An Algorithm For Ranking The Web Pages Of Search Engine

N. N. Das
CSE/IT Dept., ITM University
Gurgaon, India

Shivani Gupta
CSE/IT Dept., ITM University
Gurgaon, India

Abstract

Search engine generally returns a large number of web pages in response to user queries using algorithms. Mostly search engines use pagerank algorithm with different concepts for sorting the list of documents. The document having the large number of visitors should be at the top. This paper discusses the concept of pagerank algorithm based on the number of visits by the user on a dataset. This paper also discusses the pagerank algorithm and the steps on how a web search engine works and displayed the relevant documents on the screen.

Keywords- *Search Engine, PageRank, Web Crawler, World Wide Web, Internet, Working of Search Engine, Computer Networks*

1. INTRODUCTION

One of the greatest things about the Internet is that everyone can use it and owns it. It is a collection of networks, both big and small which can be shared worldwide. These networks connect in many different ways to form the single entity that we know as the Internet. The Internet carries an extensive range of information resources and services, like as the linked hypertext documents of the World Wide Web (WWW) and the infrastructure to support email.

The terms Internet and World Wide Web are often used in everyday speech without much distinction, means, and these terms can be used vice versa. However, the Internet and the World Wide Web are not the same. The Internet can be a global system of interconnected computer networks. In contrast, the Web is one of the applications that run on the Internet through web browser. It is a collection of text documents and other resources, which are linked through hyperlinks and URLs, usually accessed by web browsers from web servers. In short, the Web can be thought of as an application or services "running" on the Internet providing various information to the end user.

2. WEB SEARCH ENGINE

A Web Search Engine or Internet Search Engine is software code that is designed to search for information on the World Wide Web [8]. There are differences in the ways various search engines work, but they all perform three basic functions:-

1. The search engines search the documents for the specified keyword or phrases of keywords.
2. They keep the index of the words they find and from where the words found.
3. They allow users to look for words of combinations of words found in that index.

Every search engine uses different complex mathematical formulas to generate search results. The results for a specific query are based on the different algorithms which search engine uses and then

displayed them on the SERP. There are different search engine algorithms which take the key elements of a web page, including the title of the page, its content and frequency of keywords, and come up with a ranking for where to place the results on the pages. Each search engine's algorithm is unique, means every search engine uses different algorithms for its ranking. That's why a top ranking on Yahoo! does not guarantee the same results for the ranking on Google, and vice versa. To make things more complicated, the algorithms used by search engines are not only constantly used, they are also constantly undergoing modification and revision, which provides updated algorithms.

Search engines are an extremely powerful way of promoting your own website online. Consider those websites your silent Public Relations, which are quietly working in the background. Many case studies have shown that between 40% and 80% of users found what they were looking for by using the search engine feature of the Internet.The great thing about search engines is that they bring targeted traffic to the website. The people are already motivated to make a purchase from you- because they searched you out. With the right website optimization, the search engines can always deliver thewebsite to your audiences using its optimization techniques.

Crawler-based search engines use automated software programs to survey and categorize the web pages. The programs or techniques used by the search engines to access your web pages are called 'spiders', 'crawlers', 'robots' or 'bots'. A spider will find a web page, download it and processes the information presented on the web page. This is a seamless processwhich provides information. The web page processed by spider will then be added to the database of search engine. When a user performs a search, the search engine will check its corresponding database of web pages for the key words the user searched on to present a list of link results. The results (list of suggested links which you requests), are listed on pages by order of which is 'closest' (as defined by the 'bots'), to what the user wants to find. Crawler-based search engines are constantly searching the Internet for new websites and updating their database of information with these new or altered web pages. Examples of crawler based search engines are: Google (www.google.com), Ask (www.ask.com).

3. SEARCH ENGINE OPTIMIZATION

Search engine optimization (SEO) is the process of affecting the visibility of a website [1]. It is the process of improving the organic ranking of a website with leading search engines. Submit Express, a professional SEO firm helps to improve search engine rankings for their clients by modifying their websites to better reflect what search engines are looking for. There are two types of results found on search engine result pages - links that are there thanks to organic SEO and links that are there due to paid search.

