Exploration of Several Page Rank Algorithms for Link Analysis

Bipin B. Vekariya
P.G. Student, Computer Engineering Department,
Dharmsinh Desai University,
Nadiad , Gujarat, India.

Ankit P. Vaishnav
Assistant Professor, Computer Engineering Department,
Dharmsinh Desai University,
Nadiad , Gujarat, India.

***Abstract*— Web is the largest collection of information and it is growing continuously. Pages& Documents are added and deleted on frequent basis due to dynamic nature of the web. The web is serving as the major source of meaningful information related to query made by the user. The search engine applies different algorithms for link analysis to fetch most relevant pages and documents which are presented at the top of the result list. To assist the users to navigate in the result list, ranking methods are applied on the search results. Most of the ranking algorithms proposed in the literature are PageRank (PR) [1], Weighted PageRank (WPR) [5], Hyperlink-Induced Topic Search (HITS) [4], Page Ranking Algorithm Based On Number Of Visits Of Links Of Web Page [7],An Improved Page Ranking Algorithm Based On Optimized Normalization Technique [6],Improved method for computation of PageRank[8], Weighted Page Ranking Algorithm Based On Number Of Visits Of Links Of Webpage [9].With the understanding that Page Rankings for searching meaningful and relevant information from gigantic collection of web pages and numerous hyperlinks, hence primary purpose of this paper is to come up with the page rank method that gives more relevancy and accurate information to the user. i.e. Weighted Page Ranking Algorithm Based On Number Of Visits Of Links Of Webpage.Above Algorithms are reviewed, theoretically compared & analysed in order to reach a result about the most suitable algorithm for PR by observing the parameters in it.**

Keywords*— Pagerank, HITS, WPR, Pagerank With VOL, Authoritative Web Pages, Web Structure Mining.

I. INTRODUCTION

WWW is the biggest collection of information. It is the biggest and most widely known information source that is easily accessible and searchable. It consists of billions of interconnected documents called web pages, which are developed by millions of people. World web is growing continuously in last few years in terms of number of web pages. WebPages are added and deleted on frequent basis due to dynamic nature of the web. Figure 1 shows the structure of a classic Web graph.Structure of a classic Web graph can be seen as web pages as nodes, and hyperlinks as edges which connects two related nodes. The web serves as the key resource of meaningful information depending on user's request.

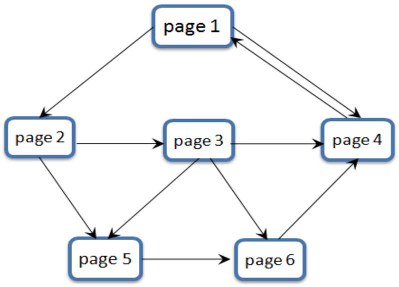


Fig.1: Classic Web Structure

Search engine in the simplest manner can be elaborate as a kind of system, which gives response with related information in the form of web pages & documents as per user's request. Google, MSN, Yahoo and other search engines keep their billions of web pages indexed & hosted on their servers across the world to serve billions of users. A user create different queries to indicate a search topic and after that in response the search engine identifies several web pages having related content that satisfies user's queries, e.g. web pages that match with the key words of user's queries. These pages are known as the result set.Quality of search engines can be evaluated not only on the number of pages that are indexed, but also on the usefulness of the ranking process that determines which pages are returned.To fetch web pages the search engine applies different algorithms for link analysis and after that the result is presented in the form of list of web pages. Figure 2 shows mechanism of basic search engine.

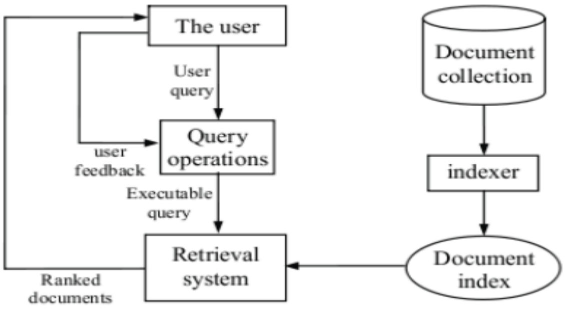


Fig.2: Basic Search Engine Mechanism [6]

WEB MINING

Aim of web mining is to discover relevant and meaningful information or knowledge from different resources. These resources mainly involve page content, hyperlink structure and usage data.

