The PageRank Algorithm

John David Stroud

CS-3330 Data Structures and Algorithms

Troy University

The PageRank Algorithm

1. Introduction

Prior to the development of the PageRank algorithm, a search on the Internet could best be described as looking for a needle in a haystack. There was no structure and one could spend hours combing through irrelevant information to find what they were looking for. By today's standards, this is an unacceptable practice. Larry Page and Sergey Brin developed the PageRank algorithm to act as an index that allows a novice user to find a topic in its most relevant form to a user in less than half a second. The PageRank algorithm was designed to seek out the most relevant topic, thus saving the end user from having to sift through a plethora of irrelevant web pages. "In short, PageRank's thesis is that a webpage is important if it is pointed to by another important page" (Langville, Meyer, 2006).

The onset of this enlightenment of information ushered out the Industrial Age and made way for the Information age. However, there were significant barriers to overcome and our society become ubiquitous with access to information. None of the information was easy to access or possessed any level of structure. This void was filled when two Stanford students introduced the world to the PageRank algorithm and set the stage to make Google one of the most dominant companies on the planet.

The goal of this paper is to present the problem of searching for information on the World Wide Web and demonstrate the solutions that the PageRank algorithm solved that forever changed the world of search engines and how we retrieve information. The paper will identify how the algorithm works and the primary purpose behind its design. We will also present the mathematics behind the algorithm, as well as identify its strengths and weaknesses. PageRank draws much of its influence from Markov and linear algebra; the complexities behind the

mathematics will be presented and discussed. In addition to the PageRank algorithm itself, modifications of the PageRank algorithm are presented. Kleinberg's Hyperlink-Induced Topic Search (HITS) method is covered briefly as it was developed near the same time frame and was committed to achieving the same purpose as the PageRank algorithm.

2. Origin of PageRank

Before PageRank content was buckling under the pressure of the massive size of the web and the death grip of spammers. The Web needed some organization, an index of some sort. The PageRank algorithm was built within the idea that a well-designed algorithm can solve the problem of sorting Web pages to allow a user to find the information that is most relevant to their particular search. In the 1990's there was no real structure or order to searching for information on the Internet; therefore it became a frustrating endeavor to find information that you were looking for. Often, you had to rely on word or mouth or suggestions from friends and colleagues who had labored over searches for hours to find what you were truly looking for.

The answer to this problem came in the form of two Stanford graduate students who developed an algorithm that would change the way we search for information and build one of the most dominant companies in the world. In 1998 the most successful search engines started using link analysis for information retrieval. "Link analysis is a technique that exploits the additional information inherent in the hyperlink structure of the web" (Langville, Mayer, 2006).

3. General Description of the PageRank Algorithm

PageRank is a method for computing a ranking of every Web page based on the graph of the Internet. The algorithm is an attempt to see how good an approximation of how "importance" can be obtained just from the link structure. Generally, highly linked pages are more "important" than pages with few links. A page has a high rank if the sum of the ranks of its backlinks is high (Page, Brin, Motwani, Winograd, 1998). The most important thing that a search engine can do is to determine how to rank the pages and put the most relevant page at the front of the list.

PageRank is a way of measuring the importance of web sites. "A numerical weighting is assigned to each element of a hyperlinked set of documents with the purpose of measuring its relative importance within the set. A page becomes more important is important pages link to it" (Roberts, 2014, p. 809). In other words, a your web page is more important if the New York Times is linked to it instead of just your close friends and family. Each page on the Web is an endorsement. The more pages you have, the more endorsements you have. What is a distinguishing feature is to have important pages link to your Web page. Your Web page will rank higher, in terms of importance if you have the New York Times linked to your Web page versus sites that have lower traffic.

Another way to describe the PageRank algorithm is that the final ranking of each page represents the probability of reaching that page by following links on the Web at random. Processes that proceed by making random choices without regard to previous decisions are called Markov processes after the Russian mathematician Andrei Markov (1856-1922) who was among the first to analyze their mathematical properties (Roberts, 2014, p.809).

4. Mathematical Definition of the PageRank Algorithm

The PageRank algorithm begins with a simple summation question. The PageRank of a page P_i, denoted rP_i, is the sum of the PageRanks of all pages pointing into P_i.

$$rP_i = c \sum_{P_j \in Bp_i} \frac{r(P_j)}{|P_j|}$$

Figure 1.1

Where Bp_i is the set of pages pointing into P_i and $|P_j|$ is the number of outlinks from P_j. Notice that the PageRank of inlinking pages $r(P_j)$ is tempered by the number of recommendations made by P_j, denoted $|P_j|$. A problem with the equation above is the PageRanks of pages inlinking to pages P_i, are unknown. To sidestep this problem, Brin and Page used an iterative procedure (Langville, Meyer, 2006, p. 89).

$$r_{k+1}P_i = \sum_{P_j \in Bp_i} \frac{r_k(P_j)}{|P_j|}$$

Figure 1.2

This process is initiated with $r_0\,(P_i) = 1/n$ for all pages P_i and repeated with the hope that the PageRank scores will eventually converge to some final stable values (Langville, Meyer, 2006, p. 89).

