A Brief Survey of Various Ranking Algorithms for Web Page Retrieval in Web Structure Mining

Yogita Garg[1], Mr. Vinod Jain[2]

[1]M.Tech Student, Department of CSE, B.S.Anangpuria Institute of Technology and Management, Alampur, India
[2]Assistant Professor, Department of CSE, B.S.Anangpuria Institute of Technology and Management, Alampur, India

***Abstract*—The World Wide Web consists millions of web pages that are interconnected. With the rapid growth of web it becomes very difficult to provide relevant information in respond to user's query. Most of the users rely on search engine to search the web. Search engine provide large amount of information in respond to user's query. In such a scenario it is the duty of service provider to provide proper, relevant and quality information to the internet user by using the web page contents and hyperlink between the web pages. So it becomes desirable to rank the pages according to relevancy. It is the job of page ranking algorithms to arrange web pages according to their relevancy to the user query. There are various algorithms for ranking the web pages. In this paper a survey of various page ranking algorithms and their merits and demerits have been discussed.**

***Keywords*—Search Engine, HITS, Page Rank, Weighted Page Rank, VOL (Visit of Link).**

I. INTRODUCTION

Web mining is the application of data mining techniques to discover patterns from the web. According to analysis targets, web mining can be divided in to three different types [2], which are Web Usage Mining, Web Content Mining and Web Structure Mining. Web Content Mining is the mining, extraction and integration of useful data, information and knowledge from web page content. Web Usage Mining is the process of finding out what users are looking for on the internet. Some user might be looking only at textual data, whereas some others might be interested in multimedia data. Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from web data in order to understand and better serve the need of web-based applications. Web Structure Mining deals with the web's hyperlink structure. It usually involves analysis of both the in links and out links of a web page. It is used in various page ranking algorithms.

The web is the largest source of data. During the past few years the World Wide Web has become the most popular way of communication and information dissemination. With the huge increase in availability of information to world-wide-web it has become difficult to access the desired information on the internet. There most of the users use search engine to navigate the web. Some popular search engines are Google, Yahoo, and Bing etc. Search engine is defined as a software program that takes input from user, searches its database and returns a set of results. The simple architecture [1] of a search engine is shown in Figure 1.

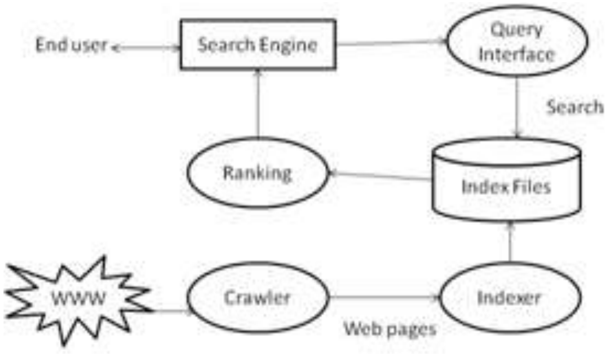

Fig. 1 Simple Architecture of a search engine

The major components of search engine are crawler, indexer and query processor. A crawler traverses the web by following hyperlinks and storing downloaded pages in a large database. It starts with seed URL and collects documents by recursively fetching links and storing the extracted URL's into a local repository. The Indexer processes and indexes the pages collected by the crawler. It extracts keywords from each page and records the URL where each word has occurred. In general Query Engine may return several hundreds or thousands of URL that match the keywords for a given query. But often users look at top ten results that can seen without scrolling. Users seldom look at results coming after first search result page, which means that results which are not among top ten are nearly invisible for general user. Therefore to provide better search result, page ranking mechanisms are used by most search engines for putting the important pages on top leaving the less important pages in the bottom of result list. So page ranking is helpful in web searching. Rankers are classified into two groups: - Content based rankers and Connectivity based rankers. Content based rankers work on the basis of number of matched terms, frequency of terms, location of terms. Connectivity based rankers work on the basis of link analysis technique, link are the edges that point to different web pages. Connectivity based rankers are independent of user's query and hence more dynamic as compared to Content based rankers. In this paper various page ranking algorithms are described based on web link structure. In this paper section II describe the work of page ranking algorithm, Section III describes the comparison of these algorithms and section IV describes the conclusion.

