Chapter 9: What's Most Important

Panos Louridas

Athens University of Economics and Business

Real World Algorithms
A Beginners Guide
The MIT Press

Outline

1. PageRank
2. PageRank Calculation

Searching in Search Engines

- When we search for a web page, there are typically many web pages that match our search terms.
- How do we know which one of them is the most relevant to our search?
- One answer to that question is to rank the web pages matching our query based on their importance.

Importance and Graphs

- The web can be represented as a graph, where pages are nodes and links between the pages are edges.
- When we search for a web page, we find a subset of the nodes of the web graph matching our query.
- How do we rank the importance of the nodes, i.e., the pages?
- Sergey Brin and Larry page provided an answer to this question, working on it as doctoral students at Stanford.
- The solution is called PageRank, was published in 1998, and lay the foundation for the success of Google.

Basic Idea

- The importance of a page depends on the importance of the pages that point to it.
- If a page P_j points to $|P_j|$ pages, then the P_j page contributes $1/|P_j|$ of its importance to each page to which it points.

The Formula

- We denote the importance of a page with $r(P_i)$.
- We will use B_{P_i} to denote the pages that point to page P_i (that have *backlinks* to it).

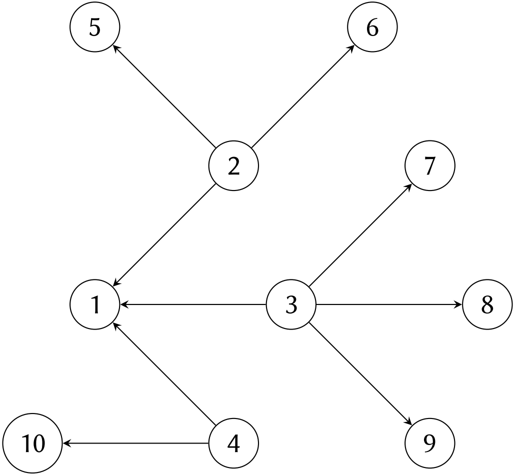
Definition

The importance of a page is defined as:

$$r(P_i) = \sum_{P_j \in B_{P_i}} \frac{r(P_j)}{|P_j|}$$

The importance thus defined is called PageRank.

Example



PageRank of Node 1

The PageRank of node 1 in the graph is:

$$r(P_1) = \frac{r(P_2)}{3} + \frac{r(P_3)}{4} + \frac{r(P_4)}{2}$$

Outline

The Chicken or the Egg?

- To calculate the PageRank of a page we need the PageRank values of the pages that point to it.
- To calculate the PageRank of those pages we need to calculate the PageRank of the pages that point to them.
- And so on and so forth.

Iterative PageRank Calculation

We can calculate the PageRank of a page iteratively as follows:

$$r_{k+1}(P_i) = \sum_{P_j \in B_{P_i}} \frac{r_k(P_j)}{|P_j|}$$

Questions

- What will the initial values be?
- Does the iterative procedure converge?
- Does it converge to a reasonable result?

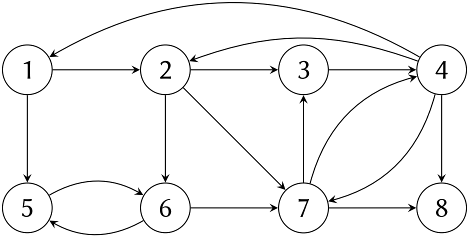
PageRank Matrix Calculation

Definition

The *hyperlink matrix* is defined as follows:

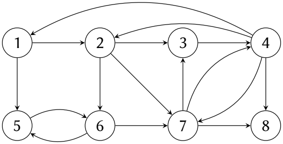
$$H[i,j] = \begin{cases} 1/|P_i|, & P_i \in B_{P_j} \\ 0, & \text{otherwise} \end{cases}$$

Hyperlink Matrix Example (1)



Hyperlink Matrix Example (2)

$$
H = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \\ P_7 \\ P_8 \end{array}
\begin{array}{c}
\begin{array}{cccccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 & P_7 & P_8 \end{array} \\
\begin{bmatrix}
0 & 1/2 & 0 & 0 & 1/2 & 0 & 0 & 0 \\
0 & 0 & 1/3 & 0 & 0 & 1/3 & 1/3 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
1/4 & 1/4 & 0 & 0 & 0 & 0 & 1/4 & 1/4 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1/2 & 0 & 1/2 & 0 \\
0 & 0 & 1/3 & 1/3 & 0 & 0 & 0 & 1/3 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
\end{array}
$$



