Mathematics of Image and Data Analysis
Math 5467

PageRank

Instructor: Jeff Calder
Email: jcalder@umn.edu

http://www-users.math.umn.edu/~jwcalder/5467

Last time

- Spectral Clustering

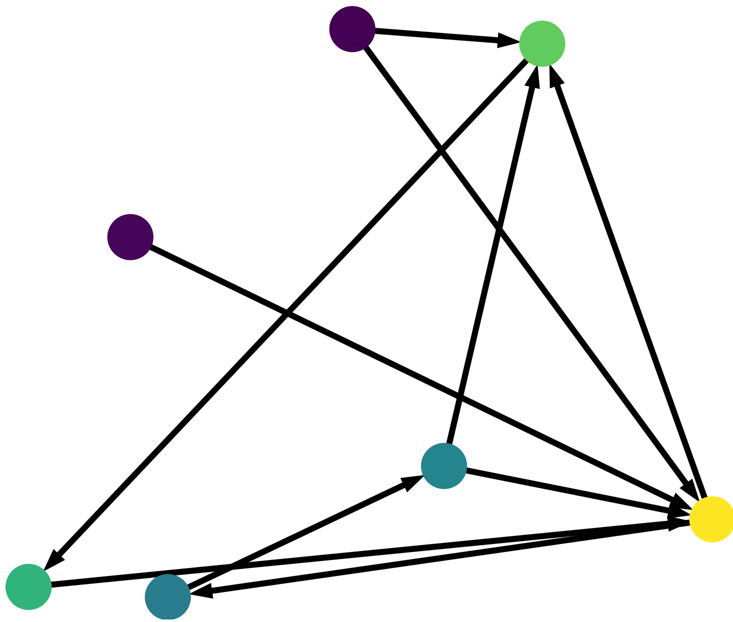
Today

- PageRank

PageRank

The PageRank algorithm ranks websites based on the link structure of the internet. It was used to sort Google search results until 2006, and has been used in

- Biology (GeneRank), chemistry, ecology, neuroscience, physics, sports, and computer systems…



PageRank

Main Idea: Take a random walk on the internet for T steps.

$$\text{Rank of site } i = \lim_{T\to\infty} \frac{1}{T}(\text{Number of times site } i \text{ is visited}).$$

Problem: Random walks can get stuck in disconnected components of the internet, and may never visit a given site i.

Solution: Every so often, the random walker teleports to a random site on the internet. The walker is called a random surfer.

Code demo

Mathematics of PageRank

To describe PageRank mathematically, we start with an adjacency matrix W

$$W(i,j) = \begin{cases} 1, & \text{if site } i \text{ links to site } j \\ 0, & \text{otherwise.} \end{cases}$$

We also have a probability transition matrix P for the random walk:

$$P(i,j) = \text{Probability of stepping from } j \text{ to } i.$$

Both P and W are $n \times n$ matrices, $n =$number of webpages.

Mathematics of PageRank

Clicking on a link at random from webpage j leads to the transition probabilities

$$P(i,j) = \frac{W(j,i)}{\underbrace{\sum_{k=1}^{n} W(j,k)}_{D(j,j)}}.$$

$$D = \begin{pmatrix} D(1,1) & & 0 \\ & \ddots & \\ 0 & & D(n,n) \end{pmatrix}$$

Exercise 1. Show that $P = W^T D^{-1}$, where D is the diagonal matrix with diagonal entries $D(i,i) = \sum_{j=1}^{n} W(i,j)$. △

Random surfer

Let $\alpha \in [0, 1)$ be the random walk probability, and let $v \in \mathbb{R}^n$ be the teleportation probability distribution. That is, $v(i) \geq 0$ for all i, and $\sum_i v(i) = 1$.

Random surfer dynamics: When at website j, the random surfer chooses the next site as follows:

1. With probability α the surfer clicks an outgoing link at random, that is, the surfer navigates to website i with probability $P(i, j)$.
2. With probability $1 - \alpha$ the surfer teleports to website i with probability $v(i)$.

Teleportation

Teleportation distribution: Common choices are

- $v(i) = 1/n$ for all i (jump to a site uniformly at random).
- (Localized PageRank) $v(i) = \delta_{ij}$ (always jump back to site j).

Localized PageRank ranks all sites based on their similarity to site j.

The PageRank vector

For $k \geq 0$ define

$$x_k(i) = \text{Probability that the random surfer is at page } i \text{ on step } k.$$

Definition 2. The **PageRank** vector x is

$$x(i) = \lim_{k \to \infty} x_k(i),$$

provided the limit exists.