Organic SEO describes search engine optimization efforts that do not involve sponsored listings or paid campaigns of any sort [6]. Organic search engine optimization is the process of improving the volume or quality of traffic to a web site from search engines by means of natural search engine optimization efforts, specifically including the optimization of on-page content.

Paid search refers to sponsored search results- ads displayed typically on SERP because advertisers have pre-paid campaigns with search engines that helpin promoting their services or websites. These sponsored advertisements often appear above or alongside the organic search results. Programs like as Google AdWords allow advertisers to bid on keywords they want to match up with their advertisement. When a user enters one of those keywords in a search query, the advertiser's link and brief description of that may appear on the following SERP. If a user clicks on the link, the advertiser should pay a certain amount of money associated with the PPC campaign.

There can be some search engine optimization techniques which can be practiced on website: White SEO, Black Hat SEO, and Gray Hat SEO.

White hat SEO is the technique through which the rules of a search engine are followed, and ensure that the article which we post is our own and in accordance with the title, and this will increase our ranking in search engines, visitors will check it and will return again. White hat seo is like creating content for the search, not for the search engines by making content easily accessible to the bots.

Black hat SEO is a technique which is the opposite of white hat SEO, which is trying to improve website ranking in search engines in ways that are not allowed to search engines, such as bluffing or breaking the scatter machine. Some black hat seo techniques are: keywords searching, hiding text and links, doorway and cloaked pages, livestock link, make Comment Spamming on other people's websites.

Grey hat SEO (not white and not black but gray) is a technique that does not fully use the second technique above mentioned, or maybe combine the two of the techniques. Perhaps this is also the transformation from white to black or black to white. It could be argued, this technique is a technique that uses black hat techniques to achieve results.

Here are various SEO tools through which everyday SEO tasks can be easily handled. We can analyze keywords, research outlinks, do on-page analysis, find accessibility issues and track rankings all in one easy-to-use management platform. Some SEO tools used are campaign manager, followerwonk, link analysis, on-page analysis, rank tracking, crawl test, SEO toolbar, keyword analysis, mozscape API, google alerts, SEO for Firefox, URL Info, Seo browser, Copyscape etc.

4. WORKING OF SEARCH ENGINE

When a user enters a query, the list of web results or documents relevant to that query or keywords is displayed. Users normally tend to visit websites that are at the top of this list as they perceive those to be more relevant to the query. If you have ever wondered why some of these websites rank much better than the others then you must know that it is because of a powerful web marketing technique called Search Engine Optimization (SEO). Several activities involved in SEO in order to deliver search results – crawling, indexing, processing, calculating, relevancy, retrieving [7].

First, search engines crawl the Web to see what is there in the database. This task is performed by a piece of software, called a crawler or a spider (or Googlebot, Yahoo SLURP). Spiders follow links

from one page to another and index everything they find in searching. Having in mind the number of pages on the Web, it is impossible for a crawler to visit a site daily just to see if a new page has appeared or if an existing page has been updated, sometimes crawlers may not end up visiting your site for a month or two.

After a page is crawled, the next step is to index its content. The indexed page is stored in a huge database, from where it can later be retrieved and used for the next step. Essentially, the process of indexing is identifying the words and expressions that best describe the page and assigning the page to particular keyword or phrases of keywords. It will not be possible to process such amounts of information but generally search engines can deal with this work. Sometimes they might not get the meaning of a page right but if you help them with optimization, it will be easier for them to classify your pages correctly and for you – to get higher rankings [7].

When a request comes the search engine processes it, means search engine compares the search string with the indexed pages in database. When the search string got the relevant result from the database, the result is returned to the search engine. There are hundreds of web pages in the database so the relevancy of each page in its index is calculated with the search string in the query.There are various algorithms for calculating the relevancy. Each of these algorithms has different relative weights for common factors like frequency of keyword, links, or Meta tags. That is why different search engines give different search results for the same search string. What is more, it is a very well-known fact that all major search engines, like Yahoo!, Google, Bing, etc. periodically updatetheir algorithms and if you want to be at the top, you also need to adapt your pages to the latest changes.

