RANKING INTERNET SEARCH RESULTS WITH LINEAR ALGEBRA

by Greg Cox, Julian Pagtama, Alonso Santana and David Talley

INTRODUCTION

About 9,740,000 results (0.11 seconds)

Linear algebra - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Linear_algebra**
Linear algebra is the branch of mathematics charged with investigating the properties of finite as well as countably infinite dimensional vector spaces and linear ...
↳ List of linear algebra topics - Basis - Rank - Projection

Why does Wikipedia's "Linear algebra" page show up first when you search for "Linear algebra" in Google? Google uses a proprietary ranking algorithm based on standard linear algebra that assigns an importance score to each page on the world wide web based on the idea of "backlinks," or how many other pages link to a particular page. By converting the 8 billion pages of the world wide web into a giant system of linear equations, and applying the fundamental principles of eigenvectors, Google can determine which results are the most important for a given search query. Google calls this their "PageRank" algorithm, but they're not the only ones to use it.

METHOD

The method for determining the importance score for a given web of pages is as follows:

1 Each page gets exactly 1 "vote," which is further reduced by the number of outgoing links the page has. For example, if Page A has 10 outgoing links, each outgoing link counts as 0.1 votes, which is multiplied by the page's calculated rank. As a result, each page's calculated rank is the sum of the weighted vote value of every page that links to it, creating a system of linear equations.

2 Create the link matrix A based off of the system of linear equations, and convert it into a column-stochastic matrix M via the following conversion formula:

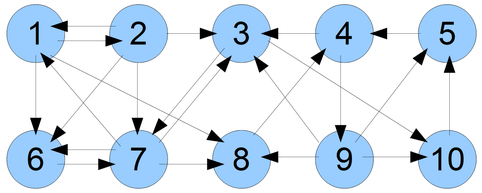
$$M = (1-m)A + mS$$

Where A is the original matrix, and S is an identically sized nxn matrix whose values are all 1/n. This formula removes all disjoint components of the link graph. The variable *m* is an arbitrary value between 0 and 1 that shifts M between A and S. We use 0.15.

3 Find the eigenvector for the new matrix, and the coordinates within that eigenvector are the importance scores for each of the pages in the result set. The accepted method for finding the eigenvector for an arbitrarily sized matrix, especially a large one like the ones Google processes on a regular basis, is by using the Power Method, an iterative process that starts with a rough approximation of the target eigenvector and iteratively refines the approximation until it approaches an accurate value. The Power Method is beyond the scope of this poster.

RESULTS

We start with an initial link graph, which represents the pages in a closed web:



From there, we calculate the system of linear equations:

$X_1 = X_2/4 + X_7/4$
$X_2 = X_1/3$
$X_3 = X_2/4 + X_4/2 + X_7/4 + X_9/4$
$X_4 = X_5/1 + X_8/1$
$X_5 = X_9/4 + X_{10}/1$
$X_6 = X_1/3 + X_2/4 + X_7/4$
$X_7 = X_2/4 + X_3/2 + X_6/1$
$X_8 = X_1/3 + X_7/4 + X_9/4$
$X_9 = X_4/2$
$X_{10} = X_3/2 + X_9/4$

Which gives us the matrix to the right:

$$A=\begin{bmatrix}
0 & 1/4 & 0 & 0 & 0 & 0 & 1/4 & 0 & 0 & 0 \\
1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1/4 & 0 & 1/2 & 0 & 0 & 1/4 & 0 & 1/4 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/4 & 1 \\
1/3 & 1/4 & 0 & 0 & 0 & 0 & 1/4 & 0 & 0 & 0 \\
0 & 1/4 & 1/2 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
1/3 & 0 & 0 & 0 & 0 & 0 & 1/4 & 0 & 1/4 & 0 \\
0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 1/4 & 0
\end{bmatrix}$$

