© March 2016 | IJIRT | Volume 2 Issue 10 | ISSN: 2349-6002

# A Comparative Study of Web Page Ranking Algorithms

Priyanka Tuteja, Meena Chaudhary, Sarika Gambhir
*Computer Science & Engineering,*
*Manav Rachna College of Engineering, Faridabad, Haryana*

***Abstract*** **- The World Wide Web WWW is a huge resource of hyperlinked and heterogeneous information which comprises of billions of web pages . To retrieve required information from World Wide Web, search engines perform various tasks based on its architecture and provide relevant and quality information to the internet users in response to its query. by using the web page contents and hyperlink between the web pages. Web mining is an active research area in present scenario. It is defined as the application of data mining techniques on the World Wide Web to find hidden information, This hidden information i.e. knowledge which contained in content of web pages or in link structure of World Wide Web or in web server logs. This paper deals with analysis and comparison of web page rankingalgorithms based on various parameter for the ranking of the web pages. Based on the analysis of various rankingalgorithms, a comparative search is done to search out out theirrelative strengths and limitations and furtherscope of analysis in web page ranking algorithmic rule.**

***Index Terms*****—- WWW; Data mining; Web mining; Search engine.**

## I. INTRODUCTION

The World Wide internet (Web) is widespread and interactive medium to propagate data nowadays. the web is large, diverse, dynamic, cosmopolitan global information service center. As on nowadays computer network is that the largest information repository for knowledge reference. With the ascension of the web, users get simply lost within the wealthy link structure. Providing relevant info to the users to cater to their desires is that the primary goal of web site owners. Therefore, finding the content of the net and retrieving the users'. interests and desires from their behavior became progressively vital. once a user makes a query from search engine, it usually returns an outsized range of pages in response to user queries. This result-list contains several relevant and digressive pages in line with user's. Figure 1 shows a working of a typical search engine, which shows the flow graph for a searched query by a web user.

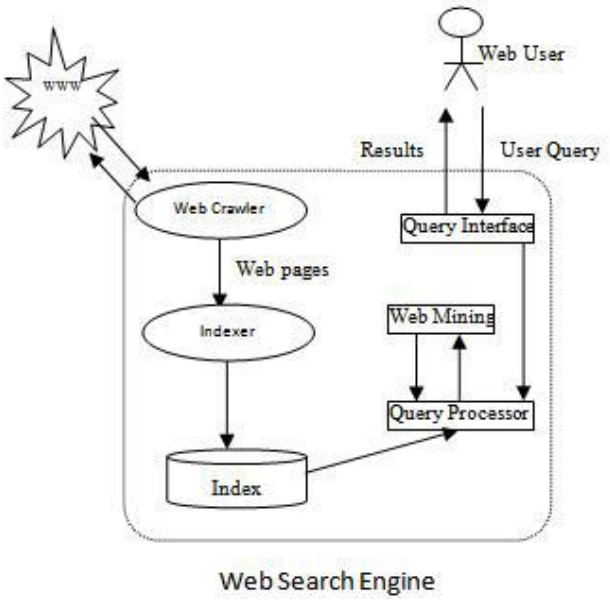


Figure 1: Working of Search Engine

An economical ranking of query words features a major role in efficient finding out query words. There square measure numerous challenges related to the ranking of web content such that some web pages are created just for navigation purpose and some pages of the online don't possess the standard of self descriptiveness. For ranking of webpages , several algorithms are planned within the literatures . The motive behind this paper to research the presently important algorithms for ranking of web content to seek out out their relative strengths, limitations and supply a future direction for the analysis within the field of economical algorithm for ranking of the web pages. The remaining part of this paper is organized as follows : section II includes web mining concepts ,categories and technologies have been discussed. Section III provides a detailed overview of some page ranking algorithms, section IV summarizes the techniques , advantages and limitations of a number of the vital webpage ranking algorithms, section V discuss the comparison of some of varied web page ranking algorithms and a conclusion is given in section VI.

## II. WEB MINING

Web mining is the technique to classify the web pages and internet users by taking into consideration the contents of the page and behavior of internet user in the past. Web Mining is the application of data mining techniques to discover and retrieve useful information from the WWW documents and services. Web mining can be divided into 3 categories namely web content mining, web structure mining and web usage mining[2] as shown in Fig 2.