Some Notation

$$\pi = \begin{bmatrix} \pi[0] \\ \pi[1] \\ \vdots \\ \pi[n-1] \end{bmatrix} = \begin{bmatrix} r(P_1) \\ r(P_2) \\ \vdots \\ r(P_n) \end{bmatrix}$$

$$\pi^T = \begin{bmatrix} r(P_1) & r(P_2) & \cdots & r(P_n) \end{bmatrix} = \begin{bmatrix} \pi[0] & \pi[1] & \cdots & \pi[n-1] \end{bmatrix}$$

T = transpose.

Power Method

Definition

The *power method* is the following iterative calculation:

$$\pi_{k+1}^T = \pi_k^T H$$

Explanation

If we have two matrices C and D, the product is:

$$E[i,j] = \sum_{t=0}^{n-1} C[i,t]D[t,j]$$

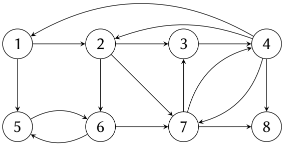
Therefore element i of $\pi_{k+1}^T = \pi_k^T H$ is:

$$\begin{aligned}\pi_{k+1}[i] &= \sum_{t=0}^{n-1} \pi_k[t]H[t,i] \\ &= \pi_k[0]H[0,i] + \pi_k[1]H[1,i] + \cdots + \pi_k[n-1]H[n-1,i]\end{aligned}$$

Example

$$r_{k+1}(P_1) = \frac{r_k(P_4)}{4}$$
$$r_{k+1}(P_2) = \frac{r_k(P_1)}{2} + \frac{r_k(P_4)}{4}$$
$$r_{k+1}(P_3) = \frac{r_k(P_2)}{3} + \frac{r_k(P_7)}{3}$$
$$r_{k+1}(P_4) = r_k(P_3) + \frac{r_k(P_7)}{3}$$
$$r_{k+1}(P_5) = \frac{r_k(P_1)}{2} + \frac{r_k(P_6)}{2}$$
$$r_{k+1}(P_6) = \frac{r_k(P_2)}{3} + r_k(P_5)$$
$$r_{k+1}(P_7) = \frac{r_k(P_2)}{3} + \frac{r_k(P_4)}{4} + \frac{r_k(P_6)}{2}$$
$$r_{k+1}(P_8) = \frac{r_k(P_4)}{4} + \frac{r_k(P_7)}{3}$$

$$\pi^T H =$$
$$\begin{bmatrix} r_k(P_1) & r_k(P_2) & r_k(P_3) & r_k(P_4) & r_k(P_5) & r_k(P_6) & r_k(P_7) & r_k(P_8) \end{bmatrix}$$
$$\times \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \\ P_7 \\ P_8 \end{array} \begin{array}{c} \begin{array}{cccccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 & P_7 & P_8 \end{array} \\ \begin{bmatrix} 0 & 1/2 & 0 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 0 & 1/3 & 1/3 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1/4 & 1/4 & 0 & 0 & 0 & 0 & 1/4 & 1/4 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 1/3 & 1/3 & 0 & 0 & 0 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$$



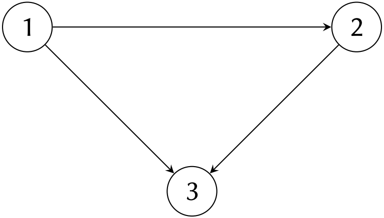
Questions for the Power Method

- Does the power method converge? If yes, the vector to which it converges is called a *stationary vector*.
- If the power method converges, does the stationary vector contain reasonable PageRank values?

Power Method Initialization

- Initially we can set all PageRank values equal to $1/n$, where n is the number of pages.
- In other words, we start by giving equal importance to all pages.

