Master of Data Science Online Programme
Course: Linear Algebra
SGA #2: Week 6

Student: Andrei Batyrov (Fall 2022 / Winter 2023)
arbatyrov@edu.hse.ru

Higher School of Economics (HSE)
National Research University
Faculty of Computer Science

February 16, 2023

Contents

List of Figures

Problem 3

Use PageRank to find the most influenced vertex of the directed graph defined by the adjacency matrix

$$M = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 \end{pmatrix}. \tag{1}$$

The probabilities are equal. Note that the adjacency matrix is not the matrix that was considered in the videos.

Solution

This problem can be solved with the Random walk on graphs approach [4]. First let's represent (1) as a directed graph. Let $Pages = \{ A, B, C, D \}$ be a set of pages, $|Pages| = 4$, corresponding to the number of columns of (1). Let $Links$ be a set of links between the pages. Then the random walk graph can be described as follows: $\mathcal{G}\langle V(\mathcal{G}) = Pages, E(\mathcal{G}) = Links \rangle$. A vertex in j-th column has a directed edge to a vertex in i-th row, if $a_{ij} = 1$, and does not if $a_{ij} = 0$. For example, there is a link (edge) from page (vertex) A to page (vertex) C, since $m_{31} = 1$. The random graph is shown in Figure 1.

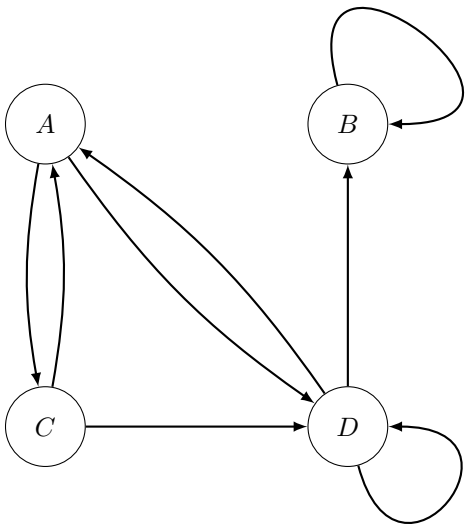

Figure 1: Random walk graph

Now we can construct the Markov transition matrix P of the same size as (1), where the sum of all elements along any column is equal to 1, i.e. each element in each column is the probability to walk along an edge (visit the link) from one vertex (page) to another. For example, for page A there are equal probabilities to visit either page C or page D, i.e. $P(A \to C) = P(A \to D) = 1/\sum_{i=1}^{4} a_{i1} = 1/(0+0+1+1) = 1/2$. All possible directed edges with their probabilities are listed below:

1. $P(A \to C) = 1/\sum_{i=1}^{4} a_{i1} = 1/(0+0+1+1) = 1/2$
2. $P(A \to D) = 1/\sum_{i=1}^{4} a_{i1} = 1/(0+0+1+1) = 1/2$
3. $P(B \to B) = 1/\sum_{i=1}^{4} a_{i2} = 1/(0+1+0+0) = 1$
4. $P(C \to A) = 1/\sum_{i=1}^{4} a_{i3} = 1/(1+0+0+1) = 1/2$
5. $P(C \to D) = 1/\sum_{i=1}^{4} a_{i3} = 1/(1+0+0+1) = 1/2$
6. $P(D \to A) = 1/\sum_{i=1}^{4} a_{i4} = 1/(1+1+0+1) = 1/3$

7. $P(D \to B) = 1/\sum_{i=1}^{4} a_{i4} = 1/(1+1+0+1) = 1/3$
8. $P(D \to D) = 1/\sum_{i=1}^{4} a_{i4} = 1/(1+1+0+1) = 1/3$.

So, our Markov transition matrix is

$$P = \begin{pmatrix} 0 & 0 & \frac{1}{2} & \frac{1}{3} \\ 0 & 1 & 0 & \frac{1}{3} \\ \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{3} \end{pmatrix}. \tag{2}$$

The Markov process allows to find the current state of the system by knowing the previous state. The PageRank algorithm employs the Markov process to find the stationary state of the system, in this case, the eigenvector g of probabilities of visiting each vertex (page) of the random walk graph while infinitely walking along the edges of the graph, such that

$$g = \lim_{k \to \infty} P_\alpha^k X_0, \tag{3}$$

where

- $P_\alpha = (1-\alpha)P + \alpha Q$,
- P is our transition matrix (2),
- α is the factor of "falling out" of the Markov process, i.e. it is the probability that instead of walking along the graph, a user can transit to some random page with the equal probability of $1/4$, since $|\text{Pages}| = 4$,
- Q is a matrix of size of P with all elements equal to the equal probability of $1/4$,
- k is the number (index) of the discrete state of the system, and
- X_0 is the vector of the initial state of the system. Let the initial vector X_0 be a vector of equal probabilities of having visited each of the four vertices (pages), i.e. $X_0 = \begin{pmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{pmatrix}$.

In this solution let's study the case when a user is not randomly directed from one page to another with the probability of α. For other scenarios see the section Further Considerations below. So, we have a pure Markov process, $\alpha = 0$, $P_\alpha = P$, and the matrix Q is not needed. As an example, let's find the state X_1. As per (3), $X_1 = \lim_{k \to 1} P^k X_0 = PX_0 = \begin{pmatrix} 0 & 0 & 1/2 & 1/3 \\ 0 & 1 & 0 & 1/3 \\ 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 1/3 \end{pmatrix} \begin{pmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{pmatrix} = \begin{pmatrix} 5/24 \\ 1/3 \\ 1/8 \\ 1/3 \end{pmatrix} = \begin{pmatrix} 0.208(3) \\ 0.(3) \\ 0.125 \\ 0.(3) \end{pmatrix}$. We can see that the probabilities of transiting to vertices (pages) B and D are greater than the same probabilities for A and C. Eventually, we need to find (3) for a theoretically infinite number of walks along the edges (page visits). To find $\lim_{k \to \infty} P^k$, as per [3], we need to decompose P^k as a product of three matrices:

$$P^k = TD^kT^{-1}, \tag{4}$$

where T is the transition matrix of eigenvectors of P, and D is the diagonal matrix in the basis of eigenvectors of P. Let's find this decomposition with the following algorithm.