The last step for the search engine to display the result on the screen is retrieving. It only means to display the resultant web pages in the database to the search string, in the browser. This was the major difference between search engines because of which they show the different results for the same query or we can say, for the same phrases of keywords. With this we can say, Google is the most promising search engine, then Yahoo, Bing, etc.

As in Figure1, user enters the query in the search engine then it goes to the web server. The web server sends the query to the index servers. The content inside the index server is similar to the index in the back of a book – it tells which pages contain the words that match any particular query term. Then the query travels to the document server, which actually

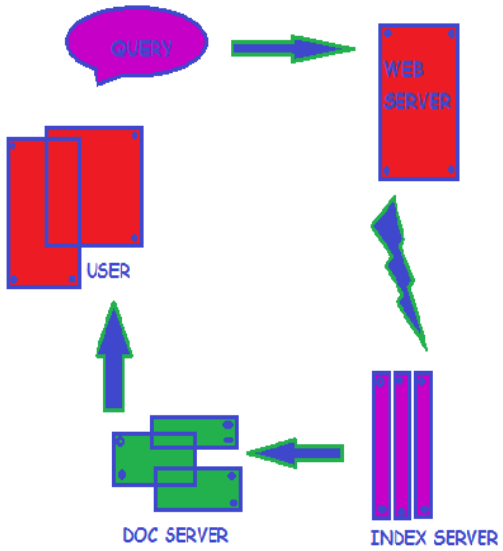

Figure 1 Working

retrieve the stored documents. Snippets are generated to describe each search result. Then the search results are returned to the user in a fraction of second.

5. PAGERANK ALGORITHM

The original Page Rank Algorithm was designed by Lawrence Page and Sergey Brin [5]. The Page Rank is based on the link analysis in which the pages are ranked on the basis of number of outgoing and incoming links [2]. The Page Rank is entirely based on the link structure of World Wide Web. Google uses the page rank algorithm, means Google searches the pages based on their number of outgoing and incoming links. Because of the page rank algorithm Google is such an effective search engine.

The limiting probability that an infinite number of random surfer visits any page is its pagerank. A page has high rank if the other pages with high rank linked to it. It is given by:-

PR (A) = (1-d) + d (PR (T_i)/C (T_i) + ... + PR (T_n)/C (T_n))

Where

- PR (A) is the Pagerank of page A,
- PR (T_i) is the Pagerank of pages T_i which link to page A,
- C (T_i) is the number of outbound links on page T_i and
- d is a damping factor which can be set between 0 and 1 usually 0.85.

The Pagerank ranks the page individually, not the whole website. The pagerank of page A is determined by the pages linked to page A. The pagerank of pages T_i which link to page A is determined by number of outbound links C (T) to page T.

The weight of pagerank T_i is added up and outcome is that if an additional inbound link is added to page A then pagerank of page A is increased. Finally, the outcome of weighted pagerank of all the pages is multiplied by damping factor, d (which lies between 0 and 1).

The simplified version of pagerank is:-

$$PR(u) = c \sum_{v \epsilon B(u)} \frac{PR(v)}{N(v)}$$

Where

- u represents a web page
- B(u) represents the set of web pages pointing to u
- PR (u) represents rank score of page u
- PR (v) represents rank score of page v
- N(v) represents the number of outgoing links link to page v
- C is the factor for normalization.

The values assigned to outgoing links of a page are in turn used to calculate the ranks of pages to which that page points.

All users do not follow the direct links on web, so the modified version of page rank is:-

$$PR(u) = (1 - d) + d \sum_{v \in B(u)} \frac{PR(v)}{N(v)}$$

Where d is dampening factor that is usually set to 0.85, d can be the probability of users following the direct links and (1-d) can be page rank distribution following non-directly links.

