Lecture 9

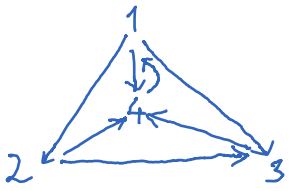
Wednesday, February 5, 2025 11:29 AM

In 1998, while being PhD students at Stanford University, Lawrence Page and Sergey Brin published a paper "The anatomy of a large-scale hypertextual Web search engine" on the journal of *Computer Networks and ISDN systems*. They discovered a profound algorithm to rank web pages. They called it PageRank algorithm, a name that attributes one of the coauthor and also refers to the web *pages*. This algorithm is the basis of the Google search engine.

The overall idea is that web pages are ranked based on their importance. The most important page is ranked first. Each web page will be given a weight between 0 and 1. The sum of the weights of all web pages is equally to 1. Some natural rules apply:

- Any link from a web page to itself does not count.
- Multiple links from web page A to web page B are only counted as one.

Note that the rank of a web page is not simply determined by the number of inbound links to it, but also by the rank of the pages that link to it. Consider the following internet:



Pages 4 has inbound links from all other pages, so it should be the most important page. In term of inbound links, page 1 and 2 each has 1 link, page 3 has 2 links, page 4 has 3 links. Based on these numbers, you may be haste to say that page 3 should be ranked above page 1. But the page that links to page 1 is page 4, which is the most important page. The two pages that link to page 3 are less important than page 4.

Brin and Page's solution (simple version): Imagine that you freely navigate from one website to another by clicking any link on the page that you are on. All outbound links have the same chance of being clicked on. *The importance of the web page is defined by the chance that you land on it*. Suppose there are n web pages, labelled from 1 to n. The internet above can be represented by a matrix:

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

The entry A_{ij} at row i and column j is the probability of going to page i from page j. Matrix A is called a *transition matrix*. You can notice that A is a stochastic matrix.

Denote by E_i the chance that you are on page i, and F_i the event that you click on a link that takes you to page i. The probability of landing on page i is

$$P(E_i) = P(E_1 \cap F_i) + P(E_2 \cap F_i) + \cdots + P(E_n \cap F_i) = P(F_i|E_1)P(E_1) + \cdots + P(F_i|E_1)P(E_1)$$

Let $p_i = P(E_i)$. Then $p_i = A_{i1}p_1 + A_{i2}p_2 + \cdots + A_{in}p_n$. In terms of matrix multiplication,

$$p = Ap, \text{ where } p = \begin{bmatrix} p_1 \\ p_2 \\ \dots \\ p_n \end{bmatrix}.$$

Therefore, p is a probability stationary vector of A. You can check that A^5 has all positive entries. Therefore, according to Perron-Frobenius theorem, A has a unique probability stationary vector. This vector is called the *PageRank vector* of the internet.

To find this vector, you solve the equation $(A - I_4)p = 0$. Associated matrix:

$$\begin{bmatrix} -1 & 0 & 0 & 1 & 0 \\ 1/3 & -1 & 0 & 0 & 0 \\ 1/3 & 1/2 & -1 & 0 & 0 \\ 1/3 & 1/2 & 1 & -1 & 0 \end{bmatrix} \xrightarrow{RREF} \begin{bmatrix} 1 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1/3 & 0 \\ 0 & 0 & 1 & -1/2 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

All the stationary vectors are $v = \begin{bmatrix} t \\ (1/3)t \\ (1/2)t \\ t \end{bmatrix}$. For this vector to be a probability vector, $t = 6/17$.

Therefore, the only probability stationary vector is $p = \begin{bmatrix} 6/17 \\ 2/17 \\ 3/17 \\ 6/17 \end{bmatrix}$. This is the PageRank vector of the internet. Consequently, page 1 and page 4 are ranked equally highest. Page 3 is ranked next. Page 2 is ranked lowest.

Brin and Page's solution (general version): there is a drawback with the simple version mentioned above. In the previous example, A, A^2, A^3, A^4 each has at least one zero entry. It is possible to have a situation where no powers A^k have all positive entries. This is indeed the case if the internet has disconnected clusters.

The consequence is that the probability stationary vector may not be unique. In such a case, it is ambiguous how to order the pages. Brin and Page's solution is to introduce a damping parameter $d \in [0,1]$. A typical choice of d is $d = 0.85$. The transition matrix is

$$A = dA' + (1-d)\frac{1}{n}\begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}$$

where A' is the regular (without damping) transition matrix. As long as $d < 1$, A will have all positive entries.

To see the rationale behind the damping parameter, imagine that you freely navigate from one website to another in the following manner: with probability d, you click on any link on the page that you are on (all outbound links still have the same chance of being clicked on). With probability $1 - d$, you randomly jump to any of the page on the internet. This is a simple and clever solution to "connect" all disconnected cluster.