Title

Investigating the correlation between the ranking of web pages by Eigenfactor Score and Google PageRank

Introduction

Following the advancement in technology, the internet has become a staple in gaining knowledge. A large database of information is available online, ranging from numerous sources such as scientific journals to web articles, giving many the access to information from various fields and subjects. This increases the accessibility to information, allowing many to better their knowledge. While undergoing my course of study, the internet has always been a viable option for my learning, often allowing me to access web pages providing useful explanations that answer my queries and supplementary information to increase my mastery and knowledge of subjects. While searching for information online through the search engine Google, I noticed that the top websites are often sufficient and directly address my queries. Because my searches are processed in a highly efficient manner, this piqued my curiosity regarding the mathematical algorithm behind the method in which Google ranks web pages to direct information specifically tailored to web searches, improving the accessibility to the information we require. Because scientific journals are often referred to during science-based subjects, I chanced upon an existing webpage ranking system made for scholarly journals. This ranking system was based on the bibliometric data, such as the ingoing and outgoing citations of a journal. As such, I was curious as to how this system of citation ranking is applicable to web searches, in comparison with Google's PageRank algorithm, and also its possible applications to web searches.

Mathematical Background

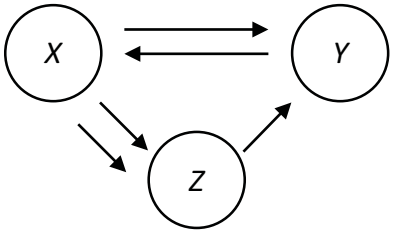
In this investigation, a set of top-ranking journals will be investigated. These journals will be analyzed for the web page data, mainly the outgoing hyperlinks, on their web pages. The citations from these journals will then be used determine the Eigenfactor Score and a bibliometric ranking of the web pages. The web page data collected will then be used to calculate the PageRank value, where a PageRank ranking will be generated from this algorithm's values. Both ranking sequences will then be compared, presenting a relationship between both methods of analysis.

Graph Theory and Matrices:

Both algorithms are formed on the basis of graph theory, the study of properties and application of graphs. A graph is defined as the relationship between different objects. A visual representation consists of a series of vertices (circles) that are connected by edges (arrows), where the edges represent the

relationship between the vertices.[1] These vertices can represent individual events or objects, while the edges reflect how each event is linked and the number of relationships each event or object has with another. This is displayed in Sample Graph G, where *X*, *Y* and *Z* are the vertices and the arrows represent the edges that reflect how each vertice is linked to each other.

Sample Graph *G*



$$\begin{array}{c} \\ X \\ Y \\ Z \end{array} \begin{array}{c} \begin{array}{ccc} X & Y & Z \end{array} \\ \begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \end{array}$$

Where the XY^{th} entry of the adjacency matrix is the number of links from the X^{th} journal to Y^{th} journal

Graph *G* can be presented alternatively in the form of an adjacency matrix, seen in the figure on the right. In this case, an adjacency matrix is extracted by counting number of edges between two adjacent vertices that share at least one common edge.[2] A better understanding of a matrix can be seen in the analysis of the matrix. The first column displays the number of outgoing hyperlinks from *X*, to *Y* and *Z*, while the second and third columns represent the number of outgoing hyperlinks from *Y* and *Z*, to the other vertices. Thus, since there is no edge leading to *X* itself, *X* does not have a relationship with itself. This is similar for *Y* and *Z*, resulting in a central diagonal column that is equal to 0. This suggests that there are no relationships between each repeated vertex.

In this investigation, the web pages are considered as the vertices and the edges between these vertices will be the outgoing hyperlinks on these pages, forming a graph reflecting the relationship between web pages. This resultant graph is directed, having specific directions for each edge, resulting in a non-symmetric adjacency matrix. These matrices created are used in the following algorithms.

Google's PageRank Algorithm:

PageRank is an algorithm often used in search engine optimization, providing an analysis of the relationship between web pages. It is simplified to be a voting system determining the relevance of web pages through their ingoing hyperlinks. Each hyperlink directed to a page represents a vote supporting the page's importance.[3] If a page has no ingoing hyperlinks, votes will not be deducted, but rather the result will be considered as zero votes. The basis behind this algorithm can be seen in graph theory,

[1] "Graph Theory Tutorial: Adjacency Matrix," 2016, accessed January 8, 2017, http://people.revoledu.com/kardi/tutorial/GraphTheory/Adjacency-Matrix.html.
[2] "Adjacency Matrix," May 26, 2007, accessed January 8, 2017, http://mathworld.wolfram.com/AdjacencyMatrix.html.
[3] "Pagerank Explained Correctly with Examples," accessed January 8, 2017, http://www.cs.princeton.edu/~chazelle/courses/BIB/pagerank.htm.

where each web page is a node and the ingoing hyperlinks of a web page are the directed edges between the nodes, representing the relationship and interaction between web pages through the ingoing hyperlinks.[4] Google's simplified PageRank algorithm is as follows:

$$Pr(x) = (1 - d) + d\left[\frac{Pr(W_1)}{C(W_1)} + \cdots + \frac{Pr(Wn)}{C(Wn)}\right]$$

Pr = PageRank function of a web page

Pr(x) = PageRank value for *x*

d = damping factor, generally assumed to be approximately 0.85

Pr(Wn) = PageRank value for webpage *n*

C(Wn) = Outbound links of webpage *n*

This algorithm consists of a recurring function, *Pr(x)*, which inputs values produced from its output values. This is used to produce a probability distribution reflecting the chances that one may end up on a web page when accessing a hyperlink. To increase the accuracy of results, a PageRank calculation undergoes several iterations, applying the formula to each set of data repetitively, until a certain value is achieved. This modifies the calculated values to fluctuate less from their true theoretical values, where there is a convergence of the iteration values. However, a true theoretical value is mathematically unattainable and is set to a certain difference between each iteration, at a small convergence value (ϵ), which is considered to be negligible. A damping factor is applied to consider the chances that a web surfer would eventually stop surfing. This value has been widely tested by many researches and is often assumed to be 0.85.[5] An application of the PageRank algorithm is displayed in the example below.

Use of PageRank algorithm:

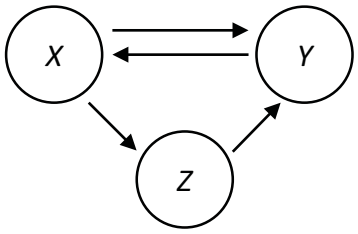


Diagram displaying set of web pages

The previous diagram represents a small web, consisting three web pages *X*, *Y* and *Z*. In this set, it is apparent that *X* has 1 outgoing hyperlink to *Y* and 2 to *Z*, while *Y* has 1 outgoing hyperlink to *X*, and *Z* has 1 outgoing hyperlink to *Y*.

[4] "Google PageRank - Algorithm," accessed January 8, 2017, http://pr.efactory.de/e-pagerank-algorithm.shtml.

[5] Meghabghab, G., & Kandel, A. (2008). Search engines, link analysis, and user's web behavior. Berlin, Heidelberg: Springer.