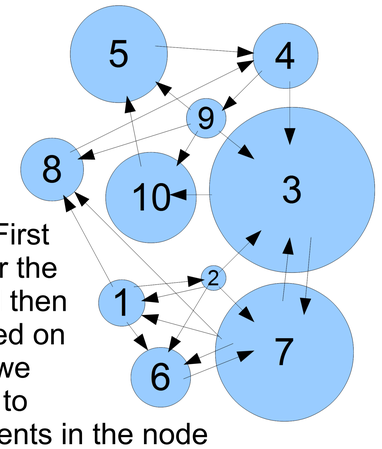
We then run A through our formula to the left under step 2, using 0.15 for *m* and 1/10 for all of the values of S, which produces the following matrix M:

$$M=\begin{bmatrix}
3/200 & 91/400 & 3/200 & 3/200 & 3/200 & 3/200 & 91/400 & 3/200 & 3/200 & 3/200 \\
179/600 & 3/200 & 3/200 & 3/200 & 3/200 & 3/200 & 3/200 & 3/200 & 3/200 & 3/200 \\
3/200 & 91/400 & 3/200 & 11/25 & 173/200 & 3/200 & 91/400 & 3/200 & 91/400 & 3/200 \\
3/200 & 3/200 & 3/200 & 3/200 & 3/200 & 3/200 & 3/200 & 173/200 & 3/200 & 3/200 \\
3/200 & 3/200 & 3/200 & 3/200 & 3/200 & 3/200 & 3/200 & 3/200 & 91/400 & 173/200 \\
179/600 & 91/400 & 3/200 & 3/200 & 3/200 & 3/200 & 91/400 & 3/200 & 3/200 & 3/200 \\
3/200 & 91/400 & 11/25 & 3/200 & 3/200 & 173/200 & 3/200 & 3/200 & 3/200 & 3/200 \\
179/600 & 3/200 & 3/200 & 3/200 & 3/200 & 3/200 & 91/400 & 3/200 & 91/400 & 3/200 \\
3/200 & 3/200 & 3/200 & 11/25 & 3/200 & 3/200 & 3/200 & 3/200 & 3/200 & 3/200 \\
3/200 & 3/200 & 11/25 & 3/200 & 3/200 & 3/200 & 3/200 & 3/200 & 91/400 & 3/200
\end{bmatrix}$$

M is now what is called a "column-stochastic" matrix, which means that all of the elements in each of its columns sum up to 1. M also now has no disjoint components, meaning every page contained in M now has at least one backlink, and at least one outgoing link. From here, we can apply the Power Method to find what is called the dominant eigenvector, or the eigenvector whose eigenvalue is furthest from 0, and the values in that eigenvector represent the importance scores of each page in the web above. The target eigenvector is to the right.

X1	0.164
X2	0.088
X3	0.584
X4	0.230
X5	0.343
X6	0.211
X7	0.488
X8	0.222
X9	0.140
X10	0.320

SUMMARY



Using the method described previously, we were able to determine the individual importance scores of a group of linked nodes. First we created equations for the score of each node, and then we formed a matrix based on those equations. Then we performed a conversion to remove disjoint components in the node graph, and finally we found a suitable eigenvector using the Power Method that properly ranked each page according to its backlinks. The new graph above represents a new graph where each node is sized according to its importance score.

CONCLUSION

Linear algebra has countless uses in the real world, but Google's use of eigenvectors and the Power Method were a first in the world of internet search technology, hence the phrase "The $25,000,000,000 Eigenvector." The prominent use of linear algebra from Google's PageRank algorithm comes from the fact that each page's importance score also relies on the importance scores of the pages that link to it, and because one page's unknown importance score relies on another page's equally unknown importance score, the values can be represented as unknown variables, which are perfect for use in a system of linear equations. There are other page ranking algorithms out there, and some have nothing to do with linear algebra, but link analysis in general heavily depends on matrices, which makes it a perfect linear algebra problem.

ACKNOWLEDGEMENTS

"Page rank," by Wikipedia
http://en.wikipedia.org/wiki/Page_rank

"The $25,000,000,000 Eigenvector," by Kurt Bryan and Tanya Leise
http://www.rose-hulman.edu/~bryan/googleFinalVersionFixed.pdf

"Power Method for Approximating Eigenvalues," by Ron Larson
http://college.cengage.com/mathematics/larson/elementary_linear/4e/shared/downloads/c10s3.pdf

"The Use of the Linear Algebra by Web Search Engines," by Amy N. Langville and Carl D. Meyer
http://meyer.math.ncsu.edu/Meyer/PS_Files/IMAGE.pdf