PageRank

Nikola Samardzic
UCLA Undergraduate
nikola.s@ucla.edu

May 25, 2017

1 What is PageRank?

PageRank is defined in the following way:

Let N be the total number of pages within a network. Let $i \to j$ imply that there is a hyperlink on page i to page j, and let $i \not\to j$ imply that page i does not have a link to page j. We define a **relationship matrix** Q:

$$Q_{ji} = \begin{cases} 1/k, & i \to j \\ 0, & i \not\to j \end{cases}$$

Now observe that Q is the transitional matrix of the Markov Chain defined by pages and their mutual hyperlinks. Assuming that this Markov Chain is irreducible, there is a stable probability distribution

$$\begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_N \end{bmatrix}$$

In practice it has been determined that p_i is proportional to the amount of time a user spends on webpage i. Therefore, the bigger p_i is the more relevant the i^{th} webpage is considered.

Now, the matrix Q is slightly modified

$$P = \alpha \begin{bmatrix} Q_{11} & Q_{12} & \dots & Q_{1N} \\ Q_{21} & Q_{22} & \dots & Q_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ Q_{N1} & Q_{N2} & \dots & Q_{NN} \end{bmatrix} + \frac{1-\alpha}{N} \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix}$$

For $0 < \alpha < 1$ P is a transition matrix of an irreducible, aperiodic Markov Chain, which implies that its stable probability distribution unique. In practice, α is usually 0.85 or $1 - 1/N$. It is important to note that the choice of α might influence the final ranking of the pages.

We can think about the matrix P as an improved version of Q. It generously gives each page an initial ranking of $(1-\alpha)/N$. If page i has a ranking of p_i, and there is another page j for which $i \to j$, then the contribution of page i to the rank of page j is αp_i. The total ranking of a specific page could be found by solving a system of linear equations

$$\begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_N \end{bmatrix} = \alpha \begin{bmatrix} Q_{11} & Q_{12} & \dots & Q_{1N} \\ Q_{21} & Q_{22} & \dots & Q_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ Q_{N1} & Q_{N2} & \dots & Q_{NN} \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_N \end{bmatrix} + \frac{1-\alpha}{N} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

2 Solving the system

The most popular method for solving this system uses an iterative method that finds the biggest eigenvalue (and corresponding eigenvector).

Starting from an arbitrary vector x, the k^{th} iteration calculates $P^k x$. In order to approximate the stable probability distribution we introduce a normalizor: Let's look at the normalized value of this vector, where the norm is defined relative to the standard $\mathbb{R}^n$ vector dot product. We know that

$$\lim_{k\to\infty} \frac{A^{k+1}x}{||A^{(k)}x||} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{bmatrix},$$

i.e. the stable probability distribution on the Chain.

The most processor power is spent on the computation of the matrix and vector products. The number of iterations required depend on the ratio $|\lambda_2|/|\lambda_1|$. Havelivala and Kamvar have proven that the upper bound for $|\lambda_2|$ is α, which implies that the value of α is proportional to the speed of convergence. For example, for $\alpha = 0.85$ the iterative method converges in 50 iterations on a Markov Chain containing 80 million pages.