(Original) Google PageRanks

Stephen Vardeman

Analytics Iowa LLC

ISU Statistics and IMSE

General objective

This might be thought of as of some general interest beyond the particular application of ranking Web pages, if one abstracts the notion of summarizing features of a directed graph with N nodes (N Web pages in the motivating application) where edges point from some nodes to other nodes (there are links on some Web pages to other Web pages). The basic idea is that one wishes to rank the nodes (Web pages) by some measure of importance.

Notation

If $i \neq j$ define

$$L_{ij} = \begin{cases} 1 & \text{if there is a directed edge pointing from node } j \text{ to node } i \\ 0 & \text{otherwise} \end{cases}$$

and define

$$c_j = \sum_{i=1}^{N} L_{ij} = \text{the number of directed edges pointed away from node } j$$

(There is the question of how we are going to define L_{jj}. We may either declare that there is an implicit edge pointed from each node j to itself and adopt the convention that $L_{jj} = 1$ or we may declare that all $L_{jj} = 0$.)

Defining (linear) equations

A node (Web page) might be more important if many other (particularly, important) nodes have edges (links) pointing to it. The Google™ PageRanks $r_i > 0$ are chosen to satisfy

$$r_i = (1-d) + d \sum_j \left(\frac{L_{ij}}{c_j} \right) r_j \qquad (1)$$

(in case a $c_j = 0$ above, we'll presume that $\frac{L_{ij}}{c_j}$ is taken to be 0) for some $d \in (0,1)$ (producing minimum pagerank $(1-d)$). (Apparently, a standard choice is $d = .85$.) These are, of course, simply N linear equations in the N unknowns r_i, and for small N one might ignore special structure and simply solve these numerically with a generic solver. In what follows we exploit some special structure.

Development of a solution

Without loss of generality, with

$$\mathbf{r} = \begin{pmatrix} r_1 \\ \vdots \\ r_N \end{pmatrix}$$

we'll assume that $\mathbf{r}'\mathbf{1} = N$ so that the average rank is 1. Then, for

$$d_j = \begin{cases} \dfrac{1}{c_j} & \text{if } c_j \neq 0 \\ 0 & \text{if } c_j = 0 \end{cases}$$

define the $N \times N$ diagonal matrix $\mathbf{D} = \mathbf{diag}\,(d_1, d_2, \ldots, d_N)$. (Clearly, if we use the $L_{jj} = 1$ convention, then $d_j = 1/c_j$ for all j.)

Development cont.

Then in matrix form, the N equations (1) are (for $\mathbf{L} = (L_{ij})_{\substack{i=1,\ldots,N \\ j=1,\ldots,N}}$)

$$\begin{aligned}\mathbf{r} &= (1-d)\,\mathbf{1} + d\mathbf{LDr} \\ &= \left(\frac{1}{N}(1-d)\,\mathbf{11}' + d\mathbf{LD}\right)\mathbf{r}\end{aligned}$$

(using the assumption that $\mathbf{r}'\mathbf{1} = N$). Let

$$\mathbf{T} = \left(\frac{1}{N}(1-d)\,\mathbf{11}' + d\mathbf{LD}\right)$$

so that $\mathbf{r}'\mathbf{T}' = \mathbf{r}'$.

Development cont.

Note all entries of $\mathbf{T}$ are non-negative and that

$$\begin{aligned}\mathbf{T}'\mathbf{1} &= \left(\frac{1}{N}(1-d)\,\mathbf{1}\mathbf{1}' + d\mathbf{L}\mathbf{D}\right)'\mathbf{1} \\ &= \frac{1}{N}(1-d)\,\mathbf{1}\mathbf{1}'\mathbf{1} + d\mathbf{D}\mathbf{L}'\mathbf{1} \\ &= (1-d)\,\mathbf{1} + d\mathbf{D}\begin{pmatrix} c_1 \\ \vdots \\ c_N \end{pmatrix}\end{aligned}$$

so that if all $c_j > 0$, $\mathbf{T}'\mathbf{1} = \mathbf{1}$. We have this condition as long as we either limit application to sets of nodes (Web pages) where each node has an outgoing edge (an outgoing link) or we decide to count every node as pointing to itself (every page as linking to itself) using the $L_{jj} = 1$ convention. Henceforth suppose that indeed all $c_j > 0$.

Markov chain stationary distribution

Under this assumption, $\mathbf{T}'$ is a stochastic matrix (with rows that are probability vectors), the transition matrix for an irreducible aperiodic finite state Markov Chain. Defining the probability vector

$$\mathbf{p} = \frac{1}{N}\mathbf{r}$$

it then follows that since $\mathbf{p}'\mathbf{T}' = \mathbf{p}'$ the PageRank vector is N times the stationary probability vector for the Markov Chain. This stationary probability vector can then be found as the limit of any row of

$$\left(\mathbf{T}'\right)^n$$

as $n \longrightarrow \infty$.