POLITECNICO DI TORINO

Dipartimento di Scienze Matematiche *G.L. Lagrange*

Exercise 6: Google's PageRank

PCS - Programmazione e Calcolo Scientifico

Prof. S. Berrone, G. Teora, F. Vicini

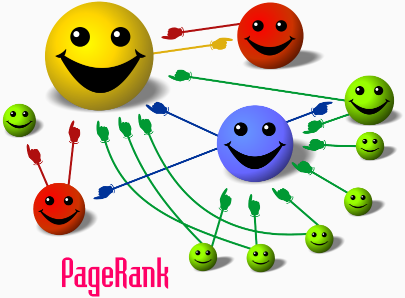
A.A. 2022/2023

Politecnico di Torino

Google's PageRank

Google's PageRank

The Google's PageRank is an algorithm used by Google to rank web pages in their search engine results.



https://it.wikipedia.org/wiki/PageRank

Web as a direct graph

A graph is a pair of sets $G = (V, E)$, where V is a set whose elements are called vertices, and E is a set of paired vertices $(v_i, v_j) \in V \times V$, whose elements are called edges or links. A direct graph is a graph in which edges have orientations, i.e. $(v_i, v_j) \neq (v_j, v_i)$.

Web

The Web can be represented as a direct graph whose n nodes are the Web pages, while its edges are the hyperlinks between pages.

A toy example:

Starting from `https://www.polito.it/`, we can build a direct graph with $n = 6$ nodes by following the hypertext references, until reaching the desired number of nodes.

1. `https://www.polito.it` points to 2, 3, 4;
2. `https://www.polito.it/themes/custom/polito/logo_meta.png` is a dangling node;
3. `https://www.polito.it/en` points to 1, 2
4. `https://www.polito.it/ateneo/qualita` points to 2, 5, 6
5. `https://www.coronavirus.polito.it/it` points to 1
6. `https://www.polito.it/impatto-sociale/masterplan` points to 2, 4

A toy example: The Adjacency Matrix

The Adjacency Matrix $A \in \mathbb{R}^{n \times n}$ of a direct graph is defined s.t.

$$A(i,j) = \begin{cases} 1 & \text{if } (v_i, v_j) \in E \\ 0 & \text{otherwise} \end{cases}$$

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

Google's PageRank

PageRank is a way of measuring the importance of website pages: more important websites are likely to receive more high-quality links from other websites.

The PageRank can be represented by a vector of probabilities $y \in \mathbb{R}^n$, whose entries are defined as

$$y[i] = \sum_{\substack{j=1,\ldots,n \\ (v_j, v_i) \in E}} \frac{y[j]}{\deg(j)},$$

$$y[i] \geq 0, \quad \forall i = 1, \ldots, n, \quad \|y\|_1 = 1,$$

where $\deg(i)$ represents the number of outcoming links from node i.

A toy example: The Adjacency Matrix

The Transition Matrix $\tilde{G} \in \mathbb{R}^{n \times n}$ of a direct graph is defined s.t.

$$\tilde{G}(i,j) = \begin{cases} \frac{1}{deg(i)} & \text{if } \deg(i) > 0 \text{ and } (v_i, v_j) \in E \\ 0 & \text{otherwise} \end{cases}$$

$$\tilde{G} = \begin{bmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \end{bmatrix}$$

A toy example

According to those definitions, the most important page is `https://www.polito.it/themes/custom/polito/logo_meta.png`.

The Google Matrix

The PageRank is represented by the left eigenvector related to the eigenvalue $\lambda = 1$ of the row-stochastic Google matrix:

$$G(i,j) = \begin{cases} \frac{1}{\deg(i)} & \text{if } \deg(i) > 0 \text{ and } (v_i, v_j) \in E \\ \frac{1}{n} & \text{otherwise} \end{cases} \tag{1}$$

Dangling nodes

If the random surfer arrives at a dangling node, it picks another URL at random and continues surfing again.

The Parametric Google Matrix

We define the Parametric Google Matrix as

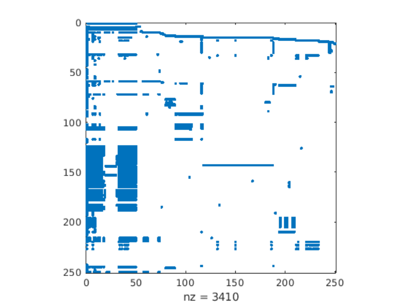
$$G(c) = cG + (1-c)ew^T, \quad c \in (0,1), \tag{2}$$

where c is a damping factor, $w \in \mathbb{R}^n$ is a probability vector called the *personalization vector* and $e \in \mathbb{R}^T$ is the all-ones vector.

The well-posedness of PageRank problem

The PageRank problem related to the Parametric Google Matrix is well-posed, i.e. $\exists! y \in \mathbb{R}^n: \ y^T G(c) = y^T$ s.t. $\|y\|_1 = 1$, $y \geq 0$. Furthermore, the eigenvalues of $G(c)$ are $\{1, c\lambda_2, \ldots, c\lambda_n\}$ and $|c\lambda_i| \leq c \ \forall i = 2, \ldots, n$, where $\{1, \lambda_2, \ldots, \lambda_n\}$ are the eigenvalues of G.

Sorting algorithm



The spy plot of the adjacency matrix of a portion of the web with 250 nodes built starting from `https://www.polito.it/`.

Reminder

Use very efficient sorting algorithms to rank web pages.