantages: Some of the advantages of the algorithm are given under as [17]:

- Quality: The quality of the pages returned by this algorithm is high as compared to page rank algorithm.
- Efficiency: According to importance of the page rank value of a page is divided among it's out link of that page. So, it is more efficient than page rank algorithm.

2) Disadvantages: Despite of some of the advantages of the algorithm there are some disadvantages are given under as [17]:

- Less Relevant: It returns less relevant pages to user query as this algorithm considers only link structure not the content of the pages.
- Popularity: It is based only on the popularity of the web page.

D. Weighted Page Content Rank Algorithm

Weighted Page Content Rank Algorithm (WPCR) is a proposed page ranking algorithm in which according to a user query a sorted order to the web pages returned by a search engine. WPCR is a numerical value based algorithm on which the web pages are given an order. This algorithm considers web structure mining as well as web content mining as their main techniques whereas in weighted page ranking only web structure mining technique is used. Web structure mining is used to calculate the importance of the page and web content mining is used to find how much a page is relevant. Importance means the popularity of the page which means how much number of pages is referred by or is pointing to a particular page. It can't be calculated with the help of in links only, out links are also to be considered. The matching of the page with the user query shows the relevancy of the page. The page is more relevant if it maximally matched to the user query [6].

Algorithm: WPCR calculator
Input: Page P, in link and Out link Weights of
All back links of P, Query Q, d (damping factor).
Output: Rank score
Step 1: Relevance calculation:
1) Find all meaningful word strings of Q (say N)
2) Find whether the N strings are occurring in P or not?
Z= Sum of frequencies of all N strings.
3) S= Set of the maximum possible strings occurring in P.
4) X= Sum of frequencies of strings in S.
5) Content Weight (CW) = X/Z
6) C= No. of query terms in P
7) D= No. of all query terms of Q while ignoring stop words.
8) Probability Weight (PW) = C/D
Step 2: Rank calculation:
1) Find all back links of P (say set B).
2) PR (P) = (1-d) +d
3) Output PR (P) i.e. the Rank score.

In this algorithm, the input parameters used in Page Rank are Back links, Weighted Page Rank uses Back links and Forward Links as Input Parameter and Weighted Page Content Rank uses Back links, Forward Link and Content as Input Parameters. WPCR algorithm uses content also as an input parameter [15].

1) Advantages: There are some of the advantages of the algorithm are given under as:

- More Relevant: It considers the content of the pages along with the links in order to find the more relevant pages.
- Popularity and Content: It is based on both the popularity as well as the content of the web page.

IV. Comparison Of Page Ranking Algorithms

The comparison of various web page ranking algorithms is shown in Table 1. The comparison is done on the basis of factors such as quality of results, methodology, key in parameter, main technique use, relevancy, and, importance and limitations.

TABLE I

Comparison Between Hits, Page Rank, Weighted Page Rank, Weighted Page Content Rank Algorithms

ALGORITHM	HITS	PAGE RANK	WEIGHTED PAGE RANK	WEIGHTED PAGE CONTENT RANK
Main Technology	Web structure mining, Web content mining	Web structure mining	Web Structure Mining	Web Content Mining
Input Parameter	Content, Back and Forward links	Back Links	Back Links and Forward Links	Content as well as links
Relevancy	More	Less	Less	More
Methodology	It computes the hubs and authorities of the relevant pages.	It computes the score for the pages at the time of indexing of pages.	Weight of web page calculated on the basis of input and outgoing links and on the basis of weight the importance of page is decided.	It computes the rank of the pages on the basis of the content of the page.
Quality of results	Less than Page Rank	Medium	Higher than Page Rank	Highest

V. Conclusion

In this paper, various ranking algorithms for web pages are discussed. An overview of various algorithms on the basis of performance metrics which is used for evaluation is presented. There are various ranking algorithms such as HITS, Page Rank, and Weighted Page Rank, Weighted Page Content Rank etc. which provides relevant result to user query easily and fatly. In Page Rank and Weighted Page Rank algorithms, main technique used is web structure mining but in Weighted Page Content Rank web structure mining as well as web content mining is used as main technique. The input parameters used in Page Rank are back links, Weighted Page Rank uses back links and forward links as input parameter and Weighted Page Content Rank uses back links, forward link and content as input parameters. Weighted Page Content Rank algorithm provides better result than other in terms of performance measures and also provides relevant result to user's query easily.

Acknowledgement

The authors wish to thank the reviewers and editors for their suggestions and constructive comments that help in bringing out the useful information and improve the content of paper.

