Search Engine Optimization and PageRank Algorithms

Nilkanth Satish Shet Shirodkar – 14103 M.Tech

Abstract: World Wide Web is a huge repository of information resources that include text, audio, video etc. As the amount of information available on web is increasing it is difficult to acquire genuine information on web. Therefore users today mainly depend upon various search engines for finding suitable answers for their queries. Search engines may return millions of pages in response to a query. It is not possible for a user to preview all the returned result set. So search engine make use of ranking algorithm to display the resultant pages in a ranked order using different page ranking algorithms. SEO is the process of getting traffic from search engines such as Google, Yahoo and Bing. local listings are shown and ranked based on what the search engine considers most relevant to users.In this paper, we will consider most popular Link based ranking algorithms namely PageRank algorithm. Relative strengths and limitations of these algorithms are explored to find out further scope of research.

Keywords: Search Engine, Worldwide web, PageRank, Ranking algorithm, Search Engine Optimization

I.INTRODUCTION

World Wide Web is a vast resource of hyperlinked and heterogeneous information including text, audio, video and metadata. It is estimated that WWW is doubling in size every six to ten months. Due to the rapid growth of information resources on World Wide Web it is difficult to manage the information on the web. Therefore it has become necessary for the users to use efficient information retrieval techniques to find and order the desired information. Search engines play an important role in searching web pages. The search engine gathers, analyzes, organizes the data from the internet and offers an interface to retrieve the network resources. Search engines are "programs" that search documents for specified keywords and returns a list of the documents where the keywords were found. The search results are generally presented in a line of results often referred to as search engine results pages (SERPs).The major components of a Search Engine are the Crawler, Indexer,
Query Processor. A crawler or spider is a program that traverses the web by following hyperlinks and storing downloaded pages in a large database. The crawler starts with a seed URL and collects documents by recursively fetching links and storing the extracted URL‟s into a local repository. Indexer extracts the terms from each web page and records the URL where each word has occurred.Query Processor is responsible for receiving and filling search requests from the user.When a user fires a query, query engine searches the web page in the index created by the indexer and returns a list of URL‟s of the web pages that match with the user query.

In general Query Engine may return several hundreds and thousands of URL in response to a user query which includes a mixture of relevant and irrelevant information. Since no user can read all web pages returned in response to the user query, Page Ranking mechanisms are used by most search engines for putting the important pages on top leaving less important in the bottom of the result list. Popular Page Ranking algorithms used are Page Rank alogorithm, Hypertext Induced Topic Search (HITS), Weighted Page Rank algorithm, Page Content Rank etc.

II.RANKING ALGORITHMS

Web-page ranking is an optimization technique used by search engines for ranking hundreds and thousands of web pages in a relative order of importance. To rank a web page different criteria are used by ranking algorithms. For example some algorithms consider the link structure of the web page while others look for the page content to rank the web page. Broadly Page Ranking algorithms can be classified into two groups Content-based Page Ranking and Connectivity-based Page Ranking.

Content-based Page Ranking: In this type of ranking the pages are ranked based on their textual. Factors that influence the rank of a page are :

- Number of matched terms with the query string
- Frequency of terms i.e the number of times the search string appears in the page. The more time the string appears, the better is the page ranking
- Location of terms i.e 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.

Connectivity-based Page Ranking (Link based): This type of ranking work on the basis of link analysis technique. They view the web as a directed graph where the web pages form the nodes and the hyperlinks between the web pages form the directed edges between these nodes. There are two famous link analysis methods:

- PageRank Algorithm
- HITS Algorithm and

III. PAGERANK

The PageRank algorithm was developed at Stanford University by Larry Page and Sergey Brin in 1996. PageRank algorithm , named after Larry Page and used by the Google Internet search engine uses the link structure of the web to determine the importance of the web page. Here a page obtains a higher rank if sum of its back-links is high. This algorithm is based on random surfer model. The random surfer model assumes that a user randomly keeps on clicking the links on a page and if he get bored of a page then switches to another page randomly. Thus, a user under this model shows no bias towards any page or link. PageRank(PR) is the probability of a page being visited by such user under this model. For each web page, Page Rank value is pre-computed . For this over 25 billion web pages on the WWW are considered to assign a rank value.