The Problem with Sinks



PageRank Calculation (1)

$$\begin{bmatrix}1/3 & 1/3 & 1/3\end{bmatrix}\begin{bmatrix}0 & 1/2 & 1/2\\ 0 & 0 & 1\\ 0 & 0 & 0\end{bmatrix} = \begin{bmatrix}0 & 1/6 & 1/2\end{bmatrix}$$

$$1/3 \times 0 + 1/3 \times 0 + 1/3 \times 0 = 0$$

$$1/3 \times 1/2 + 1/3 \times 0 + 1/3 \times 0 = 1/6$$

$$1/3 \times 1/2 + 1/3 \times 1 + 1/3 \times 0 = 1/2$$

PageRank Calculation (2)

$$\begin{bmatrix} 0 & 1/6 & 1/2 \end{bmatrix} \begin{bmatrix} 0 & 1/2 & 1/2 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1/6 \end{bmatrix}$$

$$0 \times 0 + 1/6 \times 0 + 1/2 \times 0 = 0$$

$$0 \times 1/2 + 1/6 \times 0 + 1/2 \times 0 = 0$$

$$0 \times 1/2 + 1/6 \times 1 + 1/2 \times 0 = 1/6$$

PageRank Calculation (3)

$$\begin{bmatrix} 0 & 0 & 1/6 \end{bmatrix} \begin{bmatrix} 0 & 1/2 & 1/2 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$$

$$0 \times 0 + 0 \times 0 + 1/6 \times 0 = 0$$

$$0 \times 1/2 + 0 \times 0 + 1/6 \times 0 = 0$$

$$0 \times 1/2 + 0 \times 1 + 1/6 \times 0 = 0$$

Dangling Nodes

- The power method does not work with *dangling nodes*, i.e., nodes that point nowhere.
- The dangling nodes take out importance from other nodes, without giving back anything.

The Random Surfer

- Suppose that matrix H dictates the behavior of a surfer that jumps from page to page based on the probabilities of the cells in the line the surfer has landed.
- If the surfer lands on page P_i, the next page to visit will be page P_j with probability $H[i, j]$.
- Example: in H below, if the surfer is on page 6, the surfer will jump to page 5 with probability 1/2 or to page 7 with probability 1/2.

$$
\begin{array}{c}
\\
P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \\ P_7 \\ P_8
\end{array}
\begin{array}{c}
\begin{array}{cccccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 & P_7 & P_8 \end{array} \\
\begin{bmatrix}
0 & 1/2 & 0 & 0 & 1/2 & 0 & 0 & 0 \\
0 & 0 & 1/3 & 0 & 0 & 1/3 & 1/3 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
1/4 & 1/4 & 0 & 0 & 0 & 0 & 1/4 & 1/4 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1/2 & 0 & 1/2 & 0 \\
0 & 0 & 1/3 & 1/3 & 0 & 0 & 0 & 1/3 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
\end{array}
$$

Random Surfer Dead ends

- Then, if the surfer lands on a page without any outgoing links, that is, on a line with all zero cells, the surfer cannot go anywhere.
- To avoid that, we put in each cell of such lines the value $1/n$, where n is the number of pages.
- That is like giving the surfer a teleportation device that can transport the bearer to any other point in the graph in random when stuck in a page with no exit.

The Matrix A

Definition

For each matrix H we define matrix A as the matrix with all cells equal to zero, except for the lines where H has all cells zero, where we set all cells of A equal to $1/n$, where n is the number of pages.

The Matrix S

Definition

For each matrix H we define matrix S as follows:

$$S = H + A$$

Definition of Matrix A

Definition

Suppose we have the column vector w:

$$w[i] = \begin{cases} 1, & |P_i| = 0 \\ 0, & \text{otherwise} \end{cases}$$

Then matrix A is:

$$A = \frac{1}{n} w \mathbf{e}^T$$

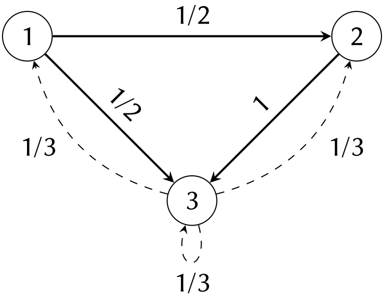
where $\mathbf{e}$ is the column vector with all elements equal to one, therefore $\mathbf{e}^T$ is the row vector with all elements equal to one. Then we have:

$$S = H + A = H + \frac{1}{n} w \mathbf{e}^T$$

Example

$$S = H + A = \begin{bmatrix} 0 & 1/2 & 1/2 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 \end{bmatrix} = \begin{bmatrix} 0 & 1/2 & 1/2 \\ 0 & 0 & 1 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}$$

