NLA 2021-2022
The PageRank algorithm

Martin Sombra

10 November 2021

The PageRank algorithm and the eigenvalue problem

PageRank is the basic algorithm of the Google search machine. It had a huge influence on the development and structure of the internet, since it determines which kind of information and services are accessed more often

There are three main steps for ranking web pages:

1. Crawl the web and locate all public pages
2. Index the data from (1) to allow to search for keywords and phrases within it
3. Rate the importance of these pages

We assume that (1) and (2) are given, and focus on (3):

How can we define and quantify the "importance" of webpages?

The score vector

Suppose that our web of interest contains n pages, each indexed by an integer $1 \leqslant k \leqslant n$

We denote by x_k the *score* of the page k: nonnegative real numbers such $x_j > x_k$ indicates that the page j is more important than the page k

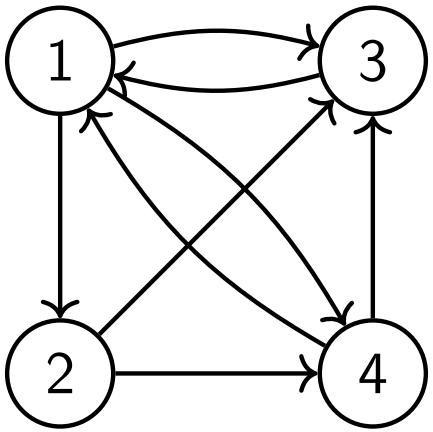
The *score vector*

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n_{\geqslant 0}$$

is normalized so that $\sum_{k=1}^{n} x_k = 1$

A directed graph

The web can be interpreted as a *directed graph*, where the pages correspond to the nodes and the links to the arrows, e.g.

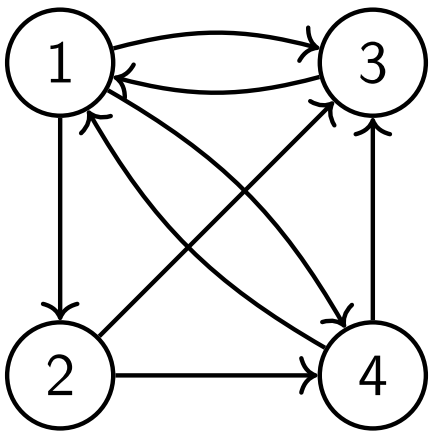


The basic idea

The web is understood as a "democracy" where pages "vote" for the importance of the other pages by linking to them

The simplest approach would consist in *counting links*: in the example this would give

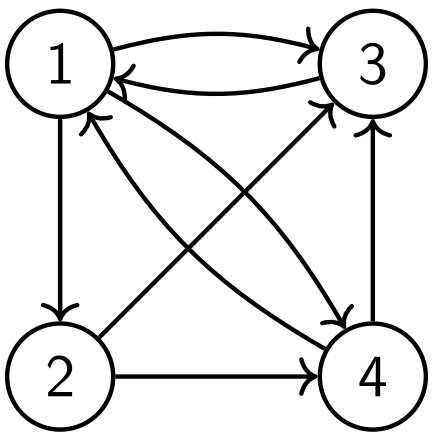
$$x_1 = \frac{2}{8}, \quad x_2 = \frac{1}{8}, \quad x_3 = \frac{3}{8}, \quad x_4 = \frac{2}{8}$$



The basic idea (cont.)

But this forgets an important aspect: receiving a link from an important page should count more!

For instance, pages 1 and 4 have the same number of backlinks (links pointing to them), but page 1 is linked by the important page 3, whereas 4 is linked by the less important page 1



The basic idea (cont.)

Another aspect to be taken into account is that a page should not increase its influence by increasing its number of outgoing links

If the page j has n_j links, then each should contribute with the score x_j/n_j to the page they link

$\rightsquigarrow$ the overall influence of the page j is its score x_j

The basic idea (cont.)