6. RESULTS

The pagerank algorithm is used by the search engine to rank the websites in the prescribed manner [1]. Google uses pagerank algorithm, but in what order it uses this algorithm is a secret. There can be many

ideas or we can say concepts through which web pages can be ordered. Like as, if there are a list of websites related to the software organizations, then those websites can be ordered by the paid amount, means the software organization which have been paid the most got the highest rank in search engine optimization and so on. I have created a dataset for the colleges and the list of colleges are ordered based on the number of visits to the college link by the user. Let's see how it works:

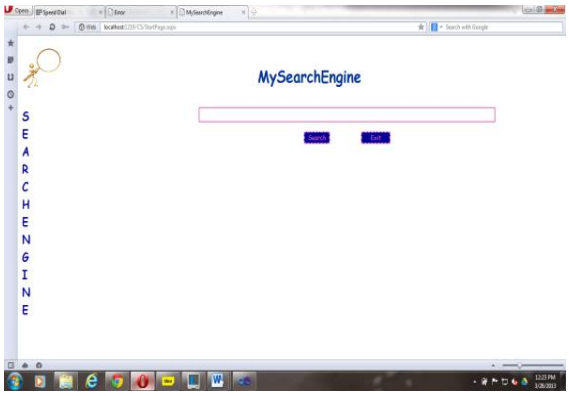

Figure 2 Structure of search engine

Figure 2 shows the basic structure of a search engine containing two buttons and textbox. One button is for search and the other is for exiting the browser. When the keyword is entered in the textbox then the relevant list of documents is displayed. As in, the keyword 'colleges' is entered the list of colleges is displayed on the screen shown in Figure 3.

Figure 3 List of web pages

When the link for the college is clicked it has shown the details for that corresponding college as shown in Figure 4.

Figure 4 Detail for the link

When the user clicks for a single web page many times then the counter for that web page is got increased and it comes on the top. It means the web page got a higher rank than other web pages on click event. As shown in Figure 5 the second web page becomes first on click because it got higher rank from the first and other web pages.

Figure 5 Sorted lists of web pages

In this way the pagerank algorithm can be implemented and a search can be designed based on this concept. This shows that there are many ways for the implementation of pagerank algorithm.

7. CONCLUSION

This paper explains the technique of search engine optimization which improves the ranking of web pages in the search engine. Then it explains the basic working of search engine which every search engine performs with different six activities such as, crawling, indexing, processing, calculating, relevancy, retrieving. Through these activities the user retrieves the list of web pages according to his query which is entered in the search engine.

The pagerank algorithm is also explained in the paper and how does it implement with the different concepts. The concept of ranking of the web pages is based on the number of visitors. How does the pagerank algorithm rank the websites for a search engine is all mentioned in the paper?

8. FUTURE WORK

The pagerank algorithm can also be implemented using the concept of recent visit to the web page by the user. It means the web page which is last visit by the user can be on the top, second last visit on the second top and so on. There can be many different ways for ranking the web pages and implementing the pagerank algorithm.

9. REFERENCES

[1] Gyanendra Kumar, Neelam Duhan, A. K. Sharma. *Page Ranking Based on Number of Visits of Links of Web Page* (2011).

[2] Neelam Duhan, A. K. Sharma, Komal Kumar Bhatia. *Page Ranking Algorithms: A Survey* (2009).

[3] Kai-Hsiang Yang and Chi-Chien Pan and Tzao-Lin Lee. *Approximate Search Engine Optimization for Directory Service*.

[4] Nursel YalçÕna, Utku Köseb*What is search engine optimization: SEO? (*2010).

[5] Brin, Sergey and Page Lawrence. *The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems*, April 1998.

[6] Search Engine Optimization at: http://www.submitexpress.com/search-engine-optimization.html.

[7] What Is SEO at: http://www.webconfs.com/seo-tutorial/introduction-to-seo.php?

[8] Web search engine at: http://en.wikipedia.org/wiki/Web_search_engine.

[9] Nripendra Das. *A Comparative Analysis of Link Oriented Algorithms of Web Mining on a Dataset* (2012).

[10] How does SEO work at: http://www.optimum7.com/internet-marketing/new-articles-content/how-does-seo-work.html