# Comparative study of various Page Ranking Algorithms in Web Structure Mining (WSM)

**Neelam Tyagi, Simple Sharma**

***Abstract— As the web is escalating day by day, so people rely on the search engines to investigate the web. In this situation, the challenge for website owner is to provide relevant information to the users as per their needs and fulfill their requirements. The famous search engine Google used Hyperlink structure for ranking the web pages. There are various ranking algorithms are present for getting the desired result. The paper refers a preface to Web mining then trying to explain detailed Web Structure mining, and supply the link evaluation algorithms brought into play by the Web. This paper also explores different PageRank algorithms and compares those algorithms used for Information Retrieval. In Web Mining, the essentials of Web mining and the Web mining categories are explained. Different Page Ranking algorithms like PageRank (PR), WPR (Weighted PageRank), HITS (Hyperlink- Induced Topic Search) algorithms are discussed and comparison of these algorithms in context of performance has been carried out. PageRanks are designed for PageRank and Weighted PageRank algorithm for a agreed hyperlink composition.***

***Keywords — HITS, PageRank, Weighted PageRank, Web Structure.***

## I. INTRODUCTION

The World Wide Web (WWW) is trendy and interactive intermediary to telecast in turn these days. It is an enormous, contrary diverse, dynamic and mostly formless data warehouse. As on today WWW is the prevalent information depository for awareness indication. The subsequent challenges [1] in Web Mining are:

1) Web is enormous.
2) Web pages are partially structured.
3) Web information stands to be miscellany in meaning.
4) Degree of quality of the in sequence extracted.
5) Winding up of knowledge from information extracted.

It is predictable that WWW has lingering by about 2000% since its evolution and is replication in size every six to ten keywords with the catalog proceeds the URLs of the pages to the months [2]. With the swift augmentation of WWW and the user's stipulate on knowledge, it is becoming more difficult to deal with the information on WWW and gratify the user desires. Therefore, the users are in search of improved information repossession techniques and tools to position, extract, and filter and locate the essential information. Most of the users use information reclamation tools akin to search engines to find information from the WWW. There are tens and hundreds of search engines obtainable but some are popular like Google, Yahoo, Bing etc., because of their swarming and ranking methodologies. The search engines download, index and store up hundreds of millions of web pages. They response tens of millions of queries every day. So Web mining and ranking mechanism becomes very significant for effective information retrieval. The sample architecture [3] of a search engine is shown in Fig. 1.

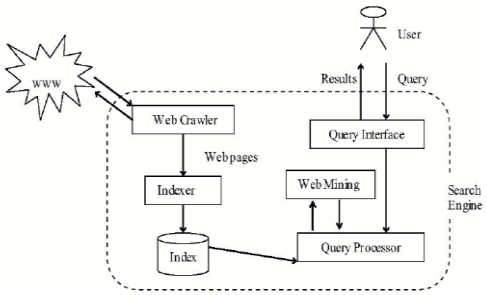


**Fig. 1 Sample Architecture of a Search Engine**

There are 3 vital components in a search engine known as Crawler, Indexer and Ranking mechanism. The Crawler is also called as a robot or spider that navigates the web and downloads the web pages. The downloaded pages are being transferred to an indexing module that parses the web pages and erect the index based on the keywords in individual pages. An alphabetical index is normally sustaining using the keywords. When a query is being floated by a user, it means the query transferred in terms of keywords on the interface of a search engine, the query mainframe section examine the query keywords with the index and precedes the URLs of the pages to the client. But before in presenting the pages to the client, a ranking mechanism is completed by the search engines to present the most relevant pages at the top and less significant ones at the substructure. It makes the search outcomes routing easier for the user. The ranking mechanism is clarified in detail later in this paper. This paper deliberates as tag along. Section II makes available the basic Web mining concepts and the three regions of Web mining. In this section Web Structure Mining is portrayed in detail as most of the Page Rank algorithms are based on the Web Structure Mining.