The Resulting Graph



The Power Method with Matrix S

Definition

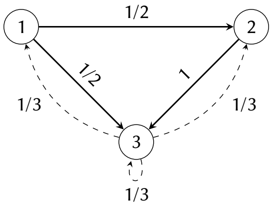
The power method with matrix S is:

$$\pi_{k+1}^T = \pi_k^T S$$

Using S

If we use S, the PageRank values of the new graph are:

$$\pi^T = \begin{bmatrix} 0.18 & 0.27 & 0.55 \end{bmatrix}$$



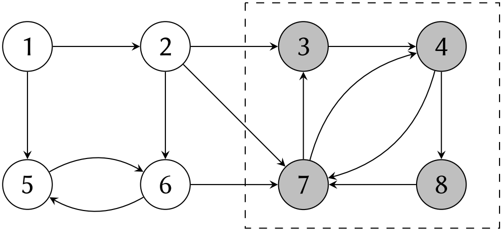
Stochastic Matrices

Definition

A matrix whose elements are greater than or equal to zero and the sum of each row is equal to one is called a *stochastic matrix*.

We call it that way because it represents a *stochastic process*, that is, a process determined by chance, such as the random surfer. The sum of the probabilities in each row is equal to one, and probabilities cannot be less than zero.

The Problem of Disconnected Cycles



Matrix S for the Previous Graph

$$
S = H = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \\ P_7 \\ P_8 \end{array}
\begin{array}{c}
\begin{array}{cccccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 & P_7 & P_8 \end{array} \\
\begin{bmatrix}
0 & 1/2 & 0 & 0 & 1/2 & 0 & 0 & 0 \\
0 & 0 & 1/3 & 0 & 0 & 1/3 & 1/3 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 1/2 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1/2 & 0 & 1/2 & 0 \\
0 & 0 & 1/2 & 1/2 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0
\end{bmatrix}
\end{array}
$$

Applying the Power Method

If we apply the power method on that graph, we will find out that it converges to the following values:

$$\pi^T = \begin{bmatrix} 0 & 0 & 0.17 & 0.33 & 0 & 0 & 0.33 & 0.17 \end{bmatrix}$$

That means that the disconnected cycle consisting of the nodes 3, 4, 7, 8 takes all PageRanks from the rest of the graph.

Solving the Disconnected Cycles Problem

- We endow the random surfer's teleportation device with yet another capability.
- The surfer follows graph S with probability a.
- With probability $(1 - a)$ the surfer jumps to a random node in the graph.
- This corresponds to everyday experience, in that when we browse, we do not always follow the links from one page to the next.

The Matrix G

We create the matrix:

$$G = \alpha S + (1 - \alpha)\frac{1}{n}\mathbf{e}\mathbf{e}^T$$

where α is the probability that the surfer will follow graph S and $(1 - \alpha)$ is the probability that it will jump to a random node in the graph.

Definition

Matrix G is called the *Google matrix*.

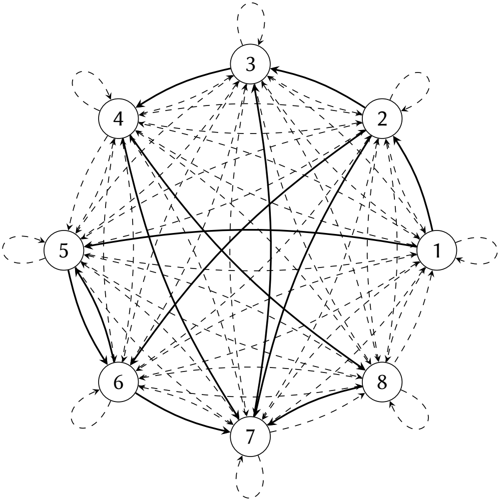
Properties of Matrix G

- G is a stochastic matrix.
- G is also a *primitive* matrix. A matrix M is called primitive if there exists a power p such that all elements of M^p are positive. G is primitive by definition, as all the elements of $G^1 = G$ are positive.

