Ranking Websites in Search Engines Using Big Data Platform

Kavya[1], Sahana K A[2], Shwetha S U[3], Supreetha[4], Kavyasri M N[5]
[1, 2,3,4,5] *Department of Computer science and engineering, Malnad College of Engineering, Hassan*

***Abstract*- WWW includes billions of web pages and a large amount of information available within the pages. To recover the information required by the WWW, search engines perform various activities depending on their architecture. These can be complicated and slow process. Current estimates indicate that there are more than 150 million Web pages with a useful life of less than a year. In addition to these important challenges, web search engines must also manage users and inexperienced pages designed to manipulate search engine ranking features. Implementing techniques used by search engines to rank the output, test them and analyze their performance. Given a query, a search engine needs to rank documents by relevance so that links to most relevant and authoritative documents could be shown on top of the list. Search engine uses many algorithms to rank the Web Pages among which we are using PageRank algorithm. Page Rank is an algorithm used by Google Search engine to rank website pages in their search results. PageRank is the way of measuring the significance of website pages and it works by counting the number and quality of links to a page to determine a rough estimate of how important the website is .The classical Page Rank algorithm assigns rank to a website on the basis of other links connected to it, so that pages linked by many other gets high rank .The search engine uses a classification algorithm to sort the results that will be displayed. In this way, the user will first have the most important and useful result.**

***Index terms*- PageRank, Search engine, Web Pages, Damping factor, in links, out links**

I.INTRODUCTION

Every search engine uses its own algorithm to rank the WebPages and making sure that only relevant results are returned for the query entered by the user and the result for a specific query is then shown on the search engine result page. To rank the WebPages, in this paper we are considering one of the trademark algorithm called PageRank which is owned by Google. PageRank works by counting the number and quality of links to a page to settle on a rough estimate of how essential the website is. The fundamental assumption is that more important websites are liable to receive more links from other websites.

The PageRank algorithm outputs a probability sharing used to embody the probability that a person at random clicking on links will appear at any particular page. PageRank can be designed for collections of documents of any size. It is assumed in several research papers that the sharing is evenly divided among all documents in the collection at the start of the computational process. The PageRank calculations need several "iterations", through the compilation to adjust estimated PageRank values to more closely reflect the theoretical accurate value. By using this PageRank algorithm we can calculate the rank score for WebPages and whichever gets the highest rank it will shown on the top of the list in search engine result page.

II. SURFER MODEL CALCULATION

The probability of clicking the link in search engine result page is expressed as a numeric value between 0 and 1. A probability of 0.5 is commonly expressed as a "50% chance" of something happening. Therefore, a Page Rank of 0.5 means there is a 50% chance that a person clicking on a link at random will be directed to the document with the 0.5 Page Rank.

Let us assume a small universe of four web pages: A, B, C and D. Links from a page to itself be neglected. Page Rank is initialized to the same value for all pages. In the original form of Page Rank, the sum of Page Rank over all pages was the total number of pages on the web at that time; therefore each page in this instance would have an initial value of 1.

The Page Rank shared from a given page to the targets of its out-bound links upon the subsequently iteration is assigned equally among all outbound links.

PR(A) = PR(B) + PR(C) + PR(D)

Where PR(A), PR(B), PR(C) and PR(D) are Page rank of node A, B, C and D respectively.

Assume that page B had a connect to pages C and A, page C had a connect to page A, and page D had links to all three pages. After the completion of this iteration, page A will have a Page Rank as defined by this equation

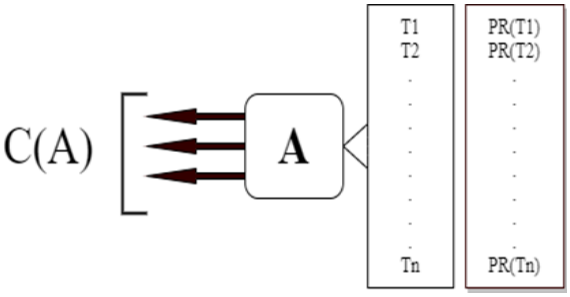
PR (A) = PR (B)/2 + PR(C)/1 + PR (D)/3

In other words, the Page Rank conferred by an outbound link is equal to the document's own Page Rank score divided by the number of outbound links L( ).