Section III explains a comprehensive synopsis of a number of link analysis algorithms. Section IV discusses the boundaries and potencies of each algorithm conversed. Finally in Section V the paper is accomplished with a beam on future implications.

Manuscript received on May 29, 2012.

**Neelam Tyagi**, Computer Science and Engineering, Manav Rachna International University, Faridabad, India.
**Simple Sharma**, Asstt. Prof(Computer Science and Engineering), Manav Rachna International University, Faridabad, India.

## II. WEB MINING

Web mining is the Data Mining technique that routinely determines or billets out the in order from web credentials. It is the extraction of appealing and latently useful patterns and implicit information from artifacts or movement associated to the World Wide Web.

### *A. Web Mining Process*

The absolute process of extracting knowledge from Web data [4] is follows in Fig.2:

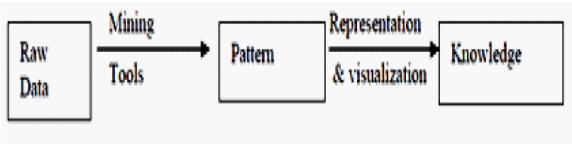


**Fig. 2 Web Mining Process**

The various steps are clarified next:

*1. Resource finding:* It is the task of retrieving intended web credentials.

*2. Information selection and pre-processing:* Robotically selecting and pre- processing definite from information retrieved Web resources.

*3. Generalization:* Robotically ascertains general patterns at individual Web site as well as multiple situates.

*4. Analysis:* Rationale and interpretation of the mined patterns.

### B. Web Mining Categories

Web mining can be categorized into three categories [4, 5]: Web content mining (WCM), Web structure mining (WSM), and Web Usage Mining (WUM) as shown in fig 3. Web content mining refers to the finding of applying information from web contents, including text, image, audio, video, etc. Web structure mining lessons the web's hyperlink structure. It usually involves analysis of the in-links and out-links of a web page, and it has been used for search engine result ranking. Web usage mining focuses on analyzing search logs or other activity logs to find appealing patterns. One of the main applications of web usage mining is to learn user summaries.

*Definitions:*

### A. Web Content Mining (WCM)

Web Content Mining is the progression of extracting useful information from the contents of web credentials. The web credentials may consists of text, images, audio, video or structured records like tables and lists. Mining can be purposeful on the web documents as well the results pages

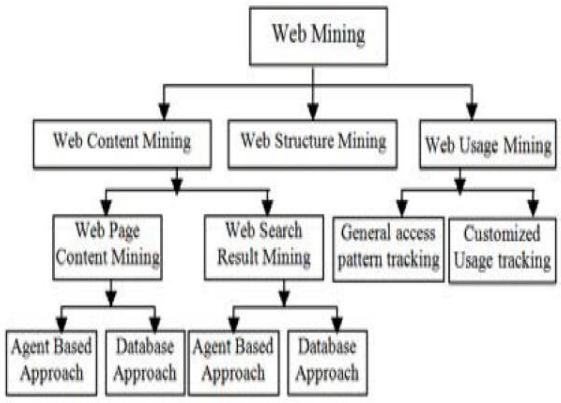


**Fig 3 Shows the general classification of Web Mining.**

fashioned from a search engine. There are bi approaches in content mining called agent based approach and database based approach. The agent based approach focus on searching appropriate information using the uniqueness of a particular domain to interpret and organize the collected information. The database approach is used for get back the semi-structure data from the web.

### *B. Web Usage Mining (WUM)*

Web Usage Mining is the method of hauling out useful information from the secondary data consequent from the interactions of the user while surfing on the Web. It extracts data accumulated in server access logs, referrer logs, agent logs, client-side cookies, user profile and Meta data.

### *C. Web Structure Mining (WSM)*

The aim of the Web Structure Mining is to generate the structural abstract about the Web site and Web page. It tries to determine the link structure of the hyperlinks at the inter document level. Basic foundation on the topology of the hyperlinks, Web Structure mining will categorize the Web pages and spawn the information like similarity and relationship between different Web sites. This type of mining can be carried out at the document level (intra-page) or at the hyperlink level (interpage). It is important to appreciate the Web data structure for Information Retrieval.