Web mining exploits several data mining techniques in order to search and extract information from WWW.

Web mining is classified into the major three categories:

A. WEB CONTENT MINING (WCM)

WCMextracts meaningful, useful and relevant information or knowledge from web page contents like text, image, audio, video, metadata, hyperlinks and extracts useful information. Web content mining can be further categorized into Web page content mining and search results mining.

Web page content mining is conventional searching of web pages with the help of content, while Search results mining is an additional search of pages found from a previous search.

WCM techniques can be exploit as concept hierarchies, synonyms, user profiles and analyzing the links between pages to improve web content mining. Web content mining uses data mining techniques for efficiency, effectiveness and scalability.

B. WEB STRUCTURE MINING

The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting between two related pages.

Web structure mining tries to discover useful knowledge from the structure of hyperlinks. It is used to identify the relationship between Web pages linked by information or direct link connection. This connection allows a search engine to pull data relating to a search query directly to the linking Web page from the Web site the content rests upon.

C. WEB USAGE MINING

Web usage mining exploit user access patterns from web usage logs, which involves progression of finding out what users are looking for on the web. Some users might be looking for only textual data, whereas some others are interested in multimedia. In order to figure out the usage pattern.

Further the rest of paper is organized as follows: section II provides a brief technical description of several Page Rank Algorithms. Section III provides theoretical comparison and analysis of those approaches based on the strength and limitations of each. And finally section IV delivers the conclusions.

II. TECHNICAL DESCRIPTION OF SEVERAL PAGE RANK ALGORITHMS

PageRank was designed to increase the effectiveness and improve the efficiency of the search engines. It is used to measure the importance of a page and to prioritize the pages returned from a traditional search engine using keyword searching. Google uses this technique. The PageRank value for a page is calculated based on the number of pages that point to it

Many of the researchers have contributed several techniques to serve the said purpose for it. Following are some of the related work accomplished by different researchers groups. In this section several approaches are briefly described. As the primary evaluation parameter of this paper is Page Ranking, only DifferentRanking methods in link analysis is explained. Other operations of Searching are not explained. One can go through the references for complete description of these methods.

A. Page Rank Algorithm [1]

This algorithm was developed by Brin and Page at Stanford University which extends the idea of citation analysis. A clear vision to PageRank is the calculation of impact of research publications in terms of the number of citations they have.

In the Web domain a page has a citation when there is another Web pages linked to it. From the point of view of the cited page this link is called a back-link. PageRank does somewhat more than just totalling back-links. It assigns different weights to back-links. Thus, a Web page can have a high rank if it has many back-links (citations) as well as if it has only few but highly rated back-links.

As shown in below formula u is taken as a webpage and B (u) the set of pages that indicate to u. Then v is taken as a webpage and Nv is the number of links from v. Finally, let c be a normalization factor for the total rank of all Web pages to be constant. Then the rank (simplified PageRank) R of u is computed like this:

$$R(u) = c\sum_{v\in B(u)} \frac{R(v)}{N_v} \qquad (1)$$

The above formula is recursive for computing page rank of any Web page. Therefore, they started with a vector of ranks initialized to some (arbitrary) values, iteratively update those ranks using above formula and wait until they unite. Experiments show that the full version of PageRank converged after around 50 recursive iterations. Figure 3 is an example of a page rank calculation.

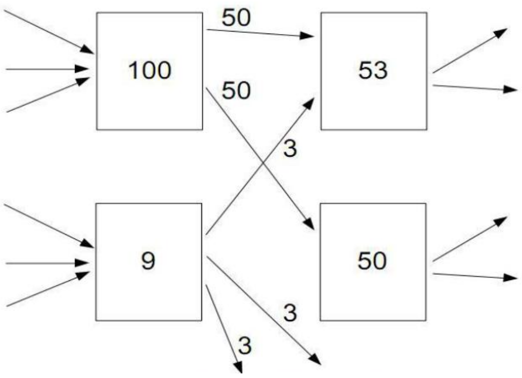


Fig 3: Example of a Page Rank Calculation

There is a little problem with so called rank sinks. Those are closed loops of pages that accumulate rank but never distribute it further, such as in Figure 4. To overcome this issue some modifications of the ranking.