A Simplified version of PageRank is defined in Equation

$$PR(A) = (1-d) + d\left(\frac{PR(T_1)}{Q(T_1)} + \dots\dots\frac{PR(T_n)}{Q(T_n)}\right)$$

Where:

PR(A) = PageRank of page A

T1….Tn= All pages that link to page A

PR(Ti) = Page rank of page Ti

Q(Ti) = the number of pages to which Ti links to

d = damping factor which can be set between 0 and 1

PR(Ti)/Q(Ti) = PageRank of Ti distributing to all pages that Ti links to.

(1-d) = To make up for some pages that do not have any out-links to avoid losing some page ranks

Damping factor: The PageRank theory holds that any imaginary surfer who is randomly clicking on links will eventually stop clicking. The probability, at any step,that the person will continue is called a damping factor d. The damping factor can be set to any value such that 0<d<1, nominally it is set around 0.85. The damping factor is subtracted from 1 and this term is then added to the product of the damping factor and the sum of the incoming PageRank scores.

A. Implementation of Page Rank Algorithm

The following steps explain the method for implementing

Page Rank Algorithm.

Step 1: Initialize the rank value of each page by 1/n.
Where n is total no. of pages to be ranked.
Suppose we represent these n pages by an Array of n elements. Then
A[i] = 1/n where 0≤ i< n

Step 2: Take some value of damping factor such that 0<d<1. e.g 0.15, 0.85 etc.

Step 3: Repeat for each node i such that 0≤ i< n. Let PR be an Array of n element which represent PageRank for each web page.
PR[i] <= 1-d
For all pages Q such that Q Links to PR[i] do
PR[i] <= PR[i] + d * A[Q]/Qn
where Qn = no. of outgoing edges of Q

Step 4: Update the values of A
A[i]= PR[i] for 0≤ i< n
Repeat from step 3 until the rank value converges i.e. values of two consecutive iterations match.

B. Advantages of PageRank

The strengths of PageRank algorithm are as follows:

- **Less Query time:** PageRank has a clear advantage over the HITS algorithm, as PageRank compute ranking at crawling time so response to user query is quick.
- **Less susceptibility to localized links:** Furthermore, as PageRank is generated using the entire Web graph, rather than a small subset, it is less susceptible to localized link.
- **More Efficient :** In contrast, PageRank computes a single measure of quality for a page at crawl time. This measure is then combined with a traditional information retrieval score at query time. Compared with HITS, this has the advantage of much greater efficiency.
- **Feasibility:** As compared to Hits algorithm the PageRank algorithm is more feasible in today's scenario since it performs computations at crawl time rather than query time..

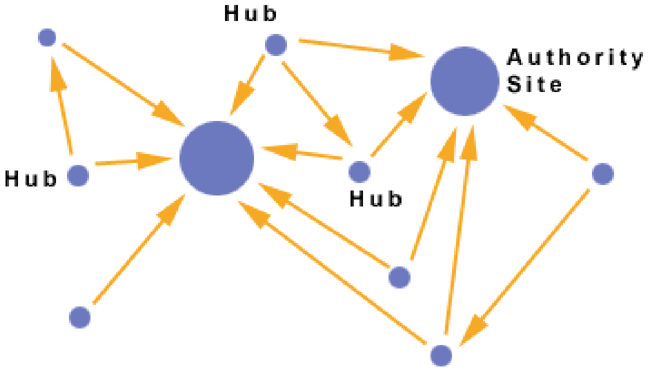
C. Disadvantages of PageRank

The following are the problems or disadvantages of PageRank:

- Less Relevant to user Query: PageRank score of apage ignores whether or not the page is relevant to the query at hand.
- Rank Sinks: The Rank sinks problem occurs when in a network pages get in infinite link cycles
- Spider Traps: Another problem in PageRank is Spider Traps. A group of pages is a spider trap if there are no links from within the group to outside the group.
- Dangling Links: This occurs when a page contains a link such that the hypertext points to a page with no outgoing links. Such a link is known as Dangling Link.
- Dead Ends: Dead Ends are simply pages with no outgoing links. PageRank doesn't handle pages with no outedges very well, because they decrease the PageRank overall.
- Circular References: If you have circle references in your website, then it will reduce your front page‟s PageRank.
- In Internet, available data is huge and the algorithm is not fast enough.

HITS Algorithm

HITS algorithm 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 named as HUB if the page point to various hyperlinks.



HITS is technically, a link based algorithm. In HITS algorithm, ranking of the web page is decided by analyzing their textual contents against a given query. After collection of the web pages, the HITS algorithm concentrates on the structure of the web only, neglecting their textual contents.

Original HITS algorithm has some problems which are given below.
(i) High rank value is given to some popular website that is not highly relevant to the given query.
(ii) Drift of the topic occurs when the hub has multiple topics as equivalent weights are given to all of the outlinks of a hub page. Figure 5 shows an Illustration of HITS process.

To minimize the problem of the original HITS algorithm, a clever algorithm is proposed by reference. Clever algorithm is the modification of standard original HITS algorithm.This algorithm provides a weight value to every link depending on the terms of queries and endpoints of the link. An anchor tag is combined to decide the weights to the link and a large hub is broken down into smaller parts so that every hub page is concentrated only on one topic. Another limitation of standard HITS algorithm is that it assumes equal weights to all the links pointing to a webpage and it fails to identify the facts that some links may be more important than the other. To resolve this problem, a probabilistic analogue of the HITS (PHITS) algorithm is proposed by reference. A probabilistic explanation of relationship of term document is provided by PHITS. It is able to identify authoritative document as claimed by the author. PHITS gives better results as compared to original HITS algorithm. Other difference between PHITS and standard HITS is that PHITS can estimate the probabilities of authorities compared to standard HITS algorithm, which can provide only the scalar magnitude of authority.

Weighted Page Rank Algorithm

Weighted Page Rank Algorithm is proposed by Wenpu Xing and Ali Ghorbani. Weighted page rank algorithm (WPR) is the modification of the original page rank algorithm. WPR decides the rank score 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 provides high value of rank to the more popular pages and does not equally divide the rank of a page among its out-link pages. Every out-link page is given a rank value based on its popularity. Popularity of a page is decided by observing its number of in links and out links.

Weighted Links Rank Algorithm

A modification of the standard page rank algorithm is given by Ricardo Baeza-Yates and Emilio Davis named as weighted links rank (WLRank). This algorithm provides weight value to the link based on three parameters i.e. length of the anchor text, tag in which the link is contained and relative position in the page. Simulation results show that the results of the search engine are improved using weighted links. The length of anchor text seems to be the best attributes in this algorithm. Relative position, which reveal that physical position does not always in synchronism with logical position is not so result oriented. Future work in this algorithm includes, tuning of the weight factor of every term for further evolution.

EigenRumor Algorithm

As the number of blogging sites is increasing day by day, there is a challenge for service provider to provide good blogs to the users. Page rank and HITS are very promising in providing the rank value to the blogs but some limitations arise, if these two algorithms are applied directly to the blogs The rank scores of blog entries as decided by the page rank algorithm is often very low so it cannot allow blog entries to be provided by rank score according to their importance. To resolve these limitations, a EigenRumor algorithm is proposed for ranking the blogs. This algorithm provides a rank score to every blog by weighting the scores of the hub and authority of the bloggers depending on the calculation of eigen vector.

