

SCOTTYRANK.JL: AN IMPLEMENTATION OF PAGERANK & HITS

SIYUAN CHEN AND MICHAEL ZHOU

E-mail addresses: siyuanc2@andrew.cmu.edu, mhzhou@andrew.cmu.edu.
Date: November 2021.

1. MATHEMATICAL BACKGROUND

1.1. Linear Algebra.

1.1.1. *Definitions.* Positive matrices are defined as matrices with positive entries.

Markov matrices are defined as square matrices with nonnegative entries and column sum 1 across all of its columns. Note that for a $n \times n$ matrix M , the latter condition is equivalent to $M^T \vec{1} = \vec{1}$, where $\vec{1} \in \mathbb{R}^n$ has all ones as components.

Positive Markov matrices are defined as, well, positive Markov matrices.

1.1.2. *Facts.* (Perron-Frobenius theorem) Let A be a positive square matrix. Let λ_1 be A 's maximum eigenvalue in terms of absolute values. Then λ_1 is positive and has algebraic (and subsequently geometric) multiplicity 1.

Let M be a Markov matrix. Let λ_1 be M 's maximum eigenvalue in terms of absolute values. Then $\lambda_1 = 1$.

Let M' be a positive Markov matrix. Let λ_1 be M' 's maximum eigenvalue in terms of absolute values. Then $\lambda_1 = 1$ and has algebraic (and subsequently geometric) multiplicity 1.

1.1.3. *Usage.* Let M be a $n \times n$ Markov matrix. Then M specifies a discrete memoryless transition process between n states, namely the process where

$$(\forall (t, i, j) \in \mathbb{N} \times [n] \times [n]) [\Pr(\text{state } i \text{ at time } t+1 \mid \text{state } j \text{ at time } t) = M_{ij}].$$

Let $\vec{v} \in \mathbb{R}^n$ such that \vec{v} has nonnegative components and $\vec{v}^T \vec{1} = 1$ (a stochastic vector). Then \vec{v} specifies an (initial) discrete probability distribution over the n states, namely the distribution where

$$(\forall i \in [n]) [\Pr(\text{state } i \text{ at time } 0) = \vec{v}_i].$$

Then the probability distribution over the n states after t steps of the transition process specified by M is precisely $M^t \vec{v}$, or equivalently

$$(\forall (t, i) \in \mathbb{N} \times [n]) [\Pr(\text{state } i \text{ at time } t) = (M^t \vec{v})_i].$$

1.2. Graph Theory.

1.2.1. *Definitions.* A simple directed graph is defined as an unweighted directed graph without self-referential edges or multiple edges between the same origin destination pair.

For a simple directed graph with n vertices, the adjacency matrix \mathcal{A} is defined to be the $n \times n$ matrix where

$$(\forall (i, j) \in [n] \times [n]) \left(A_{ij} = \begin{cases} 1 & \text{there is an edge to } i \text{ from } j \\ 0 & \text{otherwise} \end{cases} \right).$$

1.2.2. *Facts.* For a simple directed graph with n vertices and its adjacency matrix \mathcal{A} ,

$$\begin{aligned} (\forall j \in [n]) & \left[\text{number of outgoing neighbors from vertex } j = \text{out}(j) = (\mathcal{A}_{*j})^T \vec{1} \right] \\ (\forall i \in [n]) & \left[\text{number of incoming neighbors to vertex } i = \text{in}(i) = (\mathcal{A}_{i*})^T \vec{1} \right]. \end{aligned}$$

2. ALGORITHMS

2.1. The network model. Both algorithms, PageRank and HITS, model the network of interest as a simple directed graph with websites as vertices and links as edges. This implies that there will be no self-referential links, no duplicate links between the same origin and destination pair, and no priority difference between links.

2.2. PageRank.

2.2.1. The random walk. PageRank models the behavior of a typical web surfer as a damped random walk.

- (1) The surfer starts out by visiting a random site out of all sites with equal probability.
- (2) At every step, the surfer has a probability λ of continuing surfing and a complementary $1 - \lambda$ probability of losing interest, for a predetermined λ .
 - (a) If the surfer continues ...
 - (i) ... and there are links exiting the current site, the surfer clicks on a random link (and visits the site it points to) out of those links with equal probability.
 - (ii) ... and there aren't any links exiting the current site, the surfer simply visits a random site out of all sites with equal probability.
 - (b) If the surfer loses interest, they simply visits a random site out of all sites with equal probability.

To best model a typical surfer's probability of continuing surfing, λ , also known as the damping factor, is empirically determined to be around 0.85.

2.2.2. Matrix representation. Let n be the number of websites in the network of interest. Let \mathcal{A} be the adjacency matrix for the network of interest. Let $\langle \vec{v}_t \rangle_{t \in \mathbb{N}}$ be the probability distributions describing the website the surfer is visiting at time t . Let M be the transition matrix for the random walk process.

Then $\vec{v}_0 = \vec{1}/n$, M is the $n \times n$ matrix where

$$(\forall (i, j) \in [n] \times [n]) \left[M_{ij} = \begin{cases} \frac{\lambda}{\text{out}(j)} + \frac{1-\lambda}{n} & \mathcal{A}_{ij} = 1 \\ \frac{1-\lambda}{n} & \mathcal{A}_{ij} = 0 \wedge \text{out}(j) > 0 \\ \frac{\lambda}{n} + \frac{1-\lambda}{n} & \text{out}(j) = 0 \end{cases} \right],$$

and

$$(\forall t \in \mathbb{N}) (\vec{v}_t = M^t \vec{v}_0).$$

2.2.3. Definition. The PageRank score for a given website in the network of interest is defined as the probability of a typical surfer visiting that website after an indefinitely long damped random walk. In matrix form,

$$(\forall i \in [n]) \left[\text{PageRank}(i) = \lim_{t \rightarrow \infty} (\vec{v}_t)_i = \lim_{t \rightarrow \infty} (M^t \vec{v}_0)_i \right].$$

The actually ranking is produced after sorting the websites by PageRank score from high to low.