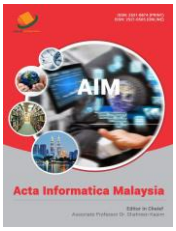
ZIBELINE INTERNATIONAL PUBLISHING

Acta Informatica Malaysia (AIM)
***DOI** : http://doi.org/10.26480/aim.02.2019.01.04*

***ISSN: 2521-0874 (Print)***
***ISSN: 2521-0505 (Online)***
***CODEN: AIMCCO***



***REVIEW ARTICLE***

# STUDY OF WEB PAGE RANKING ALGORITHMS: A REVIEW

**Aditi Chowdhary[1], Arvind Kumar[2]**

*[1]M. Tech. Scholar, Vidya College of Engineering, Meerut*
*[2]Assistant Professor, CSE, Vidya College of Engineering, Meerut*
**Corresponding Author Email: aditichaudhary032@gmail.com, arvind.kumar@vidya.edu.in*



**ARTICLE DETAILS**

***Article History:***

Received 04 January 2019
Accepted 18 February 2019
Available Online 15 March 2019

**ABSTRACT**

Due to expanding of Web, the Information is also increasing day by day. Many users generally rely on search engine to explore the web. It is a challenge for search engine to provide relevant web page and quality information against the query submitted. This paper deals in the comparative study of various web page ranking algorithms and also its advantages and disadvantages. This paper helps to find out web page ranking algorithms' relative strength and limitations for further research scope.

**KEYWORDS**

Web Page Ranking, HITS, WCM, WSM, WUM, Weight Page Rank.

## 1. INTRODUCTION

Today's Internet is growing in rapidly and the volume of information is increasing day by day. It becomes very difficult to manage information on the web. The World Wide Web consists of large hyperlinked structure or unstructured hyperlinked content. To retrieved required information from World Wide Web, User use search engine. A search engine usually takes a keyword query from user and returns a ranked list of the most relevant web documents. The search engine depends upon crawlers for the collection of required pages [1]. The Working of the web crawler is shown in figure 1.

In order to download a page, the crawler picks up its seed URL, and depending on the host protocol, downloads the page from the web server. For instance, when a user accesses an HTML page using its URL. The crawler simply sends HTTP requests for page to other machines on the Internet, just as a Web browser does when the user clicks on links. A single URL Server serves lists of URLs to a number of crawlers. Web crawler starts by parsing a specified Web page, noting any hypertext links on that page that point to other Web pages. Then parse the pages for new links, and so on, recursively.

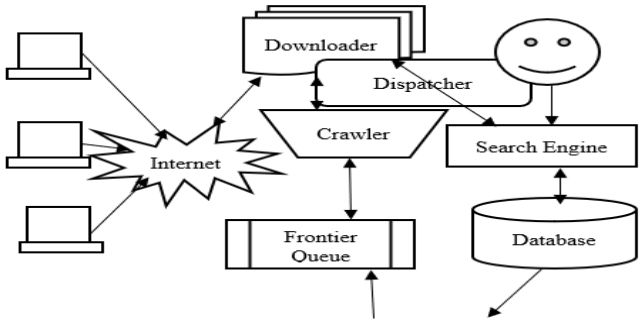



**Figure 1:** Working of Web Crawler

When the crawler visits a Web page, it extracts links to other Web pages. So, the crawler puts these URLs at the end of a queue and continues crawling to a URL that it removes from the front of the queue.

The Algorithm of the Web crawler is given below:

1) Read a URL from the set of seed URLs.
2) Determine the IP address for the host name.
3) Download the Robot.txt file which carries downloading permissions and also specifies the files to be excluded by the crawler.
4) Determine the protocol of underlying host like http, ftp, gopher etc.
5) Based on the protocol of the host, download the document.
6) Identify the document format like doc, html, or PDF etc.
7) Check whether the document has already been downloaded or not.
8) If the document is fresh one. Then Read it and extract the links or references to the other Cites from that document.
9) Else Continue.
10) Convert the URL links into their absolute IP equivalents.
11) Add the URLs to set of seed URLs.

## 2. WEB PAGE RANKING ALGORITHMS

### 2.1 Page Rank Algorithm

Sergey Brin and Larry Page introduced the page rank algorithm for ranking of web pages [2]. Web consist of the rich hyperlinked structure of web pages. The page rank algorithm utilizes the link structure of web pages to rank the pages. According to page rank algorithm if a page contains important links towards it then the links of this page towards the other page are also to be considered as important page. Let 'A' is a web page having hyperlink to web page 'B' this means that web page 'B' is important and relevant page. Page rank takes the backlink and propagates the ranking through links. If the sum of all ranks of the backlink is high then high rank is given to that page Figure 2 shows an example of backlinks: Page A is Page B and Page C, while Page C is backlink of Page D [3].