The score vector should satisfy the equations

$$x_k = \sum_{j \in L_k} \frac{x_j}{n_j}, \quad k = 1, \ldots, n,$$

with $L_k \subset \{1, \ldots, n\}$ the subset of the indices of the pages linking to the page k

The corresponding *link matrix* $A \in \mathbb{R}^{n \times n}$ is

$$A = \begin{cases} \frac{1}{n_j} & \text{if the page } j \text{ links to the page } k \\ 0 & \text{else} \end{cases}$$

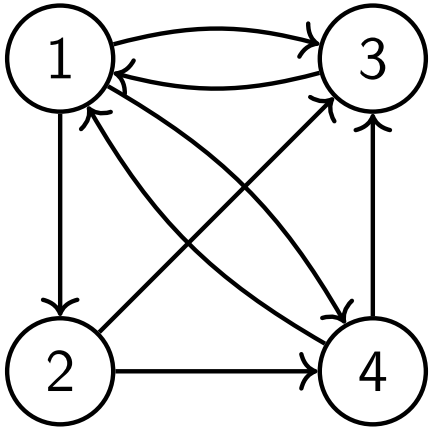
The system of linear equations above is equivalent to

$$Ax = x$$

The score vector is an eigenvector of A with eigenvalue 1

Example (cont.)

In the example



we have

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

Example (cont.)

The normalized eigenvector of A is

$$x = \begin{bmatrix} \frac{12}{31} \\ \frac{4}{31} \\ \frac{9}{31} \\ \frac{6}{31} \end{bmatrix} = \begin{bmatrix} 0.387 \\ 0.129 \\ 0.290 \\ 0.194 \end{bmatrix}$$

The page 3 (linked by all the others) is less important than the page 1:

Indeed, the page 3 gives all its "vote" to the page 1 and together with the link from the page 2, gives the page 1 the highest score

Column stochastic matrices

For simplicity: assume that the web has no *dangling nodes*, that is, pages without outgoing links

$\rightsquigarrow$ the link matrix A is *column stochastic*: its entries are nonnegative real numbers and each column sums up to 1:

$$\sum_{i=1}^{n} a_{i,j} = 1, \quad j = 1, \ldots, n$$

Column stochastic matrices (cont.)

Column stochastic matrices have $\lambda = 1$ as one of its eigenvalues: indeed, A and its transposed A^T have the same *characteristic polynomial*:

$$\chi_{A^T} = \det(A^T - t\,\mathbb{1}_n) = \det(A - t\,\mathbb{1}_n)^T = \det(A - t\,\mathbb{1}_n) = \chi_A,$$

$\rightsquigarrow$ the eigenvalues of A and of A^T coincide

Since A is column stochastic then A^T is row stochastic, and so

$$A^T \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

Hence $\lambda = 1$ is an eigenvalue of A^T, and so also an eigenvalue of A

The eigenspace

Denote by

$$V_1(A) = \{x \in \mathbb{R}^n \mid Ax = x\}.$$

the eigenspace of the eigenvalue $\lambda = 1$

Since this is an eigenvalue of A, we have that $V_1(A) \neq 0$

We would like to have

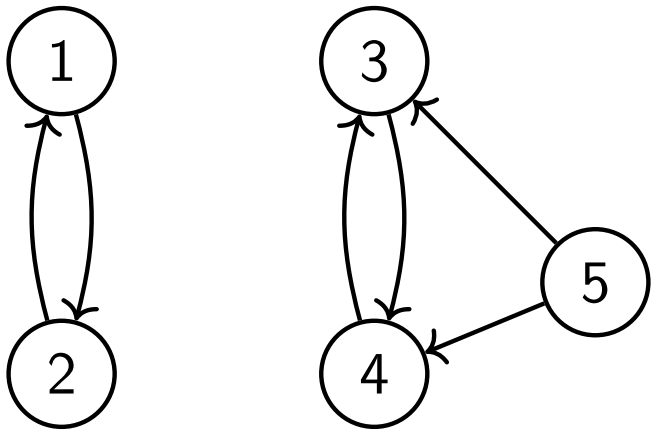
$$\dim(V_1(A)) = 1$$

so that it defines at most one normalized eigenvector

Not always true...

Example

The web



gives the link matrix

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

Example (cont.)

Both

$$x = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \text{and} \quad y = \begin{bmatrix} 0 \\ 0 \\ \frac{1}{2} \\ \frac{1}{2} \\ 0 \end{bmatrix}$$

are eigenvectors of A for the eigenvalue 1, and so $V_1(A)$ has dimension at least 2

Which of the vectors in this subspace should be used as a score?