II. LINK BASED PAGE RANKING ALGORITHMS

To present the documents in an ordered way, page ranking algorithms are applied which arrange the documents according to user's query relevancy; most relevant pages are displayed on top. Search engine use two kind of ranking factors: Query dependent factors and Query independent factors. Web structure mining use Query independent factors, as search engine display the document on the top that is popular in its in links and out links which is independent of Query. The link based algorithms view the web as a directed graph where web pages are the nodes and link between web pages are the directed edges between these nodes [3]. The important link based page ranking algorithms are given below:

A. Page Rank Algorithm

Brin and Page developed Page Rank Algorithm [4], one of the most widely used page ranking algorithms. It states that if a page has important links to it, then its link to other pages also become important. Therefore page rank takes the back links into account and propagates the ranking through links. A slightly simplified version of Page Rank is defined as:

$$PR(u) = c \sum_{v \in B(u)} \frac{PR(v)}{Nv}$$

Where u is a web page. B (u) is the set of web pages that point to u.PR (u) and PR (v) are rank scores of page u and v, respectively. Nv is the number of outgoing links of page v. c is a factor used for normalization. In Page Rank, the rank score of a page p is evenly divided among its outgoing links.

Later Page Rank was modified considering that not all the users follow the direct links on WWW. The modified version is as follows:

$$PR(u) = (1-d) + d \sum_{v \in B(u)} \frac{PR(v)}{Nv}$$

Where d is for directly linked pages and (1-d) is for none directly linked pages.

An example of back link is shown in figure 2 below. U is the back link of V & W and V & W are the back links of X.

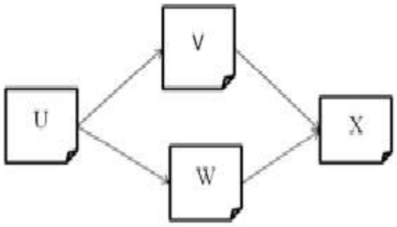

Fig 2 Illustration of back links

B. HITS Algorithm

HITS Algorithm [2] ranks the web page by processing in links and out links of the web pages. In this algorithm a web page is named as authority if the web page is pointed by many hyper links and a web page is names as HUB if the page points to various hyperlinks. Authorities and hubs are illustrated in Figure 3.

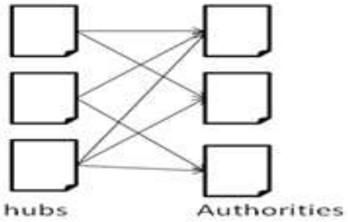

Fig. 3 Illustration of Hubs and Authorities

Hubs and authorities are assigned respective scores. Scores are computed in a mutually reinforcing way; an authority pointed to by several highly scored hubs should be a strong authority while a hub that points to several highly scored authorities should be a popular hub. 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;

$$ap = \sum_{q\in B(p)} hp$$

$$hp = \sum_{q\in B(p)} ap$$

Figure 4 shows an example of the calculation of authority and hub scores.

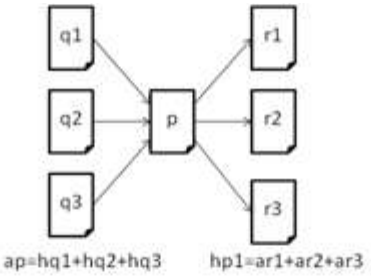

Fig. 4 An Example of HITS Operations

HITS is a purely link-based algorithm. It is used to rank pages that are retrieved from the web, based on their textual contents to a given query. Once these pages have been assembled, the HITS algorithm ignores textual content and focuses itself on the structure of the web only.

C. Weighted Page Rank Algorithm

Wenpu Xing and Ali Ghorbani proposed a Weighted Page Rank (WPR) algorithm [2] which is an addition of the Page Rank algorithm which assigns larger rank values to more important (popular) pages instead of dividing the rank value of a page evenly among its out link pages. Each out link page gets a value proportional to its popularity (its number of in links and out links). The popularity from the number of in links and out links is recorded as *Win* (*v*, *u*) and *Wout* (*v*, *u*), respectively.
Win (*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*.

$$Win(v,u) = \frac{Iu}{\sum_{p\in R(v)} Ip}$$

Where Iu and *Ip* represent the number of in links of page *u* and page *p*, respectively. *R* (*v*) denotes the reference page list of page *v*.

Wout (*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*.

$$Wout(v,u) = \frac{Ou}{\sum_{p\in R(v)} Op}$$

Where *Ou* and *Op* represent the number of out links of page *u* and page *p*, respectively. *R* (*v*) denotes the reference page list of page *v*.

Considering the importance of pages, the original Page Rank formula is modified as:

$$PR(u) = (1-d) + d \sum_{v\in B(u)} PR(v)win(v,u)wout(v,u)$$

D. PAGE RANKING BASED ON VISIT OF LINK

Gyanendra Kumar et. Al [3]. Proposed a new algorithm in which they considered user’s browsing behavior. This concept is very useful to display most valuable pages on the top of the result list on the basis of user browsing behavior, which reduce the search space to a large scale so; he proposed an improved Page Rank algorithm. In this algorithm we assign more rank value to the outgoing links which is most visited by users. In this manner a page rank value is calculated based on visits of inbound links.
The modified version based on VOL is given in equation:

$$PR(u) = (1-d) + d \sum_{v\in B(u)} \frac{Lu\, PR(v)}{TL(v)}$$

Here d is a dampening factor, u represents a web page, B (u) is the set of pages that point to u, PR (u) and PR (v) are rank scores of page u and v respectively, Lu is the number of visits of link which is pointing page u from v. TL (v) denotes total number of visits of all links present on v.

