Comparative study of Page Ranking Algorithms for Web Mining

Shruti Aggarwal
Assistant Professor, Dept. of CSE, S.G.G.S.W.U., Fatehgarh Sahib (Punjab), India
Parneet Kaur
Research Scholar, Dept. of CSE, S.G.G.S.W.U., Fatehgarh Sahib (Punjab), India, 09815720369

Abstract: The World Wide Web is growing rapidly so there should be some means to provide information relevant to user. This need results in number of search engines that aims to provide information as per users need. Web search engines relies on various Page Ranking algorithms for finding suitable answers for user queries. In this review paper various Page Rank algorithms like PageRank, Distance PageRank and HITS used for Information Retrieval are discussed and compared. We compare the algorithm on the simulation interface.

Keywords: Web Mining, PageRank, Eigenvector, HITS, Distance Rank.

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:

- Web is enormous.
- Web pages are partially structured.
- Web information stands to be miscellany in meaning.
- Degree of quality of the in sequence extracted.
- Winding up of knowledge from information extracted.

There are 3 important components in a search engine. They are Crawler, Indexer and Ranking mechanism. The crawler is also called as a robot or spider that traverses the web and downloads the web pages. The downloaded pages are sent to an indexing module that parses the web pages and builds the index based on the keywords in those pages. An index is generally maintained using the keywords. When a user types a query using keywords on the interface of a search engine, the query processor component match the query keywords with the index and returns the URLs of the pages to the user. But before showing the pages to the user, a ranking mechanism is done by the search engines to show the most relevant pages at the top and less relevant ones at the bottom [2].

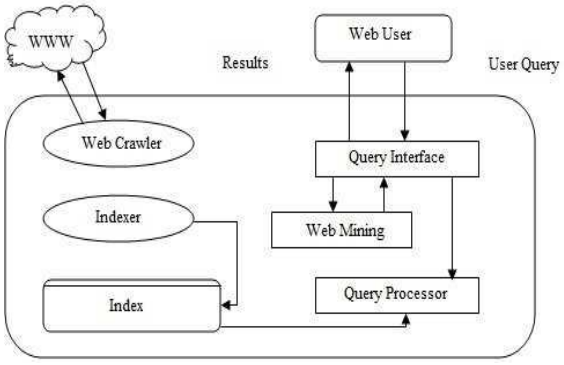


Figure: Architecture of Search Engine [2]

The motive behind this paper to analyze the popular web page ranking algorithms, their variations and to provide a comparative study of both and to highlight their relative strengths and limitations.

II. Ranking Algorithms

The web page ranking algorithms rank the search results on the basis of their relevance to the search query. PageRank algorithms rank the search results in descending order of relevance to the query string being searched. A web page's ranking for a specific query depends on factors like- its relevance to the words and concepts in the query, its overall link popularity etc. There are two categories of these algorithms viz. text based and link based [3].

1.) Text Based Ranking [4]

In this scheme pages are ranked purely on the basis of their textual content. In this ranking scheme factors that affect are:

- Number of matched words with query string
- *Location Factors* influence the rank of a page depending upon where the search string is located on that page. The search query string could be found in the title of a page or in the leading paragraphs of a page or even near the head of a page [4].
- *Frequency Factors* deal with the number of times the search string appears in the page. The more time the string appears, the better is the page ranking [4].

Most of the times, the affect of these factors is considered collectively. For example, if a search string repeatedly appears near the beginning of a page then that page should have a high rank [4].

2.) Link Analysis Algorithm [4]

Another popular class of ranking algorithms is the link-based algorithms. They view the web as a directed graph where the web pages represent the nodes and the hyperlinks between the web pages form the directed edges between these nodes [4]. Link-based ranking algorithms propagate page importance through links. The two most influential hyperlink based search algorithms are [4]:

- HITS (Hyperlink Induced Topic Search)
- PageRanking Algorithm

2.1) HITS Algorithm

HITS 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. 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.

In HITS algorithm[5], the hub and authority are calculated using the following algorithm:

1. Initialize all weights to 1
2. Repeat until the weights converge
3. for every p∈H
4. Hp=$\sum_{q \in Ip} A_q$
5. For every authority p∈ A
6. Ap=$\sum_{q \in Bp} H_q$
7. Normalize

There are two main 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 given query are collected. The next second step, Iterative step, finds hubs and authorities using the output of the sampling step using steps 4 and 7 from above algorithm.