Disconnected webs

In general, this situation arises when the web is disconnected:

Indeed, if the web consists of t subwebs, then the corresponding link matrix A splits into t blocks, and we have that

$$\dim V_1(A) \geqslant t$$

A uniform perturbation

To avoid this phenomenon, we modify the link matrix by adding to it a multiple of the *uniform matrix*

$$S = \left[\tfrac{1}{n}\right]_{i,j} = \begin{bmatrix} \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & & \vdots \\ \frac{1}{n} & \cdots & \frac{1}{n} \end{bmatrix}.$$

Then for $0 \leqslant \alpha \leqslant 1$ we set

$$M = (1 - \alpha)\, A + \alpha\, S$$

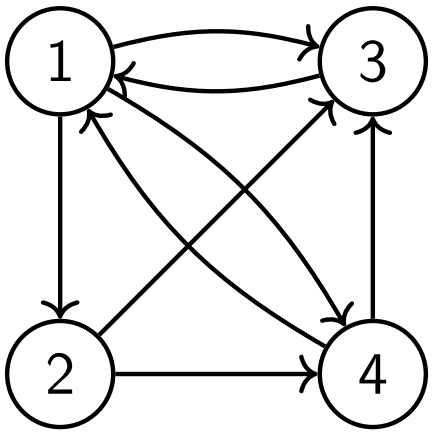
If $\alpha > 0$ then

$$\dim V_1(M) = 1$$

Google's value: $\alpha = 0.15$

Examples (cont.)

For the graph



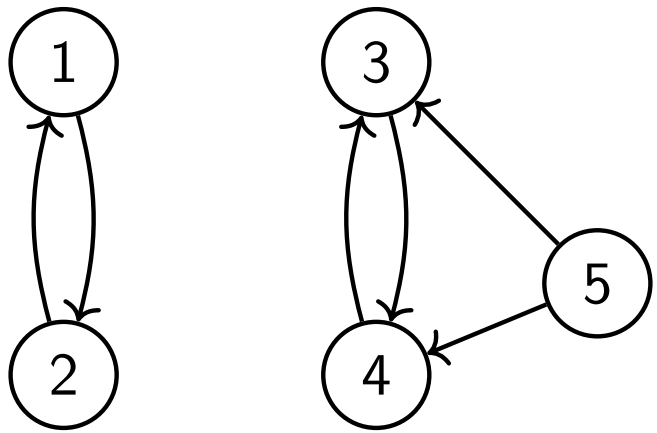
the modified link matrix M gives the scores

$$x_1 = 0.368, \quad x_2 = 0.142, \quad x_3 = 0.288, \quad x_4 = 0.202.$$

Slightly different to those for the link matrix A but giving the same order of importance to the pages

Examples (cont.)

For the graph



it gives

$$x_1 = 0.2, \quad x_2 = 0.2, \quad x_3 = 0.285, \quad x_4 = 0.285, \quad x_5 = 0.003,$$

allowing to compare the different pages

The eigenvalues of the modified link matrix

The fact that $\dim(V_1(M)) = 1$ follows from the following result:

Theorem: Let

$$\lambda_1 = 1, \lambda_2, \ldots, \lambda_n$$

be the eigenvalues of A, repeated according to their multiplicities

Then

$$|\lambda_i| \leqslant 1 \quad \text{for all } i$$

and the eigenvalues of M are

$$\lambda_1 = 1, (1-\alpha)\,\lambda_2, \ldots, (1-\alpha)\,\lambda_n$$

Proof

For each i choose $x = (x_1, \ldots, x_n) \neq 0$ such that $A\,x = \lambda_i\,x$

Then

$$\|A\,x\|_1 \geqslant |\lambda_i|\,\|x\|_1$$

and

$$\|A\,x\|_1 = \sum_{i=1}^{n} \Big| \sum_{j=1}^{n} a_{i,j}\,x_j \Big| \leqslant \sum_{j=1}^{n} \Big(\sum_{i=1}^{n} a_{i,j} \Big) |x_j| \leqslant \sum_{j=1}^{n} |x_j| = \|x\|_1$$

because A is column stochastic

Hence $|\lambda_i|\,\|x\|_1 \leqslant \|x\|_1$ and so $|\lambda_i| \leqslant 1$, as stated

Proof (cont.)