The three categories of web mining described above have its own appliance areas including site improvement, business intelligence, web personalization, site modification, usage characterization and classification, ranking of pages etc. The page ranking algorithms are generally used by search engines to find more important pages. Different page ranking algorithms are discussed in the next section which describes the functioning of these algorithms. Three important page ranking algorithms are: PageRank, Weighted PageRank, HITS which are discussed in next section.

## III. LINK ANALYSIS ALGORITHMS

Web mining technique provides the additional information through hyperlinks where different documents are associated [6].

We can examine the web as a directed labeled graph whose node are the credentials or pages and edges are the hyperlinks between them. This directed graph configuration is known as web graph.

There are several algorithms proposed based on link analysis. Three important algorithms PageRank[7], Weighted PageRank[8] and HITS (Hyper-link Induced Topic Search)[9] are discussed below.

### A. PageRank Algorithm

Brin and Page [7] developed PageRank algorithm at Stanford University based on the mention analysis. PageRank algorithm is used by the famous search engine, Google. PageRank algorithm is the most frequently used algorithm for ranking the various pages. Functioning of the Page Rank algorithm depends upon link structure of the web pages. The PageRank algorithm is based on the concepts that if a page surrounds important links towards it then the links of this page near the other page are also to be believed as imperative pages. The Page Rank reflects on the back link in deciding the rank score. Thus, a page gets hold of a high rank if the addition of the ranks of its back links is high. A simplified version of PageRank is given in Eq. 1:

$$PR(u)=C\sum_{v\epsilon B(u)}\frac{PR(v)}{N_v} \qquad (1)$$

where u represents a web page, B(u) is the set of pages that point to u, PR(u) and PR(v) are rank achieves of page u and v respectively, $N_v$ indicates the number of outgoing links of page v, c is a factor applied for normalization.

Later PageRank was tailored observing that not all users follow the direct links on WWW. The modified version is given in Eq. 2:

$$\boldsymbol{PR(u)=(1-d)+d\sum_{v\epsilon B(u)}\frac{PR(v)}{N_v}} \qquad (2)$$

where d is a dampening factor that is frequently set to 0.85. d can be thought of as the prospect of users' following the direct links and (1 –d) as the page rank distribution from non- directly linked pages.

The PageRanks form a probability distribution over the Web pages, so the sum of all Web pages' PageRank will be solitary. PageRank can be intended using a simple iterative algorithm, and keeps up a correspondence to the principal eigenvector of the normalized link matrix of the Web.

Let us take an example of hyperlink structure of four pages A, B, C and D as shown in Fig. 4. The PageRank for pages A, B, C and D can be calculated by using (2).

### B. Weighted PageRank Agorithm

Wenpu Xing and Ali Ghorbani [8] projected a Weighted PageRank (WPR) algorithm which is an addition of the PageRank algorithm. This algorithm assigns a larger rank values to the more important pages rather than separating the
rank value of a page evenly among its outgoing linked pages.

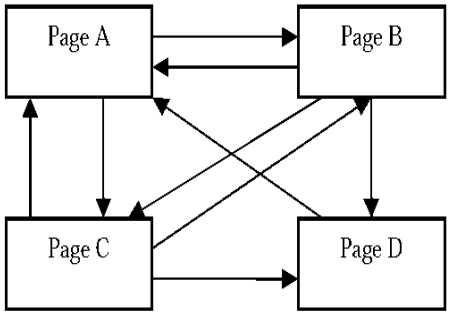


**Fig. 4 Hyperlink Structure for 4 pages PageRank Algorithm**

Each outgoing link gets a value proportional to its consequence. The importance is assigned in terms of weight values to the incoming and outgoing links and are denoted as $W^{in}_{(m,\ n)}$ and $W^{out}_{(m,\ n)}$ respectively. $W^{in}_{(m,\ n)}$ as shown in (3) is the weight of link (m, n) calculated based on the number of incoming links of page n and the number of incoming links of all orientations pages of page m.