PR (A) = PR (B)/L(B) + PR(C)/L(C) + PR(D)/L(D)

Damping Factor

The Page Rank hypothesis holds that an imaginary surfer who is arbitrarily tapping on connections will in the end quit clicking. The likelihood, at any progression, that the individual will proceed a damping factor d.



The PageRank of page A is given as follows

PR(A)= 1-d+ d(PR(B)/L(B) + PR(C)/L(C) + PR(D)/L(D) + …)

Google recalculates Page Rank scores each time it creeps the Web and revamps its file. As Google builds the quantity of archives in its accumulation, the underlying estimation of Page Rank diminishes for all records.

The equation utilizes a model of an arbitrary surfer who gets exhausted after a few ticks and changes to an irregular page. The Page Rank estimation of a page mirrors the opportunity that the irregular surfer will arrive on that page by tapping on a connection. It very well may be comprehended as a Markov chain in which the states are pages, and the advances, which are on the whole similarly likely, are the connections between pages. In the event that a page has no connections to different pages, it turns into a sink and along these lines ends the irregular surfing process. On the off chance that the irregular surfer lands at a sink page, it picks another URL indiscriminately and keeps surfing once more.

When figuring Page Rank, pages with no outbound connections is expected to connection out to every single other page in the gathering. Their Page Rank scores are along these lines separated uniformly among every single other page. At the end of the day, to be reasonable with pages that are not sinks, these irregular advances are added to all hubs in the Web. This remaining likelihood, d, is normally set to 0.85, evaluated from the recurrence that a normal surfer utilizes his or her program's bookmark highlight. In this way, the condition is as per the following:

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

Where $p_1, p_2 \ldots, p_N,$ are the pages under consideration, $M(p_i)$ is the set of pages that link to p_i, $L(p_j)$ is the number of outbound links on page p_j, and N is the total number of pages.

The Page Rank esteems are the passages of the overwhelming right eigenvector of the changed nearness grid rescaled so every segment means one. This makes Page Rank an especially exquisite measurement: the eigenvector is

$$R = \begin{bmatrix} PR(p_1) \\ PR(p_2) \\ \vdots \\ PR(p_N) \end{bmatrix}$$

Where R is the solution of the equation.

$$R = \begin{bmatrix} 1-d/N \\ 1-d/N \\ \vdots \\ 1-d/N \end{bmatrix} + d \begin{bmatrix} l(p_1,p_1) & l(p_1,p_2) & \ldots & l(p_1,p_N) \\ l(p_2,p_1) & \ddots & & \vdots \\ \vdots & & l(p_i,p_j) & \\ l(p_N,p_1) & \ldots & & l(p_N,p_N) \end{bmatrix}$$

Where the adjacency function $l(p_i,p_j)$ is the ratio between number of links outbound from page j to page i to the total number of outbound links of page j. And link is 0 if page p_j does not link to p_i, and normalized such that, for each j

$$\sum_{i=1}^{N} l(p_i,p_j) = 1$$

I.e. the components of every segment aggregate up to 1, so the framework is a stochastic lattice (for more subtleties see the calculation area underneath). Along these lines this is a variation of the eigenvector centrality measure utilized regularly in system examination.

Google's organizers, in their unique paper, revealed that the Page Rank calculation for a system comprising of 322 million connections (in-edges and out-edges) unites to inside a middle of as far as possible in 52 cycles. The assembly in a system of a large portion of the above size took roughly 45 cycles. Through this information, they closed the calculation can be scaled great and that the scaling factor for very huge systems would be generally direct in log n, where n is the size of the system.

III. LITERATURE SURVEY

A. Shaojie qiaot – SimRank method

SimRank is general similarity measure, based on a simple and intuitive graph-theoretic model. SimRank is valid in any sphere with object to object relationships. This similarity measure computes similarity between pages and uses it to several web social networks. They proposed a weighted PageRank algorithm namely SimRank, which considers the relevance of a page to the given query which can improve accuracy of scoring.