Time Rank Algorithm

An algorithm named as TimeRank, for improving the rank score by using the visit time of the web page is proposed by H Jiang et al. Authors have measured the visit time of the page after applying original and improved methods of web page rank algorithm to know about the degree of importance to the users. This algorithm utilizes the time factor to increase the accuracy of the web page ranking. Due to the methodology used in this algorithm, it can be assumed to be a combination of content and link structure. The results of this algorithm are very satisfactory and in agreement with the applied theory for developing the algorithm.

Search Engine Optimization

SEO stands for “search engine optimization.” It is the process of getting traffic from the “organic,” “editorial” or “natural” listings on search engines. All major search engines such as Google, Yahoo and Bing have such results, where web pages and other content such as videos or local listings are shown and ranked based on what the search engine considers most relevant to users.

Ways to improve your Web pages for SEO

- **URLs**

Create clean, focused and optimized URLs. Include target keywords, avoid long or automated dynamic URLs.

- **<title> tag**

Optimal length for search engines is roughly 70 characters, place important keywords towards the front of the title tag, include your brand name at the end of a title tag, and make sure to consider readability and emotional impact.

- **<meta>**

Optimal length for search engines is roughly 155 characters, make it compelling and unique to each page.

- **<h1>**

Make sure your primary keyword or phrase is enclosed in an H1 tag, make sure it’s a strong headline that reads naturally, place it near the top of your page after your opening tag.

- **Keyword ratio**

This is the percentage of times a keyword is used in a body of text, the usual amount if 2-3%.

- **SiteMap**

Make sure to include a detailed and accurate XML sitemap in order to show search engine spiders and your visitors where to find information on your site.

SEO Black hat techniques:

- Keyword stuffing

Also known as “flooding,” this is when you try to cram as many keywords into your content that it becomes unreadable and lowers the quality of your site.

- Link farm

Considered a form of spamming, a link farm is any group of web sites that all hyperlink to every site in the group.

- Article spinning

Article spinning is when the owner of a site posts a unique article but is actually an exiting article that has been re-written.

- PageRank spoofing

In the past the PageRank was easily manipulated by a HTTP 302 response or “refresh” meta tag.

SEO Tools:

- Raven Tools - All-in-one Internet marketing platform for SEO, Social Media, PPC and Content. Research, manage, monitor and report on every aspect of your campaign. www.raventools.com/SEO-Software
- RIO SEO - SEO platform to deliver global search success across Organic, Local Search, Mobile and Social Media for Top Brands & Agencies. www.rioseo.com
- Google Analytics - lets you measure your advertising ROI as well as track your Flash, video, and social networking sites and applications. www.google.com/analytics
- SEO analytics software - SEOmoz tools helps you handle everyday SEO tasks. You can analyze keywords, research backlinks, do on-page analysis, find accessibility issues and track rankings all in one easy-to-use management platform. www.seomoz.org/tools
- Traffic Travis - For all your SEO & PPC Management needs. Use Traffic Travis for both on and off page analysis as well as spying on your competitors. www.traffictravis.com
- SheerSEO - SEO software. Complete automation of SEO including tools for tracking, back links building and much more.

CONCLUSION

Based on the algorithm used, the ranking algorithm provides a definite rank to resultant web pages. A typical search engine should use web page ranking techniques based on the specific needs of the users. After going through 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 techniques have limitations particularly in terms of time response, accuracy of results, importance of the results and relevancy of results. An efficient web page ranking algorithm should meet out these challenges efficiently with compatibility with global standards of web technology.

REFERENCES

I. Sergey Brin and Larry Page, “The anatomy of a Large-scale Hypertextual Web Search Engine”, In Proceedings of the Seventh International World Wide Web Conference, 1998.

II. Hyperlink based search algorithms-PageRank and HITS By Shatakirti

III. Search Engine Optimization Starter Guide By Google.

IV. A Comparative Analysis of Web Page Ranking Algorithms by Dilip Kumar Sharma and A. K. Sharma.