Convergence of the Power Method for Matrix G

- If a matrix is stochastic and primitive, like G, then the power method converges to a unique column vector with positive values.
- In fact, the power method converges to that column vector irrespectively of the initial column vector π_1. So, we don't even need to set all initial PageRank values equal to $1/n$.
- It follows that the power method can be used to calculate the PageRank values of a graph.

The Graph Corresponding to G



The graph corresponding to G is a complete graph.

The G Matrix for the Example

$$G = \begin{bmatrix} \frac{3}{160} & \frac{71}{160} & \frac{3}{160} & \frac{3}{160} & \frac{71}{160} & \frac{3}{160} & \frac{3}{160} & \frac{3}{160} \\ \frac{3}{160} & \frac{3}{160} & \frac{29}{96} & \frac{3}{160} & \frac{3}{160} & \frac{29}{96} & \frac{29}{96} & \frac{3}{160} \\ \frac{3}{160} & \frac{3}{160} & \frac{3}{160} & \frac{139}{160} & \frac{3}{160} & \frac{3}{160} & \frac{3}{160} & \frac{3}{160} \\ \frac{3}{160} & \frac{3}{160} & \frac{3}{160} & \frac{3}{160} & \frac{3}{160} & \frac{3}{160} & \frac{71}{160} & \frac{71}{160} \\ \frac{3}{160} & \frac{3}{160} & \frac{3}{160} & \frac{3}{160} & \frac{3}{160} & \frac{139}{160} & \frac{3}{160} & \frac{3}{160} \\ \frac{3}{160} & \frac{3}{160} & \frac{3}{160} & \frac{3}{160} & \frac{71}{160} & \frac{3}{160} & \frac{71}{160} & \frac{3}{160} \\ \frac{3}{160} & \frac{3}{160} & \frac{71}{160} & \frac{71}{160} & \frac{3}{160} & \frac{3}{160} & \frac{3}{160} & \frac{3}{160} \\ \frac{3}{160} & \frac{3}{160} & \frac{3}{160} & \frac{3}{160} & \frac{3}{160} & \frac{3}{160} & \frac{139}{160} & \frac{3}{160} \end{bmatrix}$$

$$S = H = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \\ P_7 \\ P_8 \end{array} \begin{array}{c} \begin{array}{cccccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 & P_7 & P_8 \end{array} \\ \begin{bmatrix} 0 & 1/2 & 0 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 0 & 1/3 & 1/3 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 1/2 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \end{array}$$

The G matrix for S and $\alpha = 0.85 = 17/20$.
In the first line the elements are $\frac{3}{20} \times \frac{1}{8}$ and $\frac{17}{20} \times \frac{1}{2} + \frac{3}{20} \times \frac{1}{8}$.

The Power Method for Matrix G

Definition

The power method for matrix G is:

$$\pi_{k+1}^T = \pi_k^T G$$

PageRank Calculation for Matrix G

$$\pi^T = \begin{bmatrix} 0.02^+ & 0.03^+ & 0.15^+ & 0.26^- & 0.06^+ & 0.08^+ & 0.28^- & 0.13^- \end{bmatrix}$$

Efficient PageRank Calculation (1)

- The speed of convergence of the method depends on the value of α.
- If α is close to one, then the method converges slowly, and the graph G is more like graph S.
- If α is close to zero, then the method converges fast, but the graph G looks less like graph S and more like a complete graph with equal weights everywhere.
- Brin and Page selected the value $\alpha = 0.85$.

Efficient PageRank Calculation (2)

- If we do the math we find that:

$$\pi_{k+1}^T = \alpha \pi_k^T H + \pi_k^T \alpha w \mathbf{e}^T \frac{1}{n} + (1 - \alpha)\mathbf{e}^T \frac{1}{n}$$

- In reality we do not need to store matrix G.
- Although G is dense, H is very sparse (typically, about ten links per page).
- The values $\alpha w \mathbf{e}^T(1/n)$ and $(1 - \alpha)\mathbf{e}^T(1/n)$ are constant.
- The final number of arithmetic operations required is much smaller than what appears from the definition of G.