Where H_p represents the hub weight, A_p represents the Authority weight and the set of reference and referrer pages of page p denote with respect to *I(p)* and *B(p)*. The weight of authority pages is proportional to the summation of the weights of hub pages that links to the authority page. Another one is, hub weight of the page is proportional to the summation of the weights of authority pages that hub links to. Figure below shows an example of the calculation of authority and hub scores [5].

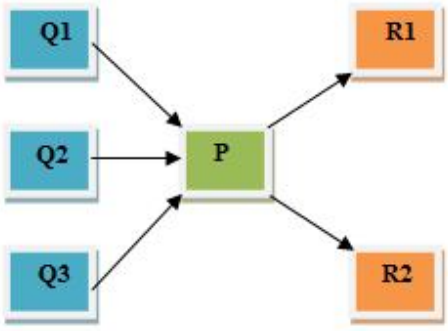


Figure: Collection of hubs and authorities [5]

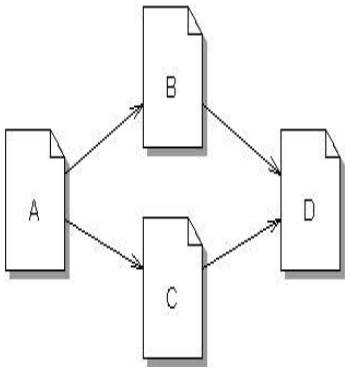
From the above equations of step 4 and 5 of algorithm, the hub and authority are calculated such as:

$$A_p = H_{Q1} + H_{Q2} + H_{Q3} \quad (1)$$

$$H_p = A_{R1} + A_{R2} \quad (2)$$

2.2) PageRank Algorithm

The PageRank algorithm, one of the most widely used page ranking algorithms, states that if a page has important links to it, its links to other pages also become important. Therefore, PageRank takes the backlinks into account and propagates the ranking through links: a page has a high rank if the sum of the ranks of its backlinks is high [6]. Figure 3 shows an example of backlinks: page A is a backlink of page B and page C while page B and page C are backlinks of page D.



An Example of Inlinks [6]

In PageRank algorithm web pages are represented as a link. The algorithm ranks the pages on the basis of how many important pages are referring that particular page.The page is given highest rank that are referred by important pages.

The simple procedure for calculating PageRank is as follows;

1 Construct the adjacency matrixA_{ij}, whose value is set to 1 if node i links to node j and 0 otherwise.

2 In order to make matrix row normal i.e. sum of each row of a matrix sums to 1.this can be done using:

$$H_i = \frac{A_i}{\sum_{k=1}^{n} A_{ik}}, \qquad (3)$$

3 Using the matrix H is insufficient for the PageRank algorithm, however, because the iteration using H alone might not converge properly — "it can cycle or the limit may be dependent on the starting vector". Part of the explanation for this is that the matrix H is not yet necessarily stochastic. A matrix is stochastic when it is a "matrix of transition probabilities for a Markov chain,"1 with the property that "all elements are non-negative and all its row sums are unity" (one). Thus, to ensure that H is stochastic, we must ensure that every row sums to one. Therefore, we define the stochastic S as

$$S = H + \frac{ae^T}{n} \qquad (4)$$

where a, is a column vector such that $a_i = 0$ if $\sum_{k=1}^{n} H_{ik} = 0$ (i.e., page i is a dangling node) and 0 otherwise, and e is a column vector of ones.

4 However, we are still not completely done, because there is no guarantee that S has a unique stationary distribution vector (i.e., there might not be a single "correct" Page Rank vector. For us to guarantee that there is a single stationary distribution vector π^T to which we can converge, we must ensure that S is irreducible as well as stochastic. A matrix is irreducible if and only if its graph is strongly connected, so we define the irreducible row-stochastic matrix G as

$$G = \alpha S + (1-\alpha)E, \qquad (5)$$

Where $0 \leq \alpha \leq 1$ damping factor and E = $\mathbf{ee}^T$ /n". G is the Google matrix

$$\pi^{(k+1)T} = \pi^k G \qquad (6)$$

Google uses the value of 0.85 for α, and we will use this same value. This will serve as the basis for further calculation of PageRank vector.