$$\boldsymbol{W^{in}_{(m,n)}=\frac{I_n}{\sum_{p\epsilon R(m)} I_p}} \qquad (3)$$

Where $I_n$ and $I_p$ are the number of incoming links of page n and page p respectively. R (m) denotes the allusion page list of page m.

$$\boldsymbol{W^{out}_{(m,n)}=\frac{O_n}{\sum_{p\epsilon R(m)} O_p}} \qquad (4)$$

$W^{out}_{(m,n)}$ is as shown in (4) is the weight of link(m, n) calculated based on the number of outgoing links of page n and the number of outgoing links of all reference pages of m. Where $O_n$ and $O_p$ are the number of outgoing links of page n and p correspondingly. The formula as proposed by Wenpu et al for the WPR is as shown in (5) which is a modification of the PageRank formula.

$$WPR(n)=(1-d)+d\sum_{m\epsilon B(n)} WPR(m)\, W^{in}_{(m,n)} W^{out}_{(m,n)} \qquad (5)$$

### *A. Comparison of WPR and PR*

To compare the WPR from the standard PageRank, they classified the resultant pages of a query into four categories based on their relevancy to the given query. They are:

- *Very Relevant Pages (VR):* The pages containing very important information allied to a given query.
- *Relevant Pages (R):* Pages are relevant but not containing important information about a given query.
- *Weak Relevant Pages (WR):* Pages may have the query keywords but they do not have the appropriate information.
- *Irrelevant Pages (IR):* Pages not having any

relevant information and query keywords.

The PageRank and WPR algorithms both supply ranked pages in the categorization order to users based on the given query. So, in the resultant list, the integer of relevant pages and their order are very important for users. The following rule has been adopted a Relevance Rule to calculate the relevancy value of each page in the list of pages. That differentiates WPR from PageRank.

*B. Relevancy Rule*: The Relevancy Rule is as shown in equation (6). The Relevancy of a page to a given query depends on its group and its location and position in the page-list. The larger the relevancy value, the better is the result.

$$\boldsymbol{k} = \sum_{\boldsymbol{i \epsilon R(p)}} (\boldsymbol{n} - \boldsymbol{i}) * \boldsymbol{W}_{\mathrm{i}} \qquad (\mathbf{6})$$

Where i denote the ith page in the result page-list R (p), n represents the first n pages chosen from the list R (p), and $W_i$ is the weight of ith page as given below:

$$W_i = \{v1, v2, v3, v4\}$$

Where v1, v2, v3 and v4 are the values assigned to a page if the page is VR, R, WR and IR respectively. The values are always v1 > v2 > v3 > v4. Investigational studies explained that WPR produces larger relevancy values than the PageRank.

### C. HITS Algorithm

Kleinberg [9] developed a WSM based algorithm named Hyperlink-Induced Topic Search (HITS) which presumes that for every query given by the user, there is a set of authority pages that are relevant and accepted focusing on the query and a set of hub pages that contain useful links to relevant pages/sites including links to many authorities. Thus, fine hub page for a subject points to many authoritative pages on that content, and a good authority page is pointed by many fine hub pages on the same subject. Hubs and Authorities are shown in Fig. 5.

Kleinberg states that a page may be a good hub and a good authority at the same time. This spherical relationship leads to the definition of an iterative algorithm called HITS (Hyperlink Induced Topic Search).

The HITS algorithm treats WWW as a directed graph G(V,E), where V is a set of Vertices representing pages and E is a set of edges that match up to links.

