COMPUTATION OF RANKING OF WEB PAGES

Gaurav Jindal*

Nidhi Kapoor**

Shanky Kalra**

Abstract:

In this paper, we present Google, a prototype of a large-scale search engine which makes heavy use of the structure present in hypertext. Google is designed to crawl and index the Web efficiently and produce much more satisfying search results than existing systems.This paper serves as a companion or extension to the "Inside Page Rank", Everyday we Google out something to found out the desire content, and upto this time, nearly half of the world just do Google on the Internet. Well in this paper we are concerned with the mechanism behind the working of PageRank Algorithm, i.e how the ranking of pages is done, how the PageRank is computed, how the in-links and out-links affects the ranking of pages. We introduce a few new results, provide an extensive reference list, and speculate about exciting areas of future research.

Keywords: Damping Factor(d), Page Rank(PR),PR(Tn),PR(Cn).

* Assistant Professor, Gitarattan International Business School, New Delhi, India

** Student, Gitarattan International Business School, New Delhi, India

I. INTRODUCTION

The year 1998 was a busy year for link analysis models. On the East Coast, a young scientist named Jon Kleinberg, an assistant professor in his second year at Cornell University, was working on a Web search engine project called HITS. His algorithm used the hyperlink structure of the Web to improve search engine results, an innovative idea at the time, as most search engines used only textual content to return relevant documents. He presented his work begun a year earlier at IBM, in January 1998 at the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms held in San Francisco, California.

Very nearby, at Stanford University, two Ph.D. candidates were working late nights on a similar project called PageRank.[1]Sergey Brin and Larry Page, both computer science students, had been collaborating on their Web search engine since 1995. By 1998, things were really starting to accelerate for these two scientists. They were using their dorm rooms as offices for the fledgling business, which later became the giant Google. By August 1998, both Brin and Page took a leave of absence from Stanford in order to focus on their growing business. In a public presentation at the Seventh International World Wide Web conference (WWW98) in Brisbane, Australia, their paper "The PageRank citation ranking: Bringing order to the Web" made small ripples in the information science community that quickly turned into waves. The connections between the two models are striking and it's hard to say whether HITS influenced PageRank, or vice versa, or whether both developed independently. Nevertheless, since that eventful year, PageRank has emerged as the dominant link analysis model, partly due to its query independence, its virtual immunity to spamming and Google's huge business success. Kleinberg was already making a name for himself as an innovative academic, and unlike Brin and Page, did not try

To develop HITS into a company. However, later entrepreneurs did; the search engine Teoma uses an extension of the HITS algorithm as the basis of its underlying technology. As a side note, Google kept Brin and Page busy and wealthy enough to remain on leave from Stanford. This paper picks up after their well-cited original 1998 paper and explores the numerous suggestions that have been made to the basic PageRank model, thus, taking the reader deeper inside PageRank. We note that this paper describes methods invented by Brin and Page, which were later implemented into their search engine

Google. Of course, it is impossible to surmise the details of Google's implementation since the publicly disseminated details of the 1998 papers Nevertheless, we do know that PageRank remains the heart of Google software and continues to provide the basis for all of web search tools as cited directly from the Google webpage.

Prior to link analysis,[2] search engines relied on traditional methods for ranking webpages. These traditional methods used only page content to retrieve pages relevant to a search query. While content is important and is still analyzed by search engines, adding link analysis was a major improvement in the quality of the ranked results returned to the user. Spamming makes this clear. Search engines that use only content to return results to users are easily spammed. Spammers, who often have financial and business interests tied up in their webpages, intentionally try to deceive search engines into giving them an unjustly high rank so that their pages appear toward the top of the results list. Since spammers have complete control over the content of their pages, they can imbed meta-tags and keywords into the HTML that search engines may use to determine the content of a page. However, the text that appears on the page may be quite different. When link analysis is employed these spammers are less effective because spammers have little or no control over which pages inlink to theirs. [4] PageRank uses the hyperlink structure of the web to view inlinks into a page as a recommendation of that page from the author of the inlinking page. Since inlinks from good pages should carry more weight than the inlinks from marginal pages each webpage is assigned an appropriate rank score,which measures the importance of the page.

Figure 1

[1] Google's PageRank algorithm, though mathematically elegant and very powerful, is very easy to understand.

The first step to understanding PageRank is to view the Web as a giant graph.

II. WORKING OF ALGORITHM

What is PageRank?

a). In short PageRank is a vote [4] by all the other pages on the Web, about how important a page is.

b). A link to a Page counts as a vote of support.