5 The PageRank algorithm can be iterated until the values converge sufficiently to the PageRank vector, i.e., until$\pi^{(k+1)T} \cong \pi^k G$. Thus, we conclude that the PageRank vector, the left Perron vector, and the stationary distribution vector of G are one and the same, and so finding the PageRank vector of G is equivalent to finding the dominant right eigenvector

of G^T , or the dominant left eigenvector (left Perron vector) of G.

We conclude that Page rank= largest left eigenvector of G i.e. Perron vector.

III. Weighted PageRank

Weighted Page Rank [7] Algorithm is proposed by **WenpuXing and Ali Ghorbani**. Weighted page rank algorithm (WPR) is a modification of the original page rank algorithm. The rank scores are decided based on the popularity of the pages by taking into consideration the importance of both the in-links and out-links of the pages. This algorithm does not equally divide the rank of a page among its out-link pages and provides high value of rank to the more popular pages. The rank value is given to every out-link page based on its popularity. Popularity of a page is decided by observing its number of in links and out links [5].

IV. Distance Rank Algorithm

A distance rank algorithm is proposed by **Ali Mohammad Zareh Bidoki and Nasser Yazdani**. This intelligent ranking algorithm based on reinforcement learning algorithm based on novel recursive method. In this algorithm, the distance between pages is considered as a distance factor to compute rank of web pages in search engine. The main goal of this ranking algorithm is computed on the basis of the shortest logarithmic distance between two pages and ranked according to them so that a page with smaller distance to assigned a higher rank [2]. The Advantage of this algorithm is that, being less sensitive, it can find pages faster with high quality and more quickly with the use of distance based solution as compared to other algorithms. If the some algorithms provide quality output then that has some certain limitations. So the limitation for this algorithm is that the crawler should perform a large calculation to calculate the distance vector, if new page is inserted between the two pages. This Distance Rank algorithm adopts the PageRank properties i.e. the rank of each page is computed as the weighted sum of ranks of all incoming pages to that particular page. Then, a page has a high rank value if it has more incoming links on a page [2].

Algorithms → Criteria	**HITS**	**PageRank**	**Weighted PageRank**	**Distance Rank**
Main Technique	WSM and WCM	WSM	WSM	WSM
Working Process	It computes the hubs and authority of the relevant pages.	This algorithm calculates the rank of pages using power method at the time of indexing.	Weight of webpage is calculated on the basis of incoming and outgoing links. From this weight or importance to page is assigned	Calculate the minimum average distance between two pages and more pages.
I/P Parameters	Inbound links and Outbound links and content	Inbounds links	Inbound links and Outbound links	Inbounds links
Complexity	<O(log N)	O(log N)	<O(log N)	O(log N)
Limitations	Topic drift & efficiency problem. Query dependent	Query independent	Query independent	Needs to work along with PR
Quality of results	Less than PR	Medium	Higher than PR	Less than PR

5 Comparison of various PageRanking Algorithms [2][3]

Table 1: Comparison of various PageRank Algorithms

VI. xConclusion

A typical search engine should use web page ranking techniques based on the specific needs of the users because the ranking algorithms provide a definite rank to resultant web pages. After going through this 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 results, it is concluded that existing algorithms have limitations in terms of

time response, accuracy of results, importance of the results and relevancy of results. The main purpose is to inspect the important page ranking based algorithms used for information retrieval and compare those algorithms. An efficient web page ranking algorithm should meet out these challenges efficiently with compatibility with global standards of web technology. The work applies the PageRank program in the Web, calculates PageRank values by PageRank algorithm and weighted PageRank values using Weighted PageRank algorithm.

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] Laxmi Choudhary and Bhawani Shankar Burdak,"Role of Ranking Algorithms for Information Retrieval", 2010.

[3] Nidhi Grover and Ritika Wason, "Comparative Analysis of PageRank and HITS Algorithms", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181Vol. 1 Issue 8, October – 2012.

[4] World Wide Web searching technique, Vineel Katipally, Leong-Chiang Tee, Yang Computer Science & Engineering Department Arizona State University.

[5] Neelam Tyagi and 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.

[6] S. Pal, V.Talwar, and P. Mitra," Web mining in soft computing framework: Relevance, state of the art and future directions", IEEE Trans. Neural Networks, 13(5):1163–1177, 2002.

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