Transition probabilities

To see how x_k transitions to x_{k+1} requires some probability. We condition on the location of the surfer at step k, and on the outcome of the coin flip, to obtain

$$x_{k+1}(i) = (1-\alpha)v(i) + \alpha \underbrace{\sum_{j=1}^{n} P(i,j)x_k(j)}_{(Px_k)_i}.$$

We can write this in matrix/vector form as

$$\underset{\substack{\downarrow \\ x}}{x_{k+1}} = (1-\alpha)v + \alpha P \underset{\substack{\downarrow \\ x}}{x_k}. \tag{1}$$

If x_k converges to a vector x as $k \to \infty$, then x should satisfy

$$x = (1-\alpha)v + \alpha P x.$$

Question: Does x_k converge as $k \to \infty$, and if so, how quickly does it converge?

Analysis of PageRank

We consider the PageRank equation

$$x = (1 - \alpha)v + \alpha P x. \tag{2}$$

Lemma 3. *Let $v \in \mathbb{R}^n$ and $0 \leq \alpha < 1$. Then there is a unique vector $x \in \mathbb{R}^n$ solving the PageRank equation (2). Furthermore, the following hold.*

(i) We have $\sum_{i=1}^n x(i) = \sum_{i=1}^n v(i)$.

(ii) If $v(i) \geq 0$ for all i, then $x(i) \geq 0$ for all i.

$$\|x\|_2 = \sqrt{x(1)^2 + \cdots + x(n)^2}$$

The ℓ_1-norm

It will be more convenient to work in the ℓ_1-norm $\|\cdot\|_1$ defined by

$$\|x\|_1 = \sum_{i=1}^{n} |x(i)|.$$

In the ℓ_1-norm, the transition matrix P is non-expansive.

$$\boxed{P(i,j) = \frac{w(j,i)}{\sum_{k=1}^{n} w(j,k)}}$$

Proposition 4. *We have* $\|Px\|_1 \leq \|x\|_1$.

Proof:

$$\|Px\|_2 = \sum_{i=1}^{n} |(Px)_i|$$

$$= \sum_{i=1}^{n} \left| \sum_{j=1}^{n} P(i,j) x(j) \right|$$

$$\leq \sum_{i=1}^{n} \sum_{j=1}^{n} |P(i,j)| \, |x(j)|$$

$$= \sum_{j=1}^{n} |x(j)| \underbrace{\sum_{i=1}^{n} P(i,j)}_{=1} = \|x\|_1 \qquad \blacksquare$$

$$\sum_{i=1}^{n} P(i,j) = \frac{\sum_{i=1}^{n} W(j,i)}{\sum_{k=1}^{n} W(j,k)} = 1$$

Proof of lemma 3

$$x = (1-\alpha)v + \alpha P x.$$

$$x - \alpha P x = (1-\alpha) v$$

$$\underbrace{(1-\alpha)^{-1}(I - \alpha P)}_{A} x = v$$

$$A x = v$$

Look at $\ker(A)$. Claim $\ker(A) = \{0\}$.

Let $z \in \ker(A)$, so $Az = 0$. Need to show $z = 0$.

$A = (1-\alpha)^{-1}(I - \alpha P)$

$$Az = 0 \longleftrightarrow z - \alpha P z = 0$$

$$\longleftrightarrow z = \alpha P z$$

$$\|z\|_1 = \|\alpha P z\|_1$$

$$= \alpha \|Pz\|_1 \le \alpha \|z\|_1$$

$$(1-\alpha)\|z\|_1 \le 0, \quad \alpha < 1$$

$$\Rightarrow \quad 1-\alpha = 0 \quad \text{or} \quad \|z\|_2 = 0$$

$$\alpha < 1 \Rightarrow \|z\|_2 = 0 \quad \blacksquare$$

(i) We have $\sum_{i=1}^{n} x(i) = \sum_{i=1}^{n} v(i)$.

(ii) If $v(i) \geq 0$ *for all* i, *then* $x(i) \geq 0$ *for all* i.