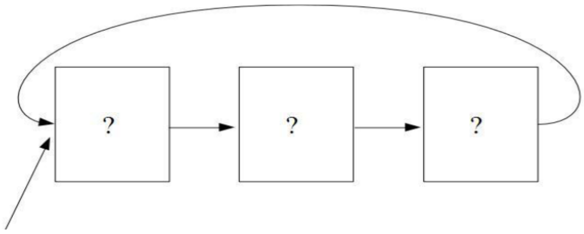


Fig 4: Rank Sink

To overcome this trouble some modifications of the ranking was carried out. If a backlink comes from an important webpage than this link is given higher weightage than those which are coming from non-important webpages. The link from one webpage to another is considered as a vote. Not only the number of votes that a page receives is important but the importance of webpages that casts the vote is also important.

Page and Brin proposed a formula to calculate the PageRank of a page A as stated below-

PR(A)=(1-d)+d(PR(T1)/C(T1)+…..+PR(Tn/C(Tn)) (2)

Here PR(Ti) is the PageRank of the Pages Ti which links to page A, C(Ti) is number of outlinks on page Ti and d is damping factor. Damping factor is used to stop other pages having too much impact. The total vote is “damped down” by multiplying it to 0.85.

The PageRank forms a probability distribution over the webpages so the sum of PageRanks of all web pages will be one. The PageRank of a webpage can be calculated without knowing the final value of PageRank of other pages. It is an iterative algorithm which follows the principle of normalized link matrix of web. PageRank of a webpage depends on the number of pages pointing to a page.

B. HITS (Hypertext Induced Topic Search) [4]

Jon Klienberg gave two forms of web pages, so called as hubs and authorities.

Hubs are the webpages that act as resource lists. A good quality hub page is a webpage which is pointing to many authoritative pages on that content. Authorities are pages having important contents. A good quality authority page is a page which is pointed by many good hub pages on the same content.

A page may be a good quality hub and a good quality authority at the same time. The HITS algorithm treats WWW as directed graph G(V,E), where V is a set of vertices representing pages and E is set of edges corresponds to link. Figure 5 shows the hubs and authorities in web.

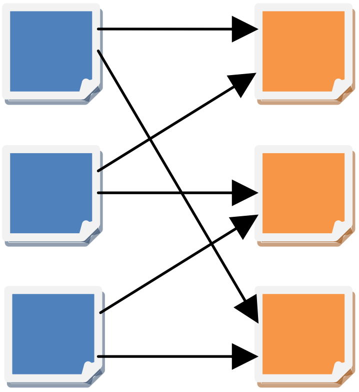


Fig. 5: Hubs and Authorities

It has two steps:

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

2. Iterative Step: In this step Hubs and Authorities are found using the output of sampling step. Following equations are used to calculate the weight of Hub and the weight of Authority.

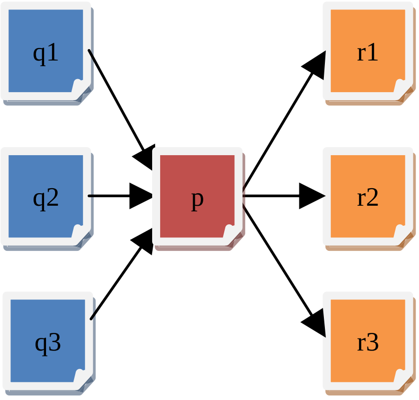
Here Hq is Hub Score of a page, Ap is authority score of a page, I(p) is set of reference pages of page p and B(p) is set of referrer pages of page p, the authority weight of a page is proportional to the sum of hub weights of pages that link to it. Similarly a hub of a page is proportional to the sum of authority weights of pages that it links to.