References

[1] Madhuri V. Joseph, Lipsa Sadath, Vanaja Rajan, "*Data Mining: A Comparative Study on Various Techniques and Methods*", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), ISSN: 2277 128X, Volume 3, Issue 2, February 2013, pp. 106-113.

[2] Divya Nasa, "*Text Mining Techniques- A Survey*", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), ISSN: 2277 128X, Volume 2, Issue 4, April 2012, pp. 50-54.

[3] Shaidah Jusoh, Hejab m. Alfawarch, "T*echniques, Applications and Challenging Issue in Text Mining*", International Journal of Computer Science Issues (IJCSI), ISSN (Online): 1694-0184, Vol. 9, Issue 6, No 2, November 2012, pp. 431-436.

[4] Shruti Aggarwal, Gurpreet Kaur, "*Improving the Efficiency of Weighted Page Content Rank Algorithm using Clustering Method*" International Journal of Computer Science & Communication Networks (IJCSCN), ISSN:2249-5789, Vol 3(4), pp. 231-239.

[5] T.Munibalaji, C.Balamurugan, "*Analysis of Link Algorithms for Web Mining*", International Journal of Engineering and Innovative Technology (IJEIT), ISSN: 2277-3754, Volume 1, Issue 2, February 2012, pp-81-86.

[6] Pooja Sharma, Deepak Tyagi, Pawan Bhadana, "Weighted Page Content Rank for Ordering Web Search Result", International Journal of Engineering Science and Technology (IJEST), ISSN: 0975-5462, Vol. 2 (12), 2010, pp. 7301-7310.

[7] Rashmi Rani, Vinod Jain, "*Weighted PageRank using the Rank Improvement*", International Journal of Scientific and Research Publications, IJSRP, ISSN: 2250-3153, Vol. 3, Issue 7, July 2013.

[8] Nidhi Grover and Ritika Wason, "*Comparative Analysis Of Pagerank And HITS Algorithms*", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 1 Issue 8, October – 2012.

[9] Debajyoti Mukhopadhyay, Pradipta Biswas, Young-Chon Kim, "*A Syntactic Classification based Web Page Ranking Algorithm*," 6th International Workshop on MSPT Proceedings, MSPT 2006, pp-87-92.

[10] Dilip Kumar Sharma, A. K. Sharma, "*A Comparative Analysis of Web Page Ranking Algorithms*," International Journal on Computer Science and Engineering, IJCSE, ISSN 0975-3397, Vol. 02, No. 08, 2010, pp-2670-2676.

[11] Ayman Farahat, Thomas Lofaro, Joel C. Mille, Gregory Rae, And Lesley A. Ward, "*Authority Rankings From Hits, Pagerank, And Salsa: Existence, Uniqueness, And Effect Of Initialization*," Siam J. Sci. Comput., Vol. 27, No. 4, pp. 1181–1201.

[12] Hema Dubey ,Prof. B. N. Roy, "*An Improved Page Rank Algorithm based on Optimized Normalization Technique*", International Journal of Computer Science and Information Technologies, IJCSIT, ISSN:0975-9646, Vol. 2 (5) , 2011, pp-2183-2188.

[13] Seifedine Kadry and Ali Kalakech, "*On the Improvement of Weighted Page Content Rank*", Journal of Advances in Computer Networks,

DOI: 10.7763/JACN.2013.V1.23, Vol. 1, No. 2, June 2013, pp-110-114.

[14] Wenpu Xing and Ali Ghorbani, "*Weighted PageRank Algorithm*", Proceedings of the Second Annual Conference on Communication Networks and Services Research, IEEE, 2004.

[15] Supreet Kaur, Usvir Kaur, "*An Optimizing Technique for Weighted Page Rank with K-Means Clustering*", International Journal of Advanced Research in Computer Science and Software Engineering, IJARCSSSE, ISSN: 2277 128X, Volume 3, Issue 7, July 2013, pp. 788-792.

[16] Amar Singh, Navjot Kaur, "*To Improve the Convergence Rate of K-Means Clustering Over K-Means with Weighted Page Rank Algorithm*", International Journal of Advanced Research in Computer Science and Software Engineering, IJARCSSSE, ISSN: 2277 128X, Volume 3, Issue 8, August 2013, pp. 143-148.

[17] Taruna Kumari, Ashlesha Gupta, Ashutosh Dixit, "*Comparative Study of Page Rank and Weighted Page Rank Algorithm*", International Journal of Innovative Research in Computer and Communication Engineering, IJIRCCE, ISSN(Online): 2320-9801, ISSN (Print): 2320-9798, Vol. 2, Issue 2, February 2014, pp-2929-2937.