PR (A) = (1-d) + d(PR(T1)/C(T1) +…..+PR(Tn)/C(Tn)) (1)

Breaking down the equation.

a). PR(Tn)- Each Page has a notion of its own self importance. That's "PR(T1)" for the first web page in the web all the way up to "PR(Tn)" for the last page.

b). C(Tn)- Each Page spreads out its vote [5]evenly amongst all of it's outgoing links.The count, or number of it's outgoing links for page 1 is "C(T1)", "C(Tn)" for page n, and so on for all pages.

c). PR(Tn)/C(Tn) - so if our page (page A) has a backlink from page "n" the share of the vote page A will get is "PR(Tn)/C(Tn)".

d). d(… - All these fractions of votes are added together but, to stop the other pages having too much influence, this total vote is "damped down" by multiplying it by 0.85 (the factor "d") .

e). (1 - d) - The (1 – d) bit at the beginning is a bit of probability math magic so the "sum of all web pages\' PageRank will be one": it adds in the bit lost by the d(…. It also means that if a page has no links to it (no backlinks) even then it will still get a small PR of 0.15 (i.e. 1 – 0.85). (Aside: the Google paper says "the sum of all pages" but they mean the "the normalised sum" – otherwise known as "the average" to you and me).

III. SIMPLE EXAMPLE

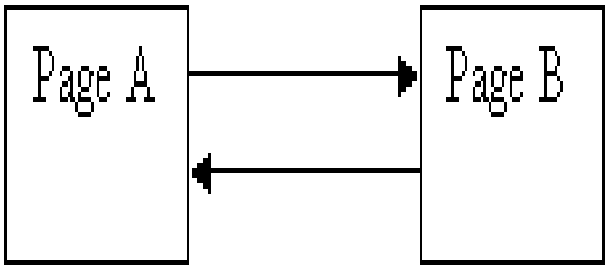

Figure 2

Each page is having one link towards to each other.

Like in this example Page A has one link towards Page B is casting one vote for one another.

So that means C(A) = 1 and C(B) = 1.

IV. EXPLANATION

[4] We don't know what their PR should be to begin with, so we will just guess 1 as a safe random number.

a). d (Damping factor)=0.85

b). PR (A)= (1 – d) + d(PR(B)/1).

c). PR(A)= 0.15 + 0.85 * 1= 1

d). PR(B)= 0.15 + 0.85 * 1= 1.

Lets do it Again with another number.

Let's try 0 and re-calculate.

a). PR(A)= 0.15 + 0.85 * 0= 0.15.

b). PR(B)=0.15+0.85*0.15=0.2775

Now we have calculated a "next best guess" so we just plug it in the equation again.

a). PR(A)= 0.15 + 0.85 * 0.2775= 0.385875

b). PR(B)= 0.15 + 0.85 * 0.385875= 0.47799375

c). PR(A)= 0.15 + 0.85 * 0.47799375= 0.5562946875

d). PR(B)= 0.15 + 0.85 * 0.5562946875=0.622845788

V. PRINCIPLE

It doesn't matter where you start your guess, once the PageRank calculations have settled down, the "normalized probability distribution" (the average PageRank for all pages) will be 1.0.

VI. APPLICATIONS

1) A version of PageRank has recently been proposed as a replacement for the traditional Institute for Scientific Information (ISI) impact factor, and implemented at eigenfactor.org. Instead of merely counting total citation to a journal, the "importance" of each citation is determined in a PageRank fashion.

2). [5] A similar new use of PageRank is to rank academic doctoral programs based on their records of placing their graduates in faculty positions. In PageRank terms, academic departments link to each other by hiring their faculty from each other (and from themselves).

3). PageRank has been used to rank spaces or streets to predict how many people (pedestrians or vehicles) come to the individual spaces or streets. In lexical semantics it has been used to perform Word Sense Disambiguation and to automatically rank Word Net Synsets according to how strongly they possess a given semantic property, such as positivity or negativity.

4). A dynamic weighting method similar to PageRank has been used to generate customized reading lists based on the link structure of Wikipedia.

VII. CONCLUSION

As we have covered the Computation of Ranking of Web Page in the Analysis of Page Rank, in this context we can say that the ranking of pages plays an important role in placing them at the topmost place of the search.

It's all about the rank of page which is done by other pages, and at the initial every page has a rank of 0.15 by the formula of (1-d) present in the algorithm.

VIII. REFERENCES

[1] Amy N. Langvillea & Carl D. Meyerb,Deeper "Inside PageRank", October 20, 2004,Internet Mathematics Vol 1, No 3,335-380.

[2] http://www.whydomath.org/node/google/math.html

[3] http://www.markhorrell.com/seo/pagerank.html

[4] Sergey Brin , Lawrence Page, "The Anatomy of a large-scale Hypertextual Web search engine", Computer Networks and ISDN Systems,Volume 30, Issues 1–7, pp 107–117 ,April 1998, [Proceedings of the Seventh International World Wide Web Conference].

[5] P. Desikan, N. Pathak, J. Srivastava and V. Kumar "Incremental PageRank Computation on evolving graphs",pp 1094 – 1095, WWW '05 [Special interest tracks and posters of the 14th international conference on World Wide Web].

[6] L. Page, S. Brin, R. Motwani, T. Winograd "The Page Rankcitation ranking bringing order to the web" Stanford Digital Libraries Working Paper, 1998.