1. Find eigenvalues λ_i of (2). As per [1], eigenvalues of a matrix are the roots of the characteristic polynomial, i.e.
$$det(P - \lambda I) = 0, \tag{5}$$

where *det* is a function of columns of a square matrix. So, our characteristic equation is

$$\begin{aligned} det\left[\begin{pmatrix} 0 & 0 & 1/2 & 1/3 \\ 0 & 1 & 0 & 1/3 \\ 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 1/3 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}\right] = 0, \\ det\left[\begin{pmatrix} 0 & 0 & 1/2 & 1/3 \\ 0 & 1 & 0 & 1/3 \\ 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 1/3 \end{pmatrix} - \begin{pmatrix} \lambda & 0 & 0 & 0 \\ 0 & \lambda & 0 & 0 \\ 0 & 0 & \lambda & 0 \\ 0 & 0 & 0 & \lambda \end{pmatrix}\right] = 0, \\ det\left[\begin{pmatrix} 0-\lambda & 0 & 1/2 & 1/3 \\ 0 & 1-\lambda & 0 & 1/3 \\ 1/2 & 0 & 0-\lambda & 0 \\ 1/2 & 0 & 1/2 & 1/3-\lambda \end{pmatrix}\right] = 0, \\ \lambda(\lambda-1)(2\lambda+1)(6\lambda-5) = 0. \end{aligned} \tag{6}$$

The equation (4) has four real roots: $\lambda_1 = 0, \lambda_2 = 1, \lambda_3 = -1/2, \lambda_4 = 5/6$. These are the eigenvalues of (2). And the matrix D is

$$D = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 \\ 0 & 0 & 0 & \lambda_4 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1/2 & 0 \\ 0 & 0 & 0 & 5/6 \end{pmatrix}. \tag{7}$$

2. Find eigenvectors of (2), As per [2], each eigenvalue has a corresponding eigenvector, which is a solution vector of the linear system $Pv_i = \lambda_i v_i \Rightarrow (P - \lambda_i I)v_i = 0$. Since we have four eigenvalues, there are four eigenvectors, each of can be obtained by solving its linear system. Let's find v_1 corresponding to λ_1:

$$\left[\begin{pmatrix} 0 & 0 & 1/2 & 1/3 \\ 0 & 1 & 0 & 1/3 \\ 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 1/3 \end{pmatrix} - \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}\right] v_1 = \begin{pmatrix} 0 & 0 & 1/2 & 1/3 \\ 0 & 1 & 0 & 1/3 \\ 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 1/3 \end{pmatrix} v_1 = 0. \tag{8}$$

The solution vector is $v_1 = x_4 \begin{pmatrix} 0 \\ -1/3 \\ -2/3 \\ 1 \end{pmatrix}$. By choosing $x_4 = 1, v_1 = \begin{pmatrix} 0 \\ -1/3 \\ -2/3 \\ 1 \end{pmatrix}$. By constructing and solving similar systems for $\lambda_2 = 1, \lambda_3 = -1/2, \lambda_4 = 5/6$, we get $v_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, v_3 = \begin{pmatrix} -1 \\ 0 \\ 1 \\ 0 \end{pmatrix}, v_4 = \begin{pmatrix} 5/8 \\ -2 \\ 3/8 \\ 1 \end{pmatrix}$. And the matrix T is

$$T = \begin{pmatrix} 0 & 0 & -1 & 5/8 \\ -1/3 & 1 & 0 & -2 \\ -2/3 & 0 & 1 & 3/8 \\ 1 & 0 & 0 & 1 \end{pmatrix}. \tag{9}$$

Now by applying (3) and (4), we get $g = \lim_{k\to\infty} P^k X_0 = \lim_{k\to\infty} TD^kT^{-1}X_0 = T\left[\lim_{k\to\infty} D^k\right] T^{-1}X_0$. It is easily seen that $\lim_{k\to\infty} D^k$ will have zeros for elements $|d_{ij}| < 1$, so $D^k = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$.

Finally,

$$
\begin{aligned}
g = TD^kT^{-1}X_0 &= \begin{pmatrix} 0 & 0 & -1 & 5/8 \\ -1/3 & 1 & 0 & -2 \\ -2/3 & 0 & 1 & 3/8 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & -1 & 5/8 \\ -1/3 & 1 & 0 & -2 \\ -2/3 & 0 & 1 & 3/8 \\ 1 & 0 & 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{pmatrix} \\
&= \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & -1 & 5/8 \\ -1/3 & 1 & 0 & -2 \\ -2/3 & 0 & 1 & 3/8 \\ 1 & 0 & 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{pmatrix} \\
&= \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} -0.6 & 0 & -0.6 & 0.6 \\ 1 & 1 & 1 & 1 \\ -0.625 & 0 & 0.375 & 0.25 \\ 0.6 & 0 & 0.6 & 0.4 \end{pmatrix} \begin{pmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{pmatrix} \\
&= \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{pmatrix} \\
&= \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}.
\end{aligned} \tag{10}
$$

So, our stationary state vector g shows that after infinite walk (visiting pages), page B has the highest probability of attracting visitors, i.e. it is the most influenced vertex in the graph.

Further Considerations

We have studied the case for $\alpha = 0$, i.e. there is no "punishment" for any page to be redirected to. In this case the page B attracts