Next set

$$e = \begin{bmatrix} \frac{1}{\sqrt{n}} \\ \vdots \\ \frac{1}{\sqrt{n}} \end{bmatrix} \in \mathbb{R}^n$$

This is a unit n-vector such that $S = e\,e^T$

Let U_1 be an $n \times (n-1)$ matrix completing e to an orthonormal $n \times n$ matrix $U = [e\,U_1]$. Then

$$U^T A\,U = \begin{matrix} 1 \\ n-1 \end{matrix}\begin{bmatrix} e^T A \\ U_1^T A \end{bmatrix}\begin{bmatrix} 1 & n-1 \\ e & U_1 \end{bmatrix} = \begin{bmatrix} e^T \\ U_1^T A \end{bmatrix}\begin{bmatrix} e & U_1 \end{bmatrix}$$

$$= \begin{matrix} 1 \\ n-1 \end{matrix}\begin{bmatrix} 1 & e^T U_1 \\ U_1^T A\,e & U_1^T A\,U_1 \end{bmatrix} = \begin{bmatrix} 1 & \mathbb{0} \\ w & T \end{bmatrix}$$

where the second equality follows from the fact that A is column stochastic, and the fourth from the fact that e is orthogonal to U_1

Proof (cont.)

Since $U^T A U$ is similar to A, it has the same eigenvalues

$\rightsquigarrow$ the eigenvalues of T are $\lambda_2, \ldots, \lambda_n$

We have that

$$U^T e = \begin{matrix} 1 \\ n-1 \end{matrix} \begin{bmatrix} e^T \\ U_1^T \end{bmatrix} e = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^n$$

and so

$$U^T M U = (1-\alpha)\, U^T A U + \alpha\, U^T e\, e^T U$$

$$= (1-\alpha) \begin{bmatrix} 1 & \mathbb{0} \\ w & T \end{bmatrix} + \alpha \begin{bmatrix} 1 \\ \mathbb{0} \end{bmatrix} \begin{bmatrix} 1 & \mathbb{0} \end{bmatrix} = \begin{bmatrix} 1 & \mathbb{0} \\ (1-\alpha)\, w & (1-\alpha)\, T \end{bmatrix}$$

Hence the eigenvalues of M are $1, (1-\alpha)\lambda_2, \ldots, (1-\alpha)\lambda_n$, as stated.

Computing the score vector

To compute the score vector x we apply the *power method*, starting from a well chosen initial vector x_0 and iterating

$$x_k \leftarrow \frac{M\, x_{k-1}}{\|M\, x_{k-1}\|_1}, \quad k \geqslant 1.$$

This iterative method converges towards x, the normalized eigenvector of A corresponding to the largest eigenvalue $\lambda = 1$

The rate of convergence is *linear* and depends on the gap between the largest eigenvalue and the other ones:

the number of correct digits in base b of the k-th approximant x_k is bounded below by

$$-\log_b \|x - x_k\|_1 \geqslant k \log_b \left(\frac{1}{|\lambda_2|}\right) + \text{constant}$$

Computing the score vector (cont.)

The theorem implies that

$$|\lambda_2| \leqslant 1 - \alpha = 0.85$$

For the base $b = 10$ we have that

$$\log_b \left(\frac{1}{|\lambda_2|} \right) \geqslant 0.07$$

and so

$$-\log_b \|x - x_k\|_1 \geqslant 0.07\, k + \text{constant}$$

Computing the score vector (cont.)

Currently there are $\approx 4 \cdot 10^8$ active public webpages

Each iteration of the power method consists of a matrix-vector multiplication. In a practical implementation, it is computed for $v \in \mathbb{R}^n_{\geqslant 0}$ with $\|v\|_1 = 1$, as

$$M\,v = (1-\alpha)\,A\,v + \alpha \begin{bmatrix} \frac{1}{n} \\ \vdots \\ \frac{1}{n} \end{bmatrix}$$

because the multiplication of v by the uniform matrix S gives

$$S\,v = \begin{bmatrix} \frac{1}{n} \\ \vdots \\ \frac{1}{n} \end{bmatrix}$$