PageRank Algorithm

by Homer Simpson

Introduction

In this research paper, I am going to explain the purpose and ideology behind the Page Rank Algorithm, specifically the Google PageRank Algorithm. Many people don't understand the idea behind the ranking of webpages within the biggest search engine in the world. With the right information about the ranking system, it becomes easier to build successful websites and sell products on the World Wide Web, simply because they will be noticed both easier and earlier than other competitors.

The PageRank Algorithm is used to give webpages a level of importance. It perceives this importance by calculating how many links go to it from other webpages. It then passes a percentage of this importance to the page which it links to.

Knowing how the Google PageRank algorithm works, one can spend more quality time to improve it. Improving the page rank, improves the positioning of your webpage in the most used search engine in the world.

The algorithm itself isn't very complex once you understand the basic concept behind it. I hope to explain the core concept as best as I can, and how to manipulate it to ones' advantage.

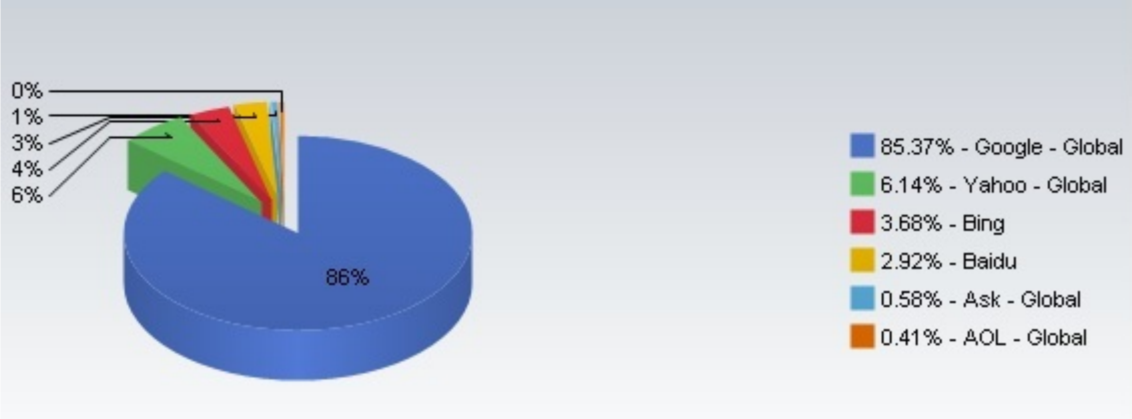

Fig 1.1 – Search Engine Popularity [1]

Background

The PageRank algorithm was created by Larry Page and Sergey Brin at Stanford University, California. The idea is based on "link popularity", as in the more links going to an individual page, the more important it is deemed to be. The higher the link popularity, the more Google deems the page as important. [2]

PageRank algorithm is the core of the Google search engine, and without it, would not exist, considering it was the original concept behind the leading search engine. Although there have been minor changes made to the algorithm, the basic concept of the formula remains the same and is still used by Google to this day. [2]

PageRank is simply a numeric value of importance that is tied to any given page on the web. The higher the importance, the higher it is going to be on the search results. Although the PageRank algorithm isn't the only formula used to rank pages, it is the most important. [6]

Webpages are given a rank by examining the link structure in which they create. Links are divided into two different types – Inlinks and Outlinks. Outlinks are links going out from a page to another. Inlinks are links going from a page and landing on your page. Each link from page to page is seen as a vote. A vote contains a variable percentage of its own pages importance that is calculated by counting how many outlinks it has and dividing it by its own importance. When page 'A' links to page 'B', 'A' casts a vote to 'B'. The vote that is casted to 'B' consists of an importance variable and is added to that linked page's importance. [6]

Generally speaking, the more inline links that you have, the more import your webpage is going to be. However, links from webpages that have a large amount of outlinks are not said to have the same importance as webpages that have few outlinks.

The algorithm ranks webpages without any human intervention in effect removing any biased rankings from the search results. Ranking updates are done periodically by Google 4 times a year. The reason for this is because in order for the algorithm to compute, it needs the most up to date link structure of the web to give the correct results.

Technique/Method/Algorithm

To begin, we'll discuss what an algorithm actually is and how it benefits us in Computer Science applications. An algorithm in Computer Science is a list of precise given instructions for computing a process on any given computer system. [8]

With this in mind, we can clearly understand that the PageRank algorithm is nothing but a set of instructions given to rank the pages on the World Wide Web as precise and accurate as possible so as to get the accurate results.

The algorithm tries to act like a normal site visitor when executed. Difference being, random links are clicked from consecutive webpages until finally moving to another page without having to click on a link. This is done when the algorithm lands on a page that has no more outlinks. Previous history of the said normal visitor has no induction as to what webpage will be visited next. This continues indefinitely. The PageRank score in this case represents the chances that a random website viewer clicks on your page.

The pages that are visited are plotted on a directed graph and are used to represent the link structure of the internet. The visited pages are represented as nodes and the pages they link to are represented by the different edges. [2]

When the directed graph is formed, the results are added to an adjacency matrix. The size of the matrix is determined by the number of pages that exist in the directed graph. The elements are added to the matrix in relation to the probability that a random user will land on a webpage ‘A’ if coming from another webpage ‘C’.

A ‘0’ is used to denote a webpage that has nil probability of a random website surfer going from that node to any other node in the directed graph. In the above situation, a user would have to either enter a different web address if they wanted to continue browsing the internet or stop. These nodes are known as dangling nodes. The action that Google takes when it comes to handling dangling nodes is undisclosed from the public population, so the exact reaction the algorithm takes with them is unknown. [2]

The full formula used by Google to compute PageRank is not publicly announced, but a basic concept is publicly available. Breaking it down to its most simple state we get the following:

$$PR(A) = (1-d) / N + d (PR(T1)/C(T1) + \ldots + PR(Tn)/C(Tn)$$

Where

The PageRank of A is denoted by PR(A)

The PageRank of y pages that link to page A is denoted by PR(Ti)

C(Ti) represents the total number of outbound links that exist on Ti.

As seen from above, Page A's PageRank is affected negatively by the number of outlinks C(t) on page A. This means that the number of outbound links on a given page plays a negative effect on the PageRank of that given page so having less outbound links plays a more positive affect. [3]

Each inlink has a weight assigned to it that is a percentage of its own webpage's weight. This weight is then added to the weight of your webpage, increasing the PageRank as a result.

In the formula, the damping factor is denoted by 'D', which is the probability that a random webpage surfer would land on your webpage. [3] The D variable can be valued at anything between 1 and 0, i.e., is almost always of a decimal value. The greater the value of D, the greater the chances are a random surfer will continue to click links on the given webpage. As seen above, the damping factor is implemented as 1-d, which means that a page is given a minimum possible PageRank, regardless of how many inbound links a page has. [3]

By looking at the formula, we can see that the sum of all pages' PageRank Ti is multiplied by 'D'. Thus, the overall benefit of a page linking to it by another page is reduced.

Websites can be manipulated to allow for linking to other webpages, without having any negative effect on PageRank. This is done by editing the HTML file in which the link is added to the page and adding a "no follow" attribute to the link. This attribute was introduced in an effort to combat Spamdexing, which is simply a method used to manipulate search engines by using keywords and phrases that have no relationship with the webpages content in an effort to drive more traffic to a particular website. The algorithm recognises these tags, and does not "follow" through on the links in which they relate to.

Strength/Weaknesses/Limitations

A major weakness within the algorithm is the fact that all the data on the internet must be precise in order to give an accurate result on PageRank of websites around the world. Inlinks and outlinks have to be counted up, necessary calculations made on them and then processed. Websites like MSN news or popular website forums have new links being added to them almost every half an hour. If the algorithm was to try and be 100% accurate, it would mean having to crawl these websites every hour in order to count how many links exist. It is almost impossible to have an exact precise amount of Inlinks and outlinks a popular website has at any given time, because new ones are always being added. For this reason, page rank updates are given periodically by Google. Google scan the web in whole once before PageRank is calculated for all websites which in effect slows down the process for both the website surfer and the webmaster who may have quality content to offer their visitors.

Another disadvantage is that people believe that the PageRank algorithm is the only thing that is important when it comes to website design and success. This is not the case. The PageRank Algorithm was not created to remove the other necessary components of a website such as meta-tags and, even high quality content. Webmasters often focus too heavily on PageRank, which in turn gives them low quality websites which hold little or no use to visitors, in turn, generating small profits. For this reason, on October 15, 2009, Google removed PageRank statistics from their popular Google Webmaster Tools application because they felt people focused on it too much. [5]

The fact that webmasters value PageRank so much on the internet, webmasters tend to post links to their websites as much as possible on other websites. This causes a lot of spam comments on the internet and becomes a hassle to moderators on different websites across the internetwork who have to delete the like. It can also cause interested viewers on websites to lose interest because of the excessive spamming that goes on.

The “no follow” attribute is a very good strength. Knowing that lots of outlinks have a negative effect on PageRank scores, webmasters would be careful as to who and where they linked to from outside of the domain. Using the “no follow” attribute, webmasters can link to where they want, knowing that there won’t be any negative effects on their own websites PageRank.

Conclusion

In conclusion, we now know that when a page has only few outlinks and many inlinks, the webpage is going to have a positive advantage given to it by the PageRank algorithm. The damping factor is the probability of a random webpage surfer landing on your page and is influenced by the number of inlinks that you have.

The PageRank algorithm is by far the most superior ranking system available on any search engine on the internet. We know this because Google is used by over 85% (see Fig1.1) of the internet users to this day, which means that users are finding the information they require with the assistance of the PageRank algorithm.

Although the PageRank algorithm has the power to rank you high in the search engine results, it should not be the only thing a webmaster focuses on, as high quality content is the key to successful websites.

The formula gives the basic concept idea behind the ranking process. But there have been alterations to it that Google have not publicly announced.

Larry Page and Sergey Brin have done an excellent job at producing this algorithm. Without it, finding relevant information on the internet would be of great difficulty, considering that other search engines were not giving results of the same high standard at the time. They truly have made history and shaped the internet into the place it is today.

Sources

[1] Fig1.1, NetMarketShare: http://www.netmarketshare.com/search-engine-market-share.aspx?qprid=4

[2] Rebecca S. Wills, Google's PageRank: The Math behind the Search Engine Algorithm, 2006 (Journal) http://www.math.cornell.edu/~web2940/Wills.pdf

[3] Székely Endre, Google and the Page Rank Algorithm, 2007, found at http://cs.ubbcluj.ro/~csatol/mach_learn/bemutato/SzekelyEndre_PageRank.pdf

[4] Sergey Brin, Lawrence Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine, (Journal) http://infolab.stanford.edu/pub/papers/google.pdf

[5] Public announcement from Google Employee in relation to removing PageRank from Webmaster Tools on Oct 15th 2009: http://www.google.com/support/forum/p/Webmasters/thread?tid=6a1d6250e26e9e48&hl=en

[6] Kurt Bryan and Tanya Leise, THE $25,000,000,000 EIGENVECTOR THE LINEAR ALGEBRA BEHIND GOOGLE (Journal), http://www.rose-hulman.edu/~bryan/googleFinalVersionFixed.pdf

[7] Tanguy Urvoy, Emmanuel Chauveau and Pascal Filoche, Tracking Web Spam with HTML Style Similarities, 2008, (Journal) http://oniros.org/papers/urvoy2008acmtweb.pdf

[8] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest and Clifford Stein, Introduction to Algorithms Third Edition, 2009