#### *A. HITS Working Method*

There are two major steps in the HITS algorithm. The first step is the Sampling Step and the second step is the Iterative Step. In the Sampling step, a set of relevant pages for the

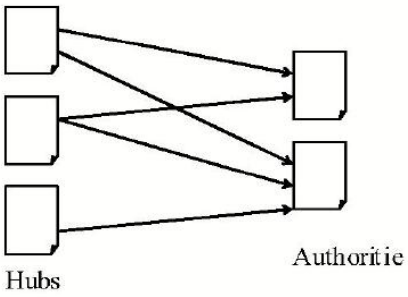


**Fig. 5 Hubs and Authorities**

given query are collected i.e. a sub-graph S of G is retrieved which is high in influence pages. This algorithm starts with a root set R, a set of S is obtained, keeping in mind that S is comparatively small, rich in relevant pages about the query and contains most of the good authorities. The second step, Iterative step, finds hubs and authorities using the output of the sampling step using (7) and (8).

$$\boldsymbol{H}_{\mathrm{p}} = \sum_{\boldsymbol{q \epsilon I(p)}} \boldsymbol{A}_{\mathrm{q}} \qquad (\mathbf{7})$$

$$\boldsymbol{A}_{\mathrm{p}} = \sum_{\boldsymbol{q \epsilon B(p)}} \boldsymbol{H}_{\mathrm{q}} \qquad (\mathbf{8})$$

Where $H_p$ is the hub weight, $A_p$ is the Authority weight, I(p) and B(p) denotes the set of reference and referrer pages of page p. The page's authority weight is proportional to the sum of the hub weights of pages that it links to it; similarly, a page's hub weight is proportional to the sum of the influence weights of pages that it links to. Fig. 6 shows an example of the calculation of authority and hub scores.

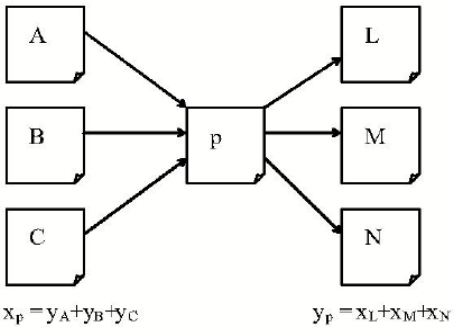


**Fig. 6 Calculation of hubs and Authorities**

### B. Constraints of HITS

The following are the constraints of HITS algorithm [10]:

- ***Hubs and authorities***: It is not simple to distinguish between hubs and authorities since many sites are hubs as well as authorities.
- ***Topic drift:*** Sometime HITS may not bring into being the most relevant documents to the user queries because of equivalent weights.

- ***Automatically generated links:*** HITS gives equivalent importance for robotically generated links which may not produce relevant topics for the user query.
- ***Efficiency:*** HITS algorithm is not efficient in real time.

A number of schemes like Probabilistic HITS, Weighted HITS etc. [11, 12, 13] have been proposed in the literature for modifying HITS.

## IV COMPARISION OF VARIOUS ALGORITHMS

On the basis of literature analysis, a comparison of various web page ranking algorithms is shown in Table 1. these Comparison is done on the basis of some vaults such as main technique use, methodology, key in parameter, relevancy, quality of results, importance and limitations. On the basis of parameters we can find the powers and limitations of each algorithm.

## V CONCLUSION

Web mining is the Data Mining technique that robotically discovers or extracts the information from web credentials. The standard search engines usually result in a large number of pages in response to users' queries, while the user always desires to get the best in a petite time. The page ranking algorithms, which are an application of web mining, play a major character in making the user search navigation easier in the results of search engine. The PageRank and Weighted Page Rank algorithm give importance to links rather than the content of the pages, the HITS algorithm anxieties on the content of web pages as well as links. Page Rank and Weighted Page Rank algorithms are used in Web Structure Mining to rank the relevant pages. After going through exhaustive analysis of algorithms for ranking of web pages against the various parameters such as methodology, input parameters, relevancy of results and importance of the outcome, it is concluded that on hand techniques have limitations particularly in terms of time response, accuracy of results, importance of the outcome and relevancy of results. An efficient web page ranking algorithm should meet out these challenges efficiently with compatibility with global principles of web technology.