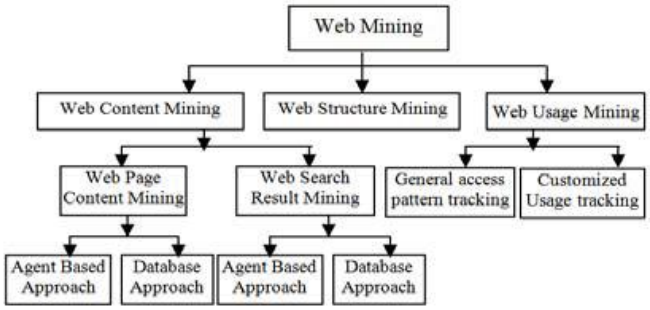


Figure 2: Classification of Web Mining

**Web Content Mining** (WCM) means mining the content web pages. It can be applied on web pages itself or on result pages obtained from a search engine. WCM can be differentiated from two different views: Information Retrieval (IR) View and Database (DB) View. In IR view, almost all the researches use bag of words to represent unstructured text, while for the semi-structured data, the HTML structure inside the documents can be used. Intelligent web agents can be used here for web mining purpose. In DB view, a web site can be transformed to represent a multi-level database and web mining tries to infer the structure of the web site from this database.

**Web Structure Mining (WSM)** tries to discover the link structure of the hyperlinks at the inter-document level in contrast to WCM that focuses on the structure of inner-document. It is used to generate structural summary about the web pages in the form of web graph where web pages act as nodes and hyperlinks as edges connecting two related pages.

**Web Usage Mining (WUM)** is used to discover user navigation patterns and the useful information from the web data present in server logs, which are maintained during the interaction of the users while surfing on the web. It can be further categorized in finding the general access patterns or in finding the patterns matching the specified parameters.[4]

## III. RANKING ALGORITHMS

### A Page Rank Algorithm :

Surgey Brin and Larry Page developed a ranking algorithmic program utilized by Google, named Page Rank (PR) once Larry Page (cofounder of Google search engine), that uses the link structure of the web to work out the importance of sites. Page Rank algorithm is that the most typically used algorithm for ranking the varied pages.operating of the Page Rank algorithmic program depends upon link structure of the web pages. The Page Rank algorithmic program relies on the ideas that if a page contains necessary links towards it then the links of this page towards the other page are to be thought-about as important pages. The Page Rank considers the back link in deciding the rank score. If the addition of the all the ranks of the back links is massive then the page then it's provided a large rank [3] . A simplified version of Page Rank is given by:

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

Where the Page Rank value for a web page u is dependent on the Page Rank 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.
An example of back link is shown in figure 3 below. U is the back link of V & W and V & W are the back links of X.

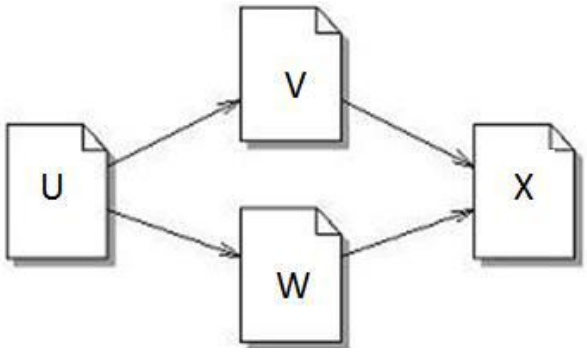


Figure 3: Illustration of back links

**B HITS Algorithm**

Kleinberg developed a WSM based mostly algorithm referred to as Hyperlink-Induced Topic Search (HITS) that ranks the web page by process in links and out links of the web pages.during this algorithm an online page is known as as authority if the web page is purposeed by several hyper links and a web page is known as as HUB if the page point to numerous hyperlinks. associate Illustration of HUB and authority square measure shown in figure 4.

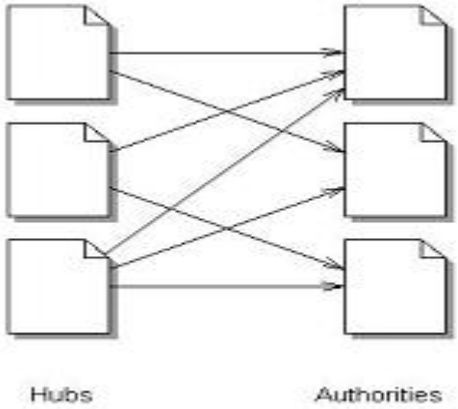


Figure 4: Illustration of Hub and Authorities

HITS is technically, a link based algorithm. In HITS [9] algorithm, ranking of the web page is decided by analyzing their textual contents against a given query. After collection of the web pages, the HITS algorithm concentrates on the structure of the web only, neglecting their textual contents. Original HITS algorithm has some problems which are given below.

(i) High rank value is given to some popular website that is not highly relevant to the given query. (ii) Drift of the topic occurs when the hub has multiple topics as equivalent weights are given to all of the outlinks of a hub page. Figure 5 shows an Illustration of HITS process.