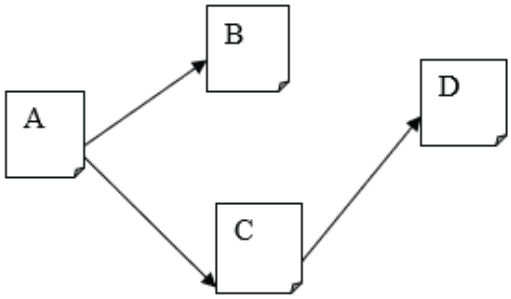



**Figure 2:** Example showing backlink of page

A Page rank is defined as:

$$\boldsymbol{PR(u) = c \sum_{v \in B(u)} \frac{PR(v)}{Nv}} \quad \ldots\ldots\ldots\ldots (\mathbf{1})$$

Where u and v are web pages, B(u) is the set of pages that point to page u. PR(u) and PR(v) are rank score of page u and v respectively. Nv denotes the number of outgoing link of page v and c is normalization factor.

In page rank, the rank score of a page say p, is equally distributed among its outgoing links and calculate iteratively. The assigned values to outgoing links are in turn used to calculate the ranks of pages to which page p, is pointing.

Due to factor that not all the user follows the existing link to any web page. Hence, equation 1 modified:

$$\boldsymbol{PR(u) = (1-d) + D \sum_{v \in B(u)} \frac{PR(v)}{Nv}} \quad \ldots\ldots (\mathbf{2})$$

***Technique:*** Based on Web Structure Mining (WSM) and consider Backlink.
***Advantage*:** Ranking is done on the basis of page importance.
***Limitations***: Ranking is at indexing time and not at the query time.

### 2.2 Weighted Page Rank Algorithm

Wenpu Xing and Ali Ghorbani proposed Weighted Page rank algorithm, which is extension of page rank algorithm [4]. This algorithm calculates the page rank on the basis of popularity of pages by taking consideration the importance of both incoming and outgoing links of pages. The popularity of a page is decided by its number of in links and out links. This algorithm does not equally distribute the rank of a page among its out links. The importance is assigned in terms of weight values to in links and out links and is denoted as:

$\boldsymbol{W^{in}_{(v,u)}}$ **:** is the weight of link (v, u) and calculate the popularity of the in links of webpage as:

$$\boldsymbol{W^{in}_{(v,u)} = \frac{I_u}{\sum_{p=R(v)} I_p}} \quad \ldots\ldots\ldots\ldots (\mathbf{3})$$

where, I_u and I_p are the number of inlinks of web page u and web page p, R(v) is reference list of page v.

$\boldsymbol{W^{out}_{(v,u)}}$: is the weight of link(v,u) and calculate the popularity of the out links of webpage as:

$$\boldsymbol{W^{out}_{(v,u)} = \frac{O_u}{\sum_{p=R(v)} O_p}} \quad \ldots\ldots\ldots\ldots (\mathbf{4})$$

where, O_u and O_p are the number of out links of web page u and web page p, R(v) is reference list of page v.

By taking the popularity of webpages into consideration the weighted page rank given as:

$$\boldsymbol{PR(u) = (1-d) + D \sum_{v \in B(u)} PR(v)\ W^{in}_{(v,u)} W^{out}_{(v,u)}} \quad \ldots\ldots\ldots\ldots \mathbf{(5)}$$

***Technique***: Based on calculation of weight of both in links and out links of page.
***Advantage:*** It higher accuracy in term of ranking because it uses content of web pages.
***Limitation:*** It is only based on popularity of web pages.

### 2.3 HITS (Hyper-Link Induced Topic Search)

Jon Kleinberg proposed this algorithm, in which two forms of web pages called as Hubs and Authorities [5]. Authorities are the pages having important contents and pointed by many hyperlinks. A good Hub page is a page which is pointing to many Authorities pages. An illustration of Hub and Authority as shown in figure 3.