E. WEIGHTED PAGERANK BASED ON VISIT OF LINKS

Simple Sharma and Neelam Tyagi [4] proposed this algorithm. In this algorithm more rank value is assigned to the outgoing links which is most visited by users and received higher popularity from number of in link. Here the popularity of out links is not considered which is considered in the original algorithm. The advanced approach in the new algorithm is to determine the user’s usage trends. The user’s browsing behavior can be calculated by number of hits (visits) of links.
The modified version based on WPR (VOL) is given as:

$$WPRvol(u) = (1-d) + d \sum_{v\in B(u)} \frac{Lu\, WPRvol(v)win(v,u)}{TL(v)}$$

Here d is a dampening factor, u represents a web page, B (u) is the set of pages that point to u, WPRVOL (u) and WPRVOL (v) are rank scores of page u and v respectively, Lu is the number of visits of link which is pointing page u from v. TL (v) denotes total number of visits of all links present on v.

III. COMPARISON OF VARIOUS PAGE RANKING ALGORITHMS

The comparison of various link based page ranking algorithms [5] [6] [7] is given in table I below:

TABLE I: COMPARISON OF VARIOUS PAGE RANKING ALGORITHMS

Algorithm	Page Rank	HITS	Weighted Page Rank	Page Rank based on VOL	Weighted Page Rank based on VOL
Basic Criteria	Graph-based ranking algorithm, consider back link and forward link in rank calculation	Consider back link and forward link for rank calculation	Consider popularity of incoming links and outgoing links of web page for rank calculation	Consider number of visit of links for calculating rank score of a web page	Consider number of visit of links and popularity of incoming links of web page for calculating rank score
Technique used	Web Structure Mining	Web Structure Mining, Web Content Mining	Web Structure Mining	Web Structure Mining	Web Structure Mining
Input Parameter	Back links	Content, Back links and Forward links	Back links and Forward Links	Visit of Links	Visit of Links and Back links
Quality of Result	Medium	Less Than Page Rank	More than Page Rank	More than Page Rank but Less than Weighted Page Rank	More than Page Rank, Weighted page Ranking and Page Rank based on VOL.
Advantages	Rank is Calculated on the basis of importance of page	Hub and Authorities scores are utilized.	It assigns larger rank value to more important pages.	It consider the user's browsing behavior during rank score calculation.	It take in to account the user's browsing Behavior as well as the Popularity of web pages.
Disadvantage	It favor older pages, because a new page, even a very good page .will not have many links unless it is part of an existing web site.	Topic drift and efficiency problem	It is based only on popularity of the web page not on the user's browsing behavior	It is based on based on user's browsing behavior but not on the popularity of the web page	It is better than existing algorithms but it also need further Improvements.

IV. CONCLUSION

Generally the search engine result in a large number of pages in response to user's queries, but the user always want the best result on the top of list. She/he does not want to waste his time to navigate through the entire search result to get the requested one. The page ranking algorithm plays an important role in making the user search navigation easier. Link based page ranking algorithms give importance to links rather than content of the page. According to page rank algorithm, rank score of a web page is divided evenly over the pages to which it links whereas Weights Page Rank Algorithm assigns larger rank values to more important pages and Weighted Page Rank Based on VOL Assigns larger rank values to pages according to visits and popularity of in links and HITS consider Both in links, out links and contents for calculating Rank score.

REFERENCES

[1] N. Duhan, A. K. Sharma and K. K. Bhatia, "Page Ranking Algorithms: A Survey", Proceedings of the IEEE International Conference on Advance Computing, 2009.

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

[3] Gyanendra Kumar, Neelam Duhan, and Sharma A. K., "*Page Ranking Based on Number of Visits of Web Pages*", International Conference on Computer & Communication Technology (ICCCT)-2011, 978-1-4577-1385-9.

[4] Neelam Tyagi, Simple Sharma, "Weighted Page Rank Algorithm based on number of visits of links of web page", International Journal of Soft Computing and Engineering (IJCSE)-ISSN: 2231-2307, Volume-2, Issue-3, July 2012

[5] 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-ISSN: 2320-9798, Volume-2, Issue 2, February 2014

[6] Neelam Tyagi, Simple Sharma," Comparative Study Of Various Page Ranking Algorithms in Web Structure Mining(WSM)" ,International Journal of Innovative Technology and Exploring Engineering(IJITEE)- ISSN:2278-3075,Volume-1,Issue-1,June 2012

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