The score value of hubs and authorities are calculated as follows:

$$A_p = \sum_{q \in B(p)} Hq \quad (3)$$

$$H_p = \sum_{q \in I(p)} Aq \quad (4)$$

Figure 6 shows an example of the calculation of authority and hub scores



Ap=Hq1+Hq2+Hq3 Hp=Ar1+Ar2+Ar3

Fig. 6: A small example of HITS calculations

HITS is a simply link-based algorithm. It is used to ranking the pages which is mainly retrieved from Web and based on their textual contents to a given query. First these pages have been place together in proper way than the HITS algorithm ignores textual content and focuses itself on the structure of the Web only.

Constraints with HITS algorithm

Following are some constraints of HITS algorithm:

1. Hubs and Authorities: It is difficult to distinguish among hubs and authorities because many sites are hubs as well as authorities.

2. Topic drift: Sometime HITS may not give the most relevant documents to the user queries because of equivalent weights.

3. Automatically generated links: HITS gives equal importance for automatically generated links which may not have relevant topics for the user query.

4. Efficiency: It is not efficient in real time.

In IBM research project, HITS was used in a prototype search engine called Clever. Due to above constraints HITS could not be implemented in a real time search engine.

C. *WEIGHTED PAGERANKALGORITHM [5]*

This algorithm was proposed by Wenpu Xing and Ali Ghorbani which is an improvement of PageRank algorithm. This Algorithm assigns the rank values to webpages according to their importance rather than dividing it evenly. The importance is assigned in terms of weight values to incoming and outgoing links. This is denoted as$W^{in}_{(m,n)}$and $W^{out}_{(m,n)}$respectively. $W^{in}_{(m,n)}$is the weight of link(m,n). It is calculated on the basis of number of incoming links to page n and the number of incoming links to all reference pages of page m.

$$W^{in}_{(m,n)} = \frac{I_n}{\sum_{p\in R(m)} I_p} \qquad (5)$$

In is number of incoming links of page n, Ip is number of incoming links of page p, R(m) is the reference page list of page m.

$W^{out}_{(m,n)}$is the weight of link(m,n). It is calculated on the basis of the number of outgoing links of page n and the number of outgoing links of all the reference pages of page m. On is number of outgoing links of page n,

$$W^{out}_{(m,n)} = \frac{O_n}{\sum_{p\in R(m)} O_p} \qquad (6)$$

Op is number of outgoing links of page p, then the weighted PageRank is given by formula:

$$PR(n) = (1-d) + d\sum_{m\in B(n)} PR(m)\, W^{in}_{(m,n)} W^{out}_{(m,n)} \qquad (7)$$

The PR and WPR algorithms both provide ranked pages in the sorting order to users based on the given query. So, in the resultant list, the number of relevant pages and their order are very important for users. Relevance rule is applied for calculation of the relevancy value in each page from the list of pages which made Weighted PageRank different from PageRank.

After study of WPR it is clear that WPR produces larger relevancy values than the PR.

D. *PAGE RANKING BASED ON NUMBER OF VISIT OF LINKS OF WEBPAGE [7]*

Page Ranking Algorithm based on VOL [7] is another extension to the standard Page Ranking Algorithm [1] and also the user perspective is considered as number of visits of inbound links. In the Page Ranking Algorithm based on VOL, author assigns more weights to the outgoing links which are most frequently visited by the user. In this algorithm it is assigned more rank value to the outgoing links which is most visited by users. In this manner a page rank value is calculate based on visits of inbound links.

The modified version based on VOL is given in the following equation

$$PR(u) = (1-d) + d\sum_{v\in B(u)} \frac{L_u\,(PR\,(v))}{TL\,(v)} \qquad (8)$$

Notations are:

Lu denotes 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.Other notations are same as in original PageRank equation.

In order to find the number of visit of the links, client side script is used. For every webpage visited, the client side script is loaded to client side from the web server. Through the script the click and keyboard events are handled. And also in the web search crawler a counter is raised with every webpage retrieved with respect to any user query. Every time a webpage is visited the counter is raised or increased by one and later on fetched to compute the ranking of the webpages. The Page Ranking algorithm based on VOL is less complex than standard PR Algorithm [1], and advantageous as it includes the user input for calculating the rank. More complex when one talk about the design issues and is limited to the inbound links for computing the page ranking.