(i)

$$x = (1-\alpha)v + \alpha P x.$$

$$x(i) = (1-\alpha)v(i) + \alpha \sum_{j=1}^{n} P(i,j)\,x(j)$$

$$\sum_{i=1}^{n} x(i) = (1-\alpha)\sum_{i=1}^{n} v(i) + \alpha \sum_{i=1}^{n}\sum_{j=1}^{n} P(i,j)\,x(j)$$

$$= (1-\alpha)\sum_{i=1}^{n} v(i) + \alpha \sum_{j=1}^{n} \underbrace{\left[\sum_{i=1}^{n} P(i,j)\right]}_{=1} x(j)$$

$$= (1-\alpha)\sum_{i=1}^{n} v(i) + \alpha \sum_{j=1}^{n} x(j)$$

$$\cancel{(1-\alpha)}\sum_{i=1}^{n} x(i) = \cancel{(1-\alpha)}\sum_{i=1}^{n} v(i) \qquad \blacksquare$$

$$\alpha < 1$$

(i) We have $\sum_{i=1}^n x(i) = \sum_{i=1}^n v(i)$.

(ii) If $v(i) \geq 0$ *for all* i*, then* $x(i) \geq 0$ *for all* i.

$$x(i) = (1-\alpha)v(i) + \alpha \sum_{j=1}^{n} P(i,j)\, x(j)$$

$$|x(i)| = \left| (1-\alpha)v(i) + \alpha \sum_{j=1}^{n} P(i,j)\, x(j) \right|$$

$v(i) \geq 0$

$$\leq (1-\alpha)v(i) + \alpha \sum_{j=1}^{n} P(i,j)\, |x(j)|$$

$$\sum_{i=1}^{n} |x(i)| \leq (1-\alpha)\sum_{i=1}^{n} v(i) + \alpha \sum_{j=1}^{n} |x(j)|$$

$$= (1-\alpha)\sum_{i=1}^{n} x(i) + \alpha \sum_{j=1}^{n} |x(j)|$$

$$\cancel{(1-\alpha)}\sum_{i=1}^{n} |x(i)| \le \cancel{(1-\alpha)}\sum_{i=1}^{n} x(i)$$

$$\Longrightarrow \sum_{i=1}^{n} |x(i)| = \sum_{i=1}^{n} x(i)$$

$$\sum_{i=1}^{n} \underbrace{\left(|x(i)| - x(i)\right)}_{} = 0$$

$= 0$

$\Rightarrow \quad x(i) = |x(i)|$

≥ 0

$\blacksquare$

Eigenvector problem

Remark 5. When v is a probability distribution, it is common to re-write the PageRank problem (2) as an eigenvector problem

$$P_\alpha x = x$$

where

$$P_\alpha := (1-\alpha)v\mathbf{1}^T + \alpha P.$$

$x(i) \geq 0$

$v(i) \geq 0$

$\sum x(i) = 1$

$\sum v(i) = 1$

$\mathbb{1}^T x = 1$

$\mathbb{1}^T v = 1$

$$x - \alpha P x = (1-\alpha)v$$

$$\alpha P x + (1-\alpha)v\mathbb{1}^T x = x$$

$$\underbrace{(\alpha P + (1-\alpha)v\mathbb{1}^T)}_{P_\alpha} x = x$$

$P_\alpha x = x$

Convergence of the PageRank iteration

Let $v \in \mathbb{R}^n$ and $0 \leq \alpha < 1$. Let x_k satisfy the PageRank iteration

$$x_{k+1} = (1-\alpha)v + \alpha P x_k,$$

and let x be the unique solution of the PageRank problem

$$x = (1-\alpha)v + \alpha P x.$$

Theorem 6. *We have*

$$\|x_k - x\|_1 \leq \underset{z}{\alpha^k} \|x_0 - x\|_1. \tag{3}$$

Since $0 \leq \alpha < 1$, this is convergence of $x_k \to x$ with a linear convergence rate of α.

Proof:

$$x_{k+1} = (1-\alpha)v + \alpha P x_k,$$

$$x = (1-\alpha)v + \alpha P x.$$

$$x_{k+1} - x = \alpha P x_k - \alpha P x$$

$$= \alpha P(x_k - x)$$

$$\|x_{k+1} - x\|_1 = \|\alpha P(x_k - x)\|_1$$

$$= \alpha \|P(x_k - x)\|_1$$

$$\|Px\|_1 \le \|x\|_1 \qquad \le \alpha \|x_k - x\|_1$$

Induction $\|x_{k+1} - x\|_1 \leq \alpha^{k+1} \|x_0 - x\|_1$ $\blacksquare$

Power iteration

Remark 7. In the eigenvector formulation discussed above, the PageRank iteration $x_{k+1} = P_\alpha x_k$ is basically the power iteration to find the largest eigenvector of P. The normalization step is not needed since $\|x_k\|_1 = 1$ for all k.

Personalized PageRank for image retrieval (.ipynb)