http://ssvrrwq.hubpages.com/hub/Understanding-the-Mathematics-behind-Googles-PageRank-linear-algebra-model

July 30, 2012

Understanding the Mathematics behind Google's PageRank (linear algebra model)

Introduction

This is intended to be a very accessible introduction to the mathematics behind PageRank. No college level mathematical knowledge is assumed, but understanding linear algebra will allow you to gain a deeper understanding.

Since so many people spend their lives on SEO, I felt that it was vital to provide an accessible introduction to how PageRank operates.

Of course, Google is always developing their algorithms, but this can still be considered the basis for their ranking algorithm. Google's actual algorithm is more complex, and the product of thousands of man-years of intellectual effort by Math and CS PHDs.

Note: the personalization factor, and other vital steps are removed to make the introduction more accessible, please see the associated paper in the conclusion if you are mathematically inclined and would like to understand the full method.

Outline of the Technique

We will model the internet as a matrix (this is basically a table with certain mathematical properties), where the rows and columns are websites, and the entries count how many times the sites link to each other.

We then multiply the entries in this matrix by one minus the probability that a random user will pick a new page at random, instead of clicking a link (Google finds this by experiment, we'll assume its somewhere around .15). I.e. we multiply each entry in the matrix by .85.

We will then compute an eigenvector of this new matrix.

A vector is a list of numbers, like [1.0, 3.0, 5,0].

In this case, the eigenvector (almost, see the paper in the conclusion) represents the PageRank of the pages from our table. For example, the 3rd website from our table, will have a PageRank at the 3rd entry in the Eigenvector.

The website's PageRank is a representation of the probability that a random surfer clicking random links will find the web page. (A mathematical proof that this is the eigenvector, when appropriately normalized, is available in the paper listed final section of this article).

	Site 1	**Site 2**	**Site 3**	**Site 4**
Site 1	0	1/2	1/2	0
Site 2	1/3	0	1/3	1/3
Site 3	0	0	0	0
Site 4	1	0	0	0

Creating the Matrix

Imagine that we are creating a new search engine using the simplified PageRank algorithm. Our prototype webspider has found 4 websites. The first website links to itself site 2 and site 3. Site 2 links to site 1,3 and 4. Site 3 is a 'dead end' and Site 4 links to site 1. We represent this data in the table to the left.

Google perturbs the dead end sites slightly from zero to make computation easier.

Computing an Eigenvector with the power method

The Power Method is then used to compute the Eigenvector. The results of the last PageRank testing run is used as the first approximation. For new sites, something close to 0 is used as the first approximation. For the first run ever, a more computationally expensive method, or an excessive number of iterations had to be used.

Conclusion and Further Reading

Obviously, this is highly simplified. If you would like to know more, I have selected a number of relevant books below.

The book on Linear Algebra is an excellent text that covers all the relevant mathematics, and touches upon PageRank.

Google's PageRank and Beyond, is an excellent book, with great depth on the subject.

Matrix Methods in Data Mining, is an advanced text, with many applications to search engines.

Additionally, if you have a math background (or once you read the linear algebra text), you may be interested in the papers "PageRank Revisited" and "Google's PageRank: The Math Behind the Search Engine" published by the ACM (the latter paper was the basis for this article).