E. *ANIMPROVED PAGE RANKING ALGORITHM BASED ON OPTIMIZED NORMALIZATION TECHNIQUE [6]*

It shows a step extension to the standard PR Algorithm [1] & add the normalization feature to the ranking. It reduces the overall complexity of the ranking process by reducing the number of iteration to calculate the page rank. Steps of the Page Ranking Algorithm based on optimized normalization technique are:

- Initially assume PageRank of all web pages to be any value, let it be1.
- First Calculate page ranks of all pages by following equation:-

PR(A) = 0.15 + 0.85 (PR(T1)/C(T1) + PR(T2)/C(T2) +……. + PR(Tn)/C(Tn)) (9)

Where T1 through Tn are pages providing incoming links to Page A,PR(T1) is the Page Rank of T1,PR(Tn) is the Page Rank of Tn,C(Tn) is total number of outgoing links on Tn.

- And calculate mean value of all page ranks by following equation:

Summation of page ranks of all web pages / number of web pages

- Now normalize page rank of each page by following formula:

Norm PR (A) = PR (A) / mean value

Where Norm PR (A) is Normalized Page Rank of page A and PR (A) is page rank of page A. and Assign PR(A)= Norm PR (A)

- Repeat until page rank values of two consecutive iterations are same.

The normalized page ranking is less complex than the standard PR Algorithm as the number of iterations is reduced by using the normalized values hence the time complexity is reduced. In this technique of page ranking the number of iterations are removed which reduces the computational complexity of the algorithm. As it is link based algorithm hence the problem of theme drift exists.

F. AN IMPROVED METHOD FOR THE COMPUTATION OF PAGERANK [8]

This paper improved the results of Web search by utilizing the link structure of the web and also given some analysis of topic bias, absolute equalization and emphasis on old web about the algorithm, which influences the ranking quality of websites. In order to solve these problems, it is proposed an improved method for the computation of PageRank on the basis of integrated factors. The experiment shown that improved method outperforms the PR [1] algorithm in the quality of the pages returned.

Authors of this paper used various improvements to existing algorithms, taking topic character, distribution of PageRank value and time factor into consideration; at last, they integrated two improvements to proposed method:

- Introduced topic Character to Eliminate Topic Bias.
- Introduced time factor to emphasis on new web.

So ultimate improved method for the computation was

$$\mathrm{PR(u)} = \mathrm{dT_u}[\textstyle\sum_{\mathrm{v}\in\mathrm{B(u)}} \mathrm{PR(V)}(\frac{\mathrm{m}}{\mathrm{N_v}} + (1-\mathrm{m})\mathrm{W_v}) + (1-\mathrm{d})] \qquad (10)$$

In the formula, $\mathrm{W_v}$ denotes the topic weight between page u and page v, $\mathrm{T_u}$ denotes the time factor of page u.

There are topic character and time factor taken into account and method for the computation of PageRank shown more excellent search performance compared to the classic PageRank algorithm.

G. WEIGHTED PAGE RANK ALGORITHM BASED ON NUMBER OF VOL OF WEBPAGE [7]

This Algorithmis used to get more relevant information according to user's query. Hence, this concept is very useful to display most valuable pages on the top of the result list on the basis of user browsing behaviour, which reduce the search space to a large scale.

The modified version based on WPR(VOL) is given in the following equation

$$\mathrm{WPR}_{vol}(\mathrm{u}) = (1-\mathrm{d}) + \mathrm{d}\textstyle\sum_{\mathrm{v}\in\mathrm{B(u)}} \frac{\mathrm{L_u}\,\mathrm{WPR}_{vol}(\mathrm{v})W^{in}_{(v,u)}}{\mathrm{TL(v)}} \qquad (11)$$

Where u represents a web page, B(u) is the set of pages that point to u, d is the dampening factor. WPRvol (u) and WPRvol (v) are rank scores of page u and v respectively, Lu denotes 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.

Visits of links are computed as explained in Page ranking based on number of visit of links of webpage method.

The top returned pages in the result list is supposed to be highly relevant to the user information needs, the ordering of pages using WPR (VOL) is more target-oriented and user can't purposefully increase the rank of a page by visiting the page multiple times because the rank of the page depends on the not on the count of visits on back linked pages.