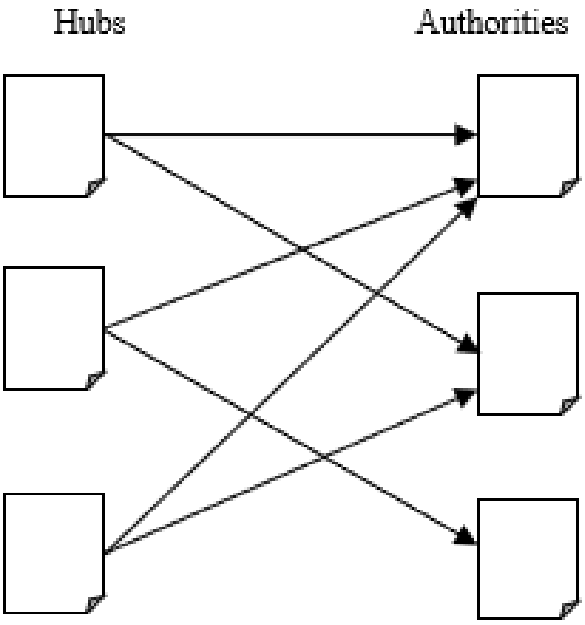



**Figure 3:** Illustration of Hub and Authorities

In this algorithm, ranking of web page is decided by analyzing the text contents against query. After collection of web pages, the Hits algorithm uses the web structure.

It has two phases:

1. Sampling: In this a set of relevant pages for the given query are collected.

2. Iterative: Find Hubs and Authorities using the output of sampling phase using following equations:

$$H_p = \sum_{q \in I(p)} A_q \quad \ldots\ldots\ldots\ldots (6)$$

$$A_p = \sum_{q \in B(p)} H_q \quad \ldots\ldots\ldots\ldots (7)$$

Where H_p is Hub weight, A_p is Authority weight, A_q is Authority score, H_q is Hub score of a page, B(p) and I(p) denotes the referrer and reference pages of page p.

A limitation of this algorithm is that it assumes equal weights to all the links pointing to a web page and it fail to identify some links may be more important than the other. To resolve this problem, a Probabilistic analogue of HITS (PHITS) is proposed [6]. This algorithm is able to identify authoritative documents.

***Technique:*** Based on calculation of Hub and Authorities score of page in order of their relevance. It uses Web Structure Mining and Web Content Mining (WCM).

***Advantage:*** High relevancy and importance of web page.
***Limitation:*** Problem of topic drift and not efficient in real time.

### 2.4 Eigen Rumor Algorithm

As the number of blogging sites is increasing rapidly, providing the rank to blog is quiet challenging. If Page rank and HITS both algorithm are applied directly to blogs, then the rank of blog is decided by page rank algorithm which cannot allow blog to be provided according to their importance. To resolve this problem, Ko Fujimura, Takafumi Inoue and Masayuki Sugisaki proposed an algorithm for ranking the blogs called Eigenrumor Algorithm [7]. This Algorithm provides a rank to every blog by weight of Hub and Authority of the bloggers depending on the calculation of eigen vector.

***Technique:*** Use the adjacency matrix, constructed from agent to object link not by page to page link. It uses Web Content Mining.
***Advantage:*** It gives high relevancy for blog.
***Limitation:*** It is only for blog not for web page.

### 2.5 Time Rank Algorithm

H. Jiang proposed this algorithm for improving the rank score by using visit time of web page [8]. This algorithm use Web usage mining to know degree of importance of web page to the user. In this algorithm the visiting time is added to the calculated score of page rank algorithm of a page. If a page contents related to the keyword of the user query then the user will stay for long time on that page, other user will leave quickly giving short visiting time to that page.

***Technique:*** It use Web Usages Mining (WUM). Visitor time is added to the computational score of page rank algorithm.
***Advantage:*** It is useful when two pages have the same link structure but different content.
***Limitation:*** It does not work efficiently when sever log is not present.

### 2.6 Distance Rank Algorithm

Ali Mohammad Zareh Bidoki and Nasser Yazdani proposed distance rank algorithm which is based on reinforcement learning which consider the logarithmic distance between pages [9]. In this algorithm ranking is done on the basis of the shortest logarithmic distance between two pages and ranked according to them. It follows the forward links. Technique: It uses Web Structure Mining and based on reinforcement learning which consider the logarithmic distance between pages.

***Advantage:*** It consider real user and find pages with high quality and more quickly.
***Limitation:*** A large calculation is done, if new page is inserted between the two pages.

### 2.7 Tag Rank Algorithm

Shen Jie, Chen Chen, Zhang Hui, Sun Rong-Shung, Zhu Yan and He Kun proposed Tag rank algorithm for ranking of web page based on social annotations [10]. This algorithm calculates the heat of the tags by using time factor of the new data source tag and annotations behavior of user. It provides better authentication for ranking of web pages.