**Table I Comparison of various Page Ranking Algorithms**

| Algorithm | Page Rank | HITS | Weighted Page Rank |
|---|---|---|---|
| **Main Technique** | Web Structure Mining | Web Structure Mining, Web Content Mining | Web Structure Mining |
| **Methodology** | This algorithm computes the score for pages at the time of indexing of the pages. | It computes the hubs and authority of the relevant pages. It relevant as well as important page as the result. | Weight of web page is calculated on the basis of input and outgoing links and on the basis of weight the importance of page is decided. |
| **Input Parameter** | Back links | Content, Back and Forward links | Back links and Forward links. |
| **Relevancy** | Less (this algo. rank the pages on the indexing time) | More (this algo. Uses the hyperlinks so according to Henzinger, 2001 it will give good results and also consider the content of the page) | Less as ranking is based on the calculation of weight of the web page at the time of indexing. |
| **Quality of results** | Medium | Less than PR | Higher than PR |
| **Importance** | High. Back links are considered. | Moderate. Hub & authorities scores are utilized. | High. The pages are sorted according to the importance. |
| **Limitation** | Results come at the time of indexing and not at the query time. | Topic drift and efficiency problem | Relevancy is ignored. |

## REFERENCES

1. M. G. da Gomes Jr. and Z.Gong, “Web Structure Mining: An Introduction”, Proceedings of the IEEE International Conference on Information Acquisition, 2005.
2. Naresh Barsagade, "Web Usage Mining And Pattern Discovery: A Survey Paper", CSE 8331, Dec.8,2003.
3. N. Duhan, A. K. Sharma and K. K. Bhatia, “Page Ranking Algorithms: A Survey”, Proceedings of the IEEE International Conference on Advance Computing, 2009.
4. Cooley, R, Mobasher, B., Srivastava, J."Web Mining: Information and pattern discovery on the World Wide Web”. In proceedings of the 9th IEEE International Conference on tools with Artificial Intelligence (ICTAI’ 97).Newposrt Beach,CA 1997.
5. R. Kosala, H. Blockeel, “Web Mining Research: A Survey”, SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining Vol. 2, No. 1 pp 1-15, 2000.
6. P Ravi Kumar, and Singh Ashutosh kumar, ”Web Structure Mining Exploring Hyperlinks and Algorithms for Information Retrieval”, American Journal of applied sciences, 7 (6) 840-845 2010.
7. S. Brin, and L. Page, “The Anatomy of a Large Scale Hypertextual Web Search Engine”, Computer Network and ISDN Systems, Vol. 30, Issue 1-7, pp. 107-117, 1998.
8. Wenpu Xing and Ali Ghorbani, “Weighted PageRank Algorithm”,

Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR '04), IEEE, 2004.

9. J. Kleinberg, "Authoritative Sources in a Hyper-Linked Environment", Journal of the ACM 46(5), pp. 604-632,1999.
10. S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, R. Kumar, P.Raghavan, S. Rajagopalan, A. Tomkins, "Mining the Link Structure of the World Wide Web", IEEE Computer Society Press, Vol 32, Issue 8 pp. 60 – 67, 1999.
11. D. Cohn and H. Chang, "Learning to probabilistically identify Authoritative Documents". In Proceedings of 17th International Conf. on Machine Learning, pages 167-174. Morgan Kaufmann, San Francisco, CA, 2000.
12. Saeko Nomura, Tetsuo Hayamizu, "Analysis and Improvement of HITS Algorithm for DetectingWeb Communities".
13. Longzhuang Li, Yi Shang, and Wei Zhang, "Improvement of HITS-based Algorithms on Web Documents", WWW2002, May 7-11, 2002, Honolulu, Hawaii, USA. ACM 1-58113-449-5/02/0005.
14. J. Kleinberg, "Hubs, Authorities and Communities", ACM Computing Surveys, 31(4), 1999.