Conceptual Analysis On Different Page Ranking Algorithms

Mr. BANDELA NARSINGAM
Asst.Professor, Dept.of CSE
TeegalaKrishnaReddy Engineering College
Hyderabad.

Mr. ANGOTH LAKSHMAN
Asst.Professor, Dept.of CSE
TeegalaKrishnaReddy Engineering College
Hyderabad.

Abstract: **As there is a huge growth in the central repository, retrieving the needed information is getting difficult. Search engines plays vital role in retrieving the information. While retrieving the information, all the search engines has to show the most promising results. For this purpose, search engines are using the page ranking mechanisms. There exists different page ranking algorithms where we will be analyzing and making a review of them. This report is prepared based on the review of different page ranking algorithms.**

Keywords: **Search Engines; Page Ranking;**

Introduction to Search engines: Most of the people using internet to have the service of Search Engines. Out of all the search engines Google, Yahoo and MSN plays a vital role and these search engines occupy 81% of the market in the search engines[1]. Out of the above percentage Google occupies major portion that is 46%. Now it is important that whether the search engines are satisfying the user needs or not. To have the better results or personalized results search engine has to use different techniques like search engine optimization, personalization, page ranking etc. To have the effective and related results related to the search queries search engines uses page ranking approaches. In this paper let us review different page rank algorithms and their implementation.

Page Rank Algorithm: Day by day information is growing with huge speed on the internet. Searching the information in the net will become a difficult task, in that getting related results is another important task. To achieve this search engines are using page ranking mechanisms. Google is a search engine which uses page ranking mechanism to give the better results to the user. The functionality of search engine is divided into 3 functions. They are Crawler, Ranking mechanism and Indexing. Crawler is a program which goes through the web and downloads the web pages. All these pages which are downloaded are sent to the indexing module so that an index is generated based on the keywords of the web pages. Whenever the user enters the search query, this search query will break down into keywords and whatever the keywords matched those documents will be presented to the user. But there may be so many pages which matches the keywords of the index which gives huge results. Sometimes all these results may not be useful to the user. To overcome these situations search engines uses page ranking approach. The resulted pages will undergo ranking mechanism and assigns the ranks to the pages by the search engine. The pages which are having highest rank will appear on the top of the results and less rank pages will follow them. It is clearly explained in the figure 1 that the functionality of the search engines is divided into two phases. First one is front end in which end user, search query, search interface, Query engine and ranking mechanism are the components of Front end process. In the Back end, the components are Crawler, Indexer, Repository and Document index are the main components. The process is clearly explained and represented by using the following figure 1.

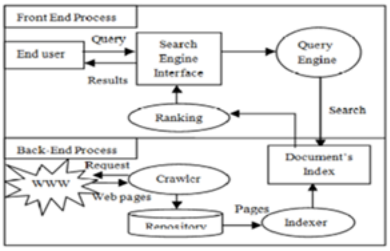

Figure 1.Architecture of the search engine.

Need of Page Ranking: If the user did not get the relevant results then nobody uses the search engines. To gain the popularity, search engine has to give the required results to the user. To achieve this page ranking mechanism should be implemented by the search engines. The following are Page Ranking algorithms which are presented in this paper. They are HITS Algorithm, Weighted Page Rank Algorithm, Distance RankAlgorithm and Eigen Rumor Algorithm and their role in information retrieval.

HITS Algorithm: The acronym for HITS is Hyper-link Induced Topic Search algorithm. This is one of the page ranking algorithm which uses in links and out links to rank the web page. We use two different terms like authority and hub. Authority is defined as, if any of the web page is pointed by more hyperlinks it is known as authority and if the web page is pointed to many hyperlinks then it is known as hub. Here hubs are used as the main resource for the algorithm and authority can be used for the content. A page may contain both hub and authority, then it can be called as a good page. An iterative algorithm is needed to calculate the weights of the authority and hub. Hits algorithm is explained in two main steps, First one is to collect the web pages which are related to the query. Second step is the iterative step where hubs and authorities are found in the output of the first step and it should continue by taking it as the input. Following expression is used to calculate the weights of the Hub and Atuthority.

The following figure gives the information about how the Authority pages and Hub pages are calculated.

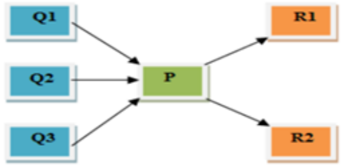

Figure 2: Calculations of Authority and Hubs.

$$\boldsymbol{HUB(H_p) = \Sigma_{q\varepsilon I_p} A_q}$$

$$\boldsymbol{AUTHORITY\ (A_P) = \Sigma_{q\varepsilon B_p} H_q}$$

Where Aq is the Authoritative score of the page and Hq is the hub score of the page. Ip is the set of reference page of page P and Bp is the set of referrer pages of page P. Hubs and Authorities are calculated by using the following expressions.

$$\boldsymbol{A_p = H_{Q1} + H_{Q2} + H_{Q3}}$$

And

$$\boldsymbol{H_p = A_{R1} + A_{R2}}$$

HITS algorithm will generate ranking pages based on the scores and generates authority and hub pages.

HITS Algorithm:

Steps:

1. **Initialize all weights to 1.**
2. **Repeat until the weights converge**
3. **For every hub pε H**
4. $\boldsymbol{Hp = \Sigma_{q\varepsilon Ip} A_q}$
5. **For every authority pεA**
6. $\boldsymbol{A_p = \Sigma_{q\epsilon Bp}}$**Hq**
7. **Normalize**

HITS is applied on a subgraph after a search is done on the complete graph.HITS defines hubs and authorities recursively. PageRank is used for ranking all the nodes of the complete graph and then applying a search. PageRank is based on the 'random surfer' idea and the web is seen as a Markov Chain.Power Iteration an efficient way to calculate with sparse matrix.

Weighted Page Rank Algorithm:

Weighted Page Rank: This algorithm is also similar to the page rank algorithm and have some extensions in this. This algorithm also contains both inlinks and out link, These links are assigned some weight based on the page rank priority. The incoming and outgoing links weight values are denoted by $w^{in}(m,n)$ and $w^{out}(m,n)$ respectively. $w^{in}(m,n)$ is calculated based on the number of inlinks to the page m and similarly $w^{out}(m,n)$ is calculated based on the number of outlinks of page m.

$W^{in}(m,n) = = \frac{In}{\Sigma P\varepsilon R(m)Ip}$

Where In represents the number of inlinks of n pages and

Ip represents the number of inlinks of page p respectively.

R(m) Represents the reference page list of page m.

Win(m,n) = $On/\Sigma P\varepsilon R(m)Op$

Where On represents the number of out links of n pages and

Op represents the number of out links of page p respectively.

R(m) Represents the reference page list of page m.

Now the page rank formula is modified as

$$WPR = (1-d) + d\Sigma_{meB} WPR(m) w^{in}(m,n) w^{out}(m,n)$$

The above equation shows that this algorithm is different from page rank algorithm. The technique used in weighted page rank algorithm is Web structure mining. Weight of web page is calculated on the basis of in links and out links and the weight of that page. The input parameters for the page rank algorithm are Back links and forward links. The results of this algorithm is better than the page rank algorithm but the limitation is that it ignores the relevancy of the search query.

Distance Ranking Algorithm: This is another algorithm which is also use to calculate the page ranks, it is proposed by Ali Mohammad ZarehBidoki and Nasser Yazdani. It is also known as intelligent algorithm where all the pages are considered as distance factors and calculate the distances of the pages through the help of search engine. Ranking is implemented by knowing the distance between the pages and based on that the rankings are applied for the pages. The values taken here for rankings are the shortest logarithmic values which represents the distance between the pages. The advantage of this algorithm is to find the pages within short time when compared to other alogrithms.This algorithm also implements the properties of the page rank algorithm. The page which is having more incoming links will have the highest page rank value.

$$Distance\, n[j] = (1-a) * Dist_{n-1}[j] + a * \min i(a * Distance_{n-1}[i] + \log(0[i])), i\varepsilon B(j)$$

Where B(j) represents the list of pages which links to the page J and

O(i) represents the number of out links in page i. After finding this all the url which are gathered will be sorted in an ascending order.

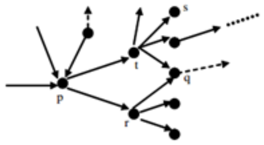

Figure 3: A portion of a graph. The distance between p and q is log(2) + log(3).

EigenRumor Algorithm: This algorithm is used for ranking the blog entry by using the weights of authority and hub pages. These weights or scores are taken by using the eigen vector calculations. Now a days there is a huge growth in the blogs, to have the effective results in providing the correct blogs to the user is a challenging task. To implement the concept of page rank and HITS on the blogs, some issues are raising. To overcome all those issues this algorithm is used. This algorithm is proposed by Ko Fujimura which calculates three vectors, they are authority vector A, a hub vector H and reputation vector R, information provisioning matrix P and information evaluation matrix E.

The issues which are concentrated by this algorithm are:

The in links to a blog are usually small. Similar to the calculation of in links in the page rank algorithm, here also it calculates the number of entries into the blogs. All these entries are calculated by page rank. But these ranks are very low and are not given according the importance of the page. These issues are overcome in this EigenRumour algorithm which is used have the connections with pageRank and HITS. All these use eigenvector calculations to have the adjacency matrix of the links. One important thing is anagent is used to represent an aspect of human being suchas a blogger, and an object is used to represent any object such as a blog entry.

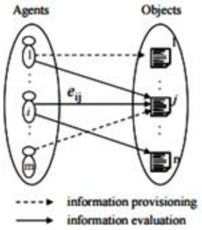

Figure 4: EigenRumor community model

Analysis of Different Page Rank Algorithms:

S. No	Algorithm Name	Approach used	Input parameters	complexity	Limitations
1.	Page Rank Algorithm	Page ranks are calculated by using the search query. All are divided into key words and counts the matches of the keyword. This process will be done at the time of the indexing.	Back links are taken as the input parameters	O(log n)	1) Results will not be generated at the query time, They are generated at the time of indexing. 2) Accuracy of results is at medium level. 3) User satisfaction will not be more.
2.	Weighted Page Rank Algorithm	This method uses the web structure mining approach where it calculate weights of the pages based on the	In links and out links both are used to know the weight of the	< O(log n)	1. It is not concentrating on the required results. 2. Quality of results are more than the page rank

		in links and out links of the page. With the help of these weights the importance of this page	page		algorithm.
3.	HITS Algorithm	It uses the concepts like web structure and content mining. It uses the Authority pages and hub pages to calculate the rank of the page.	It uses in links and out links as well as the content is also taken into consideration for the rank of the page.		1. It will have the efficiency problem. 2. It gets less results than page rank algorithm 3. Topic drift is another limitation.
4.	Distance Rank Algortihm	It uses the distance factor between the pages. These values are logarithmic values, by using these values it computes all the weights.	Forward links are the input parameters for this algorithm.	< O(log n)	1. Whenever a new page inserted between two pages then the crawler should perform a largecalculation to calculatethe distance vector.
5.	Eigen Rumour Algorithm	Adjacency matrix is constructed from the agent and object link. Page to page links are not used over here.	Input parameters are Agent and objects.	O(log n)	1. Mainly used for blog rankings. 2. Can not be implemented to calculate the page ranks. 3. Quality will be higher than the Page Rank and HITS.

Different types of data like text, images, audio, video type of data will be there called as heterogeneous type of data. To have the heterogeneous type of data, these algorithms will take the input as annotations of other type of data like images and other multimedia type of files. Annotations are the descriptions which are helpful in searching the heterogeneous data. All these search engines will take the input as the text or keywords. There are different categories are available to search the data like images, text, web etc.

REFERENCES

[1]. Danny Sullivan, Nielsen NetRatings search engine ratings, Search Engine Watch, January 2006.

[2]. Brin, S. and Page, L. (1998) The Anatomy of a Large-Scale Hypertextual WebSearch Engine. In: Seventh International World-Wide Web Conference (WWW1998), April 14-18, 1998, Brisbane, Australia. Retrieved from http://infolab.stanford.edu/~backrub/google.html

[3]. LaxmiChoudhary and Bhawani Shankar Burdak, (9 Aug 2012) Role of RankingAlgorithms for Information Retrieval Cornell University. Retrieved from http://arxiv.org/ftp/arxiv/papers/1208/1208.1926.pdf