III. THEORETICAL COMPARISON & ANALYSIS

Parameters / Algorithm	**Page Ranking Algo.[1]**	**HITS Algo.[4]**	**Weighted Page Rank Algo. [5]**	**Page Ranking Based on visit of link of pages[7]**	**Improved Page Ranking Algo.[6]**	**Improved method for computation of PR[8]**	**WPR Algorithm Based On Number of VOL of Webpage[9]**
Mining Technique used	Web Structure Mining	Web Structure Mining, Web Content Mining	Web Structure Mining	Web Structure Mining, Web usage mining	Web structure mining	Web Structure Mining, Web Content Mining	Web Structure Mining, Web usage mining
Technique	Based on Web graph of the webpages and uses the inbound links to rank the webpages	Hubs and Authority score is used to rank the webpages	Based on the popularity of Inlinks and Outlinks the page ranking is done	Enhance the standard page ranks by considering the VOL of pages	By mean values of the rank of all pages replaces the ranks of all pages than takes another round & carry on until the value of two consecutive pages becomes equal	Takes topic character and time factor into account to rank the webpages	Improve the WPR by considering the VOL of pages
I/P Parameters	Inbound links of the pages	Content, Back and Forward links	Backlinks, Forwardlinks	Inbound links, outbound links, Count of VOL.	Inbound links	Content, Inbound links of other pages	Backlinks, Forwardlinks, Count of VOL.
Strength	Ranking is done on the basis of importance of the pages	Relevancy of the pages is high	Higher relevancy because popularity of the links is considered	User input is considered therefore the relevancy of the pages is higher	Reduces the computational complexity	Higher relevancy, the basis of topic character & time factor taken	User input is considered so relevancy of the pages is higher
Relevancy of the Pages	Medium	Less than PR	Higher than PR	More than PR	More than other and less to PR	More than PR	More than WPR
Limitations	Ranking is at indexing time, Query dependent, Emphasis on old pages	Query dependent, Topic drift and all links are considered of same important	Theme drift and efficiency problem	Query and user visit dependent, Them drift, emphasis on old pages	Query dependent, Relevancy is low, Them drift, emphasis on old pages	Query dependent, them drift	Query and user visit dependent, Them drift, efficiency problem

IV. CONCLUSIONS

In today's fast growing world, to get accurate information is more important and needful. On accomplishing theoretical analysis and comparison this paper concludes that Page Ranking needed more powerful algorithm which includes maximum relevancy in result pages as per user's nature.

ACKNOWLEDGMENTS

Bipin Vekariya would like to wholeheartedly thank Asst. Prof. Ankit Vaishnav for all his diligence, guidance, encouragement, inspiration and motivation throughout. I would like to thank anonymous reviewers for his valuable suggestions.

REFERENCES

[1] S.Brin and L.Page, "The Antonomy of a Large Scale Hyper textual Web Search Engine" 7th Int.WWW Conf. Proceedings, Australia, April 1998.

[2] 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):60–67, 1999.

[3] J. M. Kleinberg. "Authoritative sources in a hyperlinked environment". *Journal of the ACM*, 46(5):604–632, September 1999.

[4] C. Ding, X. He, H. Zha, P.Husbands and H. Simon "Link Analysis: Hubs and Authorities on the World" Technical Report: 47847, 2001.

[5] W.Xing and A.Gorbani, "Weighted PageRank Algorithm" Proceedings of the Second Annual Conference on Communication Networks and Services Research, May 2004, pp. 305-314.

[6] H. Dubey and Prof. B.N. Roy, "An Improved Page Rank Algorithm

based on Optimized Normalization Technique" International Journal of Computer Science and Information technologies (IJCSIT), 2011, pp.2183-2188.

[7] G.Kumar, N. Duhan and A.K. Sharma, "Page Ranking Based on Number of Visits of Web Pages" International Conference on Computer & Communication Technology (ICCCT), 2011, pp. 11-14.

[8] Wei Huang,Bin Li, "An Improved Method for the Computationof PageRank" International Conference on Mechatronic Science, Electric Engineering and Computer August 19-22, 2011

[9] N.Tyagi and S. Sharma,"Weighted Page Rank Algorithm Based on Number of Visits of Links of Web Page,"International Journal of Soft Computing and Engineerig(IJSCE),July 2012.