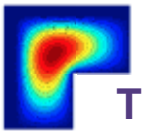


The World's Largest Matrix Computation

Google's PageRank is an eigenvector of a matrix of order 2.7 billion

by Cleve Moler

One of the reasons why Google is such an effective search engine is the PageRank™ algorithm, developed by Google's founders, Larry Page and Sergey Brin, when they were graduate students at Stanford University. PageRank is determined entirely by the link structure of the Web. It is recomputed about once a month and does not involve any of the actual content of Web pages or of any individual query. Then, for any particular query, Google finds the pages on the Web that match that query and lists those pages in the order of their PageRank.

Imagine surfing the Web, going from page to page by randomly choosing an outgoing link from one page to get to the next. This can lead to dead ends at pages with no outgoing links, or cycles around cliques of interconnected pages. So, a certain fraction of the time, simply choose a random page from anywhere on the Web. This theoretical random walk of the Web is a *Markov chain or Markov process*. The limiting probability that a dedicated random surfer visits any particular page is its PageRank. A page has high rank if it has links to and from other pages with high rank.

Let W be the set of Web pages that can reached by following a chain of hyperlinks starting from a page at Google and let n be the number of pages in W. The set W actually varies with time, but in May 2002, n was about 2.7 billion. Let G be the n-by-n connectivity matrix of W, that is, $g_{i,j}$ is 1 if there is a hyperlink from page i to page j and 0 otherwise. The matrix G is huge, but very sparse; its number of nonzeros is the total number of hyperlinks in the pages in W.

Let c_j and r_i be the column and row sums of G.

$$c_j = \sum_i g_{i,j}, \quad r_i = \sum_j g_{i,j}$$

The quantities c_k and r_k are the *indegree* and *outdegree* of the k-th page. Let p be the fraction of time that the random walk follows a link. Google usually takes $p = 0.85$. Then $1-p$ is the fraction of time that an arbitrary page is chosen. Let A be the n-by-n matrix whose elements are

$$a_{i,j} = p\, g_{i,j} / c_j + \delta, \text{ where } \delta = (1-p) / n.$$

The matrix A is not sparse, but it is a rank one modification of a sparse matrix. Most of the elements of A are equal to the small constant δ. When $n = 2.7 \cdot 10^9$, $\delta = 5.5 \cdot 10^{-11}$.

The matrix is the transition probability matrix of the Markov chain. Its elements are all strictly between zero and one and its column sums are all equal to one. An important result in matrix theory, the Perron-Frobenius Theorem, applies to such matrices. It tells us that the largest eigenvalue of A is equal to one and that the corresponding eigenvector, which satisfies the

equation

$x = Ax,$

exists and is unique to within a scaling factor. When this scaling factor is chosen so that

$\sum_i x_i = 1$

then x is the state vector of the Markov chain. The elements of x are Google's PageRank.

If the matrix were small enough to fit in MATLAB, one way to compute the eigenvector x would be to start with a good approximate solution, such as the PageRanks from the previous month, and simply repeat the assignment statement

```
x = Ax
```

until successive vectors agree to within specified tolerance. This is known as the power method and is about the only possible approach for very large n. I'm not sure how Google actually computes PageRank, but one step of the power method would require one pass over a database of Web pages, updating weighted reference counts generated by the hyperlinks between pages.

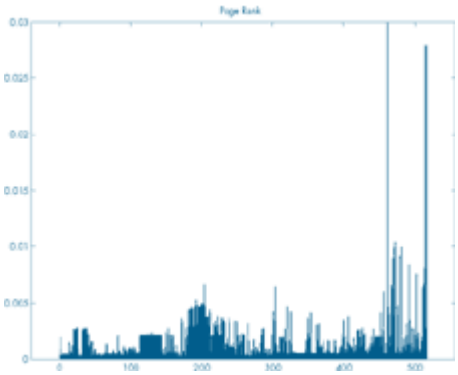
Links from the MathWorks home page to pages with high PageRank.

Our PageRank example uses a tiny subset of the Web consisting of n = 517 pages on the MathWorks Web site, starting at the home page www.mathworks.com, and following 13,531 links. The `spy` plot of the connectivity matrix G shows many cliques of interconnected pages. The portion of the `spy` plot involving columns 508:517 contains 4013 links and so is fairly dense. These columns represent the bar of buttons on the top of the home page with labels like `products` and `support` and the small copyright button on the bottom of the page that are duplicated on many other pages. The buttons point to pages like www.mathworks.com/company/copyright.shtml. Rows 114:144 of the `spy` plot all have the same structure. These rows represent links from articles off the `pressroom` page to other pressroom material.

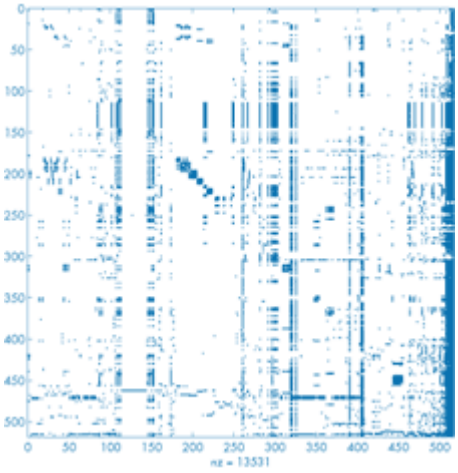
The PageRank calculation for this example produces a vector x with components that range between 0.0685 and 0.0003. Here are the top ten pages, together with their in and out

degrees. We see that the pressroom has the highest rank, even though it has fewer references than some of the other pages. MATLAB Central, the new file exchange and newsgroup access page, is ranked fifth, and *News & Notes* is ranked seventh.

	PageRank	in	out	URL
1	0.068488	57	83	www.mathworks.com/company/pressroom
2	0.027927	452	203	www.mathworks.com/siteindex.shtml
3	0.027923	456	46	www.mathworks.com/products
4	0.010445	1	136	www.mathworks.com/products/products_by_category.shtml
5	0.009961	16	34	www.mathworks.com/matlabcentral
6	0.009916	1	123	www.mathworks.com/products/product_list.shtml
7	0.009206	117	36	www.mathworks.com/company/newsletter
8	0.008997	2	118	www.mathworks.com/products/product_descriptions.shtml
9	0.008343	86	42	www.mathworks.com/company/digest
10	0.008083	423	64	www.mathworks.com/support



PageRank of the small Web sample. The top three pages are the pressroom, the site index and the products button. Click on image to see enlarged view.



Spy plot of the link structure of a small sample of the Web starting at www.mathworks.com. The dense columns on the right represent buttons that are common to many pages. Click on image to see enlarged view.

Additional information about PageRank and related topics is available on the Web.

- Google provides a brief explanation at www.google.com/technology/index.html
- A technical report by Page and Brin, together with R. Motwani and T. Winograd, is at dbpubs.stanford.edu:8090/pub/1999-66
- A paper presented by John Tomlin of the IBM Almaden Research Center, and colleagues A. Arasu, J. Novak and A. Tomkins at the 2002 World Wide Web conference is at www2002.org/CDROM/poster/173.pdf
- Jon Kleinberg, a computer science professor at Cornell University, has done a lot of work on the graph theoretic structure of the Web. His home page is www.cs.cornell.edu/home/kleinber

Next Article

related topics:

Using MathWorks Products For... | Training | MATLAB Based Books | Third-Party Products