***Technique:*** It use Web Content Mining (WCM). Visitor time is used for ranking and use of sequential clicking for sequence vector calculation with the use of random surfing model.
***Advantage:*** New information resources are indexed more effectively.
***Limitation:*** Co-occurrence factor of tag is not considered which may affect the weight of tag.

## 3. COMPARISON OF VARIOUS WEB PAGE ALGORITHMS

Comparison of various web page algorithms is done on the basis of some parameters as shown in table 1.

**Table 1:** Comparison between Various Web Page Rankings

| Algorithm | Page Rank | Weighted Page Rank | Hits | Eigen Rumor | Time Rank | Distance Rank | Tag Rank |
|---|---|---|---|---|---|---|---|
| **Author/ Year** | Sergey Brin and Larry Page, 1998 | Wenpu Xing and Ali Ghorbani, 2004 | Jon Kleinberg, 1998 | Ko Fujimura, Takafumi Inoue and Masayuki Sugisaki, 2005 | H. Jiang, 2008 | Ali Mohammad Zareh Bidoki and Nasser Yazdani, 2007 | Shen Jie, Chen Chen, Zhang Hui, Sun Rong-Shung, Zhu Yan and He Kun, 2008 |
| **Technique Used** | Web Structure Mining | Web Structure Mining | Web Structure Mining and Web Content Mining | Web Content Mining | Web Usage Mining | Web Structure Mining | Web Content Mining |
| **Methodology** | Compute the score for page at the time of indexing. | Weight is calculated on both inlinks and out links. | Compute the Hubs and Authority | Use the adjacency matrix, constructed from agent to object link not by page link | Visitor time is added to the computational score of page rank algorithm. | Based on reinforcement learning which consider the logarithmic distance between pages. | Visitor time is used for ranking and use of sequential clicking for sequence vector. |
| **Input Parameter** | Back link | Back link and forward links | Contents, Back link and forward links | Agent | Page Rank and Server Log | Forward links | Popular tags |
| **Result Quality** | Medium | Higher than page rank | Less than page rank | Higher than page rank and Hits | Moderate | High | Less |
| **Limitation** | Ranking is at indexing time and not at the query time. | Relevancy is ignored | Topic Drift and Efficiency | Only for blog not for web page. | Important page is ignored because it increases rank of those page which is open for long time | Large calculation is done, if new page is inserted between the two pages. | Co-occurrence factor of tag is not consider which may affect the weight of tag. |

## 4. CONCLUSION

This paper describes the techniques of various web page ranking to retrieve the relevant pages through search engine. These algorithms have been compared on the basis on literature survey having various parameter such as methodology, input parameter, relevancy, importance affect the ranking of web pages and also conclude that existing technique have limitations in term of accuracy of result, time response. An efficient web page ranking algorithm should meet out these challenges efficiently.

## REFERENCES

[1] Duhan, N., Sharma, A.K., Bhatia, K.K. 2009. Page Ranking Algorithms: A Survey. IEEE International Advance Computing Conference.

[2] Page, L., Brin, S., Motwani, R., Winograd, T. 1999. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report, Stanford Digital Libraries SIDL-WP-1999-0120.

[3] Ridings, C., Shishigin, M. 2002. Pagerank Uncovered. Technical Report.

[4] Xing, W., Ghorbani, A. 2004. Weighted PageRank Algorithm. In proceedings of the 2nd Annual Conference on Communication Networks & Services Research, 305-314.

[5] Kleinberg, J. 1998. Authoritative Sources in a Hyperlinked Environment. In Proceedings of the ACM-SIAM Symposium on Discrete Algorithms.

[6] Cohn, D., Chang, H. 2000. Learning to Probabilistically Identify Authoritative Documents. In Proceedings of 17th International Conference on Machine Learning, 167–174. Morgan Kaufmann San Francisco, CA.

[7] Fujimura, K., Inoue, T., Sugisaki, M. 2005. The EigenRumor Algorithm for Ranking Blogs. In WWW 2005 2nd Annual Workshop on the Weblogging Ecosystem.

[8] Jiang, H. 2008. TIMERANK: A Method of Improving Ranking Scores by Visited Time. In proceedings of the Seventh International Conference on Machine Learning and Cybernetics, Kunming.

[9] Bidoki, A.M.Z., Yazdani, N. 2007. Distance Rank: An intelligent Ranking Algorithm for Web Pages. Information Processing and Management.

[10] Jie, S., Chen, C., Hui, Z., Shuang, S.R., Yan, Z., Kun, H. 2008. Tag Rank: A New Rank Algorithm for Webpage Based on Social Web. In proceedings of the International Conference on Computer Science and Information Technology.