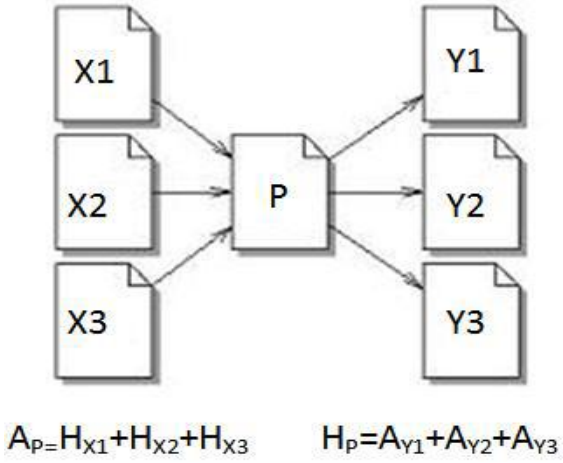


Figure 5: Illustration of HITS process

To minimize the problem of the original HITS algorithm, a clever algorithm is proposed by reference [6]. Clever algorithm is the modification of standard original HITS algorithm. This algorithm provides a weight value to every link depending on the terms of queries and endpoints of the link. An anchor tag is combined to decide the weights to the link and a large hub is broken down into smaller parts so that every hub page is concentrated only on one topic. Another limitation of standard HITS algorithm is that it assumes equal weights to all the links pointing to a webpage and it fails to identify the facts that some links may be more important than the other. To resolve this problem, a probabilistic analogue of the HITS (PHITS) algorithm is proposed by reference [11]. A probabilistic explanation of relationship of term document is provided by PHITS. It is able to identify authoritative document as claimed by the author. PHITS gives better results as compared to original HITS algorithm. Other difference between PHITS and standard HITS is that PHITS can estimate the probabilities of authorities compared to standard HITS algorithm, which can provide only the scalar magnitude of authority [1].

**C Weighted Page Rank Algorithm**

Weighted Page Rank [1] rule is planned by Wenpu Xing and Ali Ghorbani. Weighted page rank rule (WPR) is that the modification of the original page rank algorithm. WPR decides the rank score supported the popularity of the pages by taking into thought the importance of both the in-links and out-links of the pages. This rule provides high worth of rank to the more popular pages and doesn't equally divide the rank of a page among its out-link pages. each out-link page is given a

rank price supported its popularity. popularity of a page is decided by perceptive its variety of in links and out links. Simulation of WPR is finished using the web site of Saint Thomas University and simulation results show that WPR algorithm finds larger variety of relevant pages compared to standard page rank rule. As advised by the author, the performance of WPR is to be tested by using completely different websites and future work embrace to calculate the rank score by utilizing over one level of reference page list and increasing the quantity of human user to classify the web pages.

### D Weighted Links Rank Algorithm

A modification of the standard page rank algorithm is given by ricardo Baeza-Yates and Emilio Davis [7] named as weighted links rank (WLRank). This algorithm provides weight worth to the link supported 3 parameters i.e. length of the anchor text, tag during which the link is contained and relative position within the page. Simulation results show that the results of the program are improved using weighted links. The length of anchor text looks to be the simplest attributes during this rule. Relative position, that reveal that physical position doesn't always in synchronism with logical position isn't therefore result bound. Future add this algorithm includes, calibration of the burden issue of each term for more evolution.

### E Distance Rank Algorithm

An intelligent ranking formula named as distance rank is proposed by Ali mohammad Zareh Bidoki and Nasser Yazdani [5]. it's supported reinforcement learning algorithm. during this formula, the gap between pages is considered as a penalty factor. during this algorithm the ranking is done on the premise of the shortest index distance between 2 pages and ranked in step with them. The Advantage of this formula is that it will realize pages with prime quality and additional quickly with the employment of distance based resolution. The Limitation of this algorithm is that the crawler ought to perform a large calculation to calculate the distance vector, if new page is inserted between the two pages.

## IV. SUMMARY OF VARIOUS WEB PAGE RANKING ALGORITHM

By surfing the literature analysis of a number of the necessary web page ranking algorithms, it's concluded that every algorithm has some relative strengths and limitations. A tabular outline is given below in table one, which summarizes the techniques, advantages and limitations of a number of vital web page ranking algorithms.

| Author/Year | Technique | Advantages | Limitations |
|---|---|---|---|
| S. Brin et al. 1998 | Graph based algorithm based on link structure of web pages. Consider the back links in the rank calculations. | Rank is calculated on the basis of the importance of pages. | Results are computed at the indexing time not at the query time. |
| Jon Kleinberg, 1998 | Rank is calculated by computing hub and authorities score of the pages in order of their relevance. | Returned pages have high relevancy and importance. | With less efficiency and problem of topic drift |
| Wenpu Xing et al. 2004 | Based on the calculation of the weight of the page with the consideration of the outgoing links, incoming links and title tag of the page at the time of searching. | It gives higher accuracy in terms of ranking because it uses the content of the pages. | It is based only on the popularity of the web page. |
| Ricardo BaezaYates et al. 2004 | This algorithm ranks the page by providing different weights based on three attributes i.e. relative position in page, tag where link is contained & length of anchor text. | It has less efficiency with reference to precision of the search engine. | Relative position was not so effective, indicating that the logical position not always matches the physical position |
| Ali Mohammad Zareh Bidoki et al. 2007 | Based on reinforcement learning which consider the logarithmic distance between the pages. | Algorithm consider real user by which pages can be found very quickly with high quality. | A large calculation for distance vector is needed, if new page inserted between the two pages. |

Table 1 Summary of various web page ranking algorithms

## V COMPARISON OF VARIOUS WEB PAGE RANKING ALGORITHMS

Based on the literature analysis, a comparison of some of various web page ranking algorithms is shown in table 2 and in table 3. Comparison is done on the basis of some parameters such as main technique use, methodology, input parameter, relevancy, quality of results, importance and limitations.

| Algorithm | Page Rank | HITS | Weighted Page Rank | Web Page Ranking using Link Attributes | Distance Rank |
|---|---|---|---|---|---|
| **Main Technique** | Web Structure Mining | Web Structure Mining, Web Content Mining | Web Structure Mining | Web Structure Mining, Web Content Mining | Web Structure Mining |
| **Methodology** | This algorithm computes the score for pages at the time | It computes the hubs and authority of the | Weight of web page is calculated on the basis of input | it gives different weight to web links based on 3 attributes: | Based on reinforcement learning which consider the |

| | | | | | |
|---|---|---|---|---|---|
| | of indexing of the pages. | relevant pages. It relevant as well as important page as the result. | and outgoing links and on the basis of weight the importance of page is decided. | Relative position in page, tag where link is contained, length of anchor text. | logarithmic distance between the pages. |
| **Input Parameter** | Back links | Content, Back and Forward links | Back links and Forward links. | Content, Back and Forward links | 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. | more (it consider the relative position of the pages ) | Moderate due to the use of the hyperlinks. |
| **Quality of Results** | Medium | Less than PR | Higher than PR | Medium | High |
| **Importance** | High. Back links are considered. | Moderate. Hub & authorities scores are utilized. | High. The pages are sorted according to the importance. | Not specifically quoted. | High. It is based on distance between the pages. |
| **Limitation** | Results come at the time of indexing and not at the query time. | Topic drift and efficiency problem | Relevancy is ignored. | Relative position was not so effective, indicating that the logical position not always matches the physical position. | If new page inserted between two pages then the crawler should perform a large calculation to calculate the distance vector. |

Table 2 : Comparison of various web page ranking Algorithms

## V CONCLUSION

Based on the algorithm used, the ranking algorithm provides ought to use web page ranking techniques supported the a definite rank to resultant web pages. A typical search engine specific wants of the users. once researching exhaustive analysis of algorithms for ranking of web pages against the various parameters like methodology, input parameters, relevancy of results and importance of the results, it is concluded that existing techniques have limitations particularly in terms of your time response, accuracy of results, importance of the results and relevancy of results. An efficient web page ranking rule ought to meet out these challenges expeditiously with compatibility with global standards of web technology.

## REFERENCES

[1] Wenpu Xing and Ali Ghorbani, "Weighted PageRank Algorithm", In proceedings of the 2rd Annual Conference on Communication Networks & Services Research, PP. 305-314, 2004.

[2] Neelam Duhan, A. K. Sharma and Komal Kumar Bhatia, "Page
Ranking Algorithms: A Survey", In proceedings of the IEEE
International Advanced Computing Conference (IACC), 2009.

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

[4] Neelam Duhan, A. K. Sharma, Komal Kumar Bhatia, "Page
Ranking Algorithms: A Survey. YMCA Institute ofEngineering,
Faridabad, India, 2009 IEEE International Advance Computing
Conference (IACC 2009)Patiala, India, 6-7 March 2009

[5] S. Pal, V. Talwar, and P. Mitra, "Web Mining in Soft Computing
Framework : Relevance, State of the Art and Future Directions:, In
IEEE Trans. Neural Networks, 13(5), PP.1163–1177,2002.

[6] S. Chakrabarti, B. E. Dom, S. R. Kumar, P. Raghavan, S.
Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg, "Mining the
Web's Link Structure", Computer, 32(8), PP.60–67, 1999.

[7] Ricardo Baeza-Yates and Emilio Davis ,"Web page ranking using
link attributes" , In proceedings of the 13th international World Wide
Web conference on Alternate track papers & posters, PP.328-329,
2004.

[8] Dilip Kumar Sharma, A. K. Sharma," A Comparative Analysis
of Web Page RankingAlgorithms", Dilip Kumar Sharma et al. /
(IJCSE) International Journal on Computer Science and Engineering
Vol. 02, No. 08, 2010, 2670-2676.