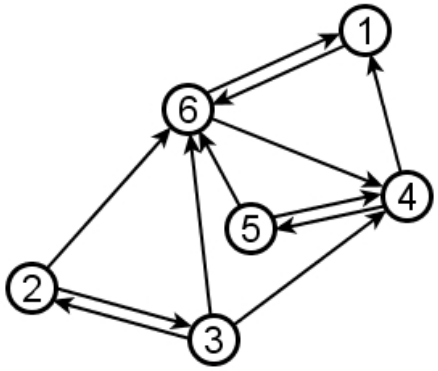

Figure 1.3 Directed graph representing web of six pages

Iteration 0	Iteration 1	Iteration 2	Rank at Iter. 2
$r_0(P_1) = 1/6$	$r_1(P_1) = 1/18$	$r_2(P_1) = 1/36$	5
$r_0(P_2) = 1/6$	$r_1(P_2) = 5/36$	$r_2(P_2) = 1/18$	4
$r_0(P_3) = 1/6$	$r_1(P_3) = 1/12$	$r_2(P_3) = 1/36$	5
$r_0(P_4) = 1/6$	$r_1(P_4) = 1/4$	$r_2(P_4) = 17/72$	1
$r_0(P_5) = 1/6$	$r_1(P_5) = 5/36$	$r_2(P_5) = 11/72$	3
$r_0(P_6) = 1/6$	$r_1(P_6) = 1/6$	$r_2(P_6) = 14/72$	2

Figure 1.4 First few iterations using Figure 1.2

5. Primary Uses

The PageRank algorithm's primary use comes from anyone who engages in a search using Google's search engine. While the user is typically unaware that they are using a sophisticated algorithm to find their information; they are nevertheless, allowing the PageRank to operate with its intended function of ranking pages of by their level importance. Hyperlinks from a users homepage to another page is my endorsement of that page. Thus, a page with more recommendations (which are realized through inlinks) must be more important than a page with a few inlinks. However, similar to other recommendation systems such as bibliographic citations or letters of reference, the status of the recommender is also important.

Another way to describe the effect of the PageRank algorithm is that the final ranking of each page represent the probability of reaching that page by following links on the Web at random. Processes that proceed by making random choices without regard to previous decisions are called Markov processes after the Russian mathematician Andrei Markov (1856 - 1922) who was among the first to analyze their mathematical probabilities (Roberts, 2014, p. 809).

A similar algorithm to the PageRank algorithm is HITS, designed by Kleinberg. If we designed the Web as a graph, the Web's hyperlink structure forms a massive directed graph. The nodes in the graph represent web pages and the links represent hyperlinks. Kleinberg took that position that the Internet could be viewed as an intricate form of populist hypermedia, in which millions of on-line participants, with diverse and often conflicting goals, are continuously creating hyperlinked content (Kleinberg, 1999). In Kleinberg's view this type of structure manifested itself into a global organization that is utterly unplanned. "Kleinberg's HITS method for ranking pages is very similar to PageRank, but it uses both inlinks and outlinks to create two popularity scores for each page" (Langville, Meyer, 2006).

6. Future Considerations

From a business perspective it is safe to assume that any entrepreneurs would like to develop a better algorithm and employ their algorithm in a manner that would give Google some serious competition. To date, no such company has been able to keep with Google's business model. One could make a strong argument that the success of PageRank was due to being one of the first algorithms of its kind or because the company behind the algorithm had a better business plan.

Matthew Richardson and Pedro Domigos from The University of Washington's Computer Science Department came up with an algorithm they deemed 'The Intelligent Surfer.' which is guided by a probabilistic model of the relevance of a page to a query. The efficient execution of their algorithm at query time is made possible by pre-computing at crawl time the necessary terms (Richardson & Domingos, 2001).

Time has given us the opportunity to critique page rank and ask ourselves if we can to build a superior algorithm with a similar and consistent result. Inefficiencies in the PageRank

algorithm have been noted and there is no short of research pointing to potential superior solutions. A problem common to both PageRank and HITS is topic drift. Because they give the same weight to all edges, the pages with the most inlinks in the network being considered (either at crawl or query time) tend to dominate, whether or not they are the most relevant to the query (Richardson, Domingos, 2001).

Richardson and Domingos propose a more intelligent surfer, who probabilistically hops from page to page, depending on the content of the pages and the query terms the surfer is looking for. In other words, when choosing among multiple out-links from a page, the directed surfer tends to follow those which lead to pages whose content has been deemed relevant to the query (Richardson & Domingos, 2001).

7. Summary

It is clear that the traffic that Google handles on their website per day that we can call the PageRank algorithm one of the most innovative solutions in history. The founders of the algorithm were faced with the problem of developing a search engine that would be more useful for all users of the Internet to conduct research and learn more efficiently. It is also important to remember that similar algorithms were being developed at the same time of PageRank's origin, however, it was the invention of Mr. Page and Mr. Brin that have enjoyed world wide acceptance and usage. Over time we have seen variations and improvement of the original form of the algorithm. However, we must stand in awe of the original form of the algorithm and ask ourselves what would the world look like today without Google?

References

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998, January 29). The PageRank Citation Raking: Bringing Order to the Web. *SpringerReference,* 1-3.

Roberts, E. S. (2014). Chapter 18.7/ Algorithms for searching the web. In *Programming Abstractions in C* (pp. 809-812). Pearson. Eric S Roberts, Stanford University

Langville, A. N., & Meyer, C. D. (2006). *Google's PageRank and beyond: The science of search engine rankings*. Princeton, NJ: Princeton University Press.

Kleinberg, J. A. (1999). *Authoritative Sources in a Hyperlinked Environment* [Scholarly project]. Department of Computer Science, Cornell University

Richardson, M., & Domingos, P. (2001). *The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank* [Scholarly project].
Department of Computer Science and Engineering, University of Washington