The advantage of using this technique is, it works in an offline fashion so the scores of pages depend on previously given pages and by using this technique it also disadvantage that, it cannot identify whether the hyperlinks of distinct pages are content correlated or not.

B. Dr.Daya gupta--User preference based rank

User preference based page ranking regulates search quality and makes user search navigation experience in the result of search engine. It uses agents to determine pages context relevancy and consider user behavior while ranking the WebPages. This method is proposed with employs web structure, web usage and web context mining techniques to order the WebPages with the help of agents and specialized crawler and to determine the relevancy of page according to given query.

The advantage of using this method is, it avoids similarity of ranking and is more dynamic in nature thereby providing users an effective method to measure the page quality and also it is more target oriented because it considers users usage trends. But the disadvantage by using this method is it has draft and requires specialized crawler.

C. Rekha singhal – weight In link PageRank

This technique dividing the weight of an in-linked WebPages distributes it to all the out-linked pages on the basis of their popularity. Weight in-link uses weights of an in-linked WebPages to calculate a new score of every individual WebPages called weight in-link score. Here in-link means the links are pointing to your website from other websites also called backlinks and outlink means the links that are pointing outwards the considered webpage. Weightage of in-linked WebPages are used to compute the rank the WebPages. This technique concentrates on structure mining of web based on the weightage of in-linked WebPages. The main objective of this proposed idea is to provide more rank to related WebPages according to their popularity in the positions of the search results.

The advantage of using this method is, it provides higher relevancy in web search results compared to search results generated by original standard PageRank algorithm. Importance of page is decided by considering weights of inlinks and outlinks. But the main disadvantage using this technique is, it ranks the web pages only on basis of popularity.

D. Fayyaz Ali-- Ratio-based weighted page rank

Fayyaz Ali and Irfan Ullah introduce a Ratio-based weighted page rank algorithm they extends the original page rank algorithm, however it neither divide the importance score of a page evenly among its outgoing links nor consider the edges coming in and leaving out of a receiving node. Rather they use the ratio based approach to divide the PageRank of a node among its referenced nodes so that every nodes receive its own share from the referee node according to its weightages. This will make sure that an important page will get rightful share from the total PageRank of the referee node.

The Advantage is Quality of the pages returned by this is high as compared to page rank algorithm. It is more efficient than PageRank because rank value of a page is divided among its outlink pages according to important of the page and limitation existing in the

form of dangling links which occurs when a page contains a link such that the hypertext points to a page with no outgoing links, and it leads to rank sink trouble occurs when in a network pages get in endless link cycles.

IV. RESULTS AND DISCUSSIONS

Proposed method	Feature	Description	Input parameters	Mining techniques
WIL PR (Weight age InLink page rank)	This gives higher relevancy in web search results than original standard page rank	Irrespective of dividing the weight of an in-linked webpage, this method distributes it to all the out linked pages on the basis of their popularity and finally a new score of every individual webpage is calculated and web pages are ranked accordingly.	In-link, Out-link	Web structure mining
User preference based page ranking	User preference based Page Rank algorithm regulates search quality &makes user search navigation experience in the results of a Search Engine.	It uses agents to determine pages context relevancy and considers user behavior while ranking web pages.	In-link, Out-link, agent and user visit counts.	Web structure, content and usage mining.
Ratio based Weighted Page Rank	It calculates ranking of web pages in terms of convergence. Convergence is the number of steps taken to get the ranks stabilized , so the change in the coming iterations is minimal	Ratio Rank approach divides the page rank of a node among its referenced nodes so that every node receives its own share from the referee node according to its weightage.	Backlinks and forward links.	Web structure mining
SimRank	Traditional page rank algorithm is improved by assigning a probability of browsing a page to be initial page rank value of each paper	It is based on similarity measure named to score web pages. This measure computes the similarity between pages &uses it to partition a web database into several web social networks.	Backlinks	Web structure mining

Table 1: comparison based on characteristics

There are many algorithms to rank the webpages in search engine result page so in our paper we are considering Google's PageRank algorithm to rank the websites. Page Ranking is the position that your websites is listed in Google when a user searches that phrase or keyword. The dataset is taken consisting of nodes and edges. By taking the damping factor, we get a page rank of the nodes. A grade of 1, means to you are on the top. But you are number 11, that means you are on page 2 of Google as most search outcome have 10 listings for every page. Ranking is set in the highest order.

V. CONCLUSION

Page Rank Algorithm is the most popular algorithm used as the basis for the very popular Google search engine. Links are the way for users to get relevant information on the internet. They are also the ranking factor for every page. So, to rank the Websites Google's search engine is a powerful tool. It uses a trademarked algorithm called PageRank, which considers not only the backlines of the webpage but link quality as well, and then by calculating PageRank, it assigns each webpage a relevancy score or value. PageRank is used mainly for Internet use like, Twitter and web crawler etc. So ultimately it used to determine the rough estimation of how important a website is.

REFERENCES

[1] Sonal Tuteja ,"Enhancement in Weighted PageRank Algorithm using VOL", IOSR Journal

of Computer Engineering (IOSR-JCE) ,e-=ISSN: 2278-0661, Volume 14, Issue 5(Sep – Oct, 2013), PP 135-141

[2] 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 (IJSCE), ISSN: 2231- 2307, Volume-2, Issue-3, July 2012

[3] Shavjie Qiaot, Yianrui Li, Hong Li Yan Zhu, "SimRank: A Page Rank Approach based on Similarity Measure", 978-1-4244-6793-8/2010, IEEE

[4] Fayyaz Ali, Irfan Ullah, Shah Khusro, "An Empirical Investigation of Page Rank and its variants in Ranking Pages on the Web", 978-1-5090- 5300-1/2016, IEEE

[5] Dr. Daya Gupta, Devika Singh, "User Preference Base Page Ranking Algorithm", ICCCCA 2016, ISBN: 978-1-5090-1666-2/2016, IEEE

[6] Y. Nawaz Ahmed Khan, Bhaskaran Raman, Shanmugasundaram Hariharan, "OCS- A system for optimizing, Clustering and Summarizing web Search Results using Intelligent Agents", 978-1-4211-4/2009, IEEE

[7] Rekha Singhal, Saurabh Ranjan Srivastava, "Enhancing the Page Ranking for Search Engine

[8] Page Ranking Algorithms: A Survey Neelam Duhan, A. K. Sharma, Komal Kumar Bhatia YMCA Institute of Engineering, Faridabad, India, 2009 IEEE International Advance Computing Conference (IACC 2009)

[9] Page Ranking Based on Number of Visits of Links of Web Page By Gyanendra Kumar , Neelam Duhan , A. K Sharma, Department of Computer Engineering, YMCA University of Science & Technology, Faridabad, India, International Conference on Computer & Communication Technology (ICCCT)-2011

[10] ALLAN BORODIN University of Toronto GARETH O. ROBERTS Lancaster University JEFFREY S. ROSENTHAL University of Toronto and PANAYIOTIS TSAPARAS University of Helsinki.

[11] Elizabeth A. Hobson; Dan Mønster & Simon DeDeo (16 Oct 2018). "Strategic heuristics underlie animal dominance hierarchies and provide evidence of group-level social knowledge". ArXiv: 1810.07215 [q-bio.PE].

[12] ENGG2012B Advanced Engineering Mathematics Notes on Page Rank Algorithm Lecturer: Kenneth Shum.

[13] Matt Cutts's blog: Straight from Google: What You Need to Know Archived 2010-02-07 at the Wayback Machine.

[14] Koby Crammer kobics@cs.huji.ac.il Yoram Singer singer@cs.huji.ac.il School of Computer Science and Engineering, Hebrew University, Jerusalem 91904, Israel.

[15] Meenakshi Bansal, Deepak Sharma, ,"Improving webpage visibility in Search Engines by enhancing keyword Density using improved On-Page Optimization technique", IJCSIT, 5347-5352, ISSN0975-9646, 2015 .

[16] A. Jain (2013), "The Role Of Off-Page Search Engine ranking, International Journal of Advanced Research in Computer Science and software Engineering Vol 3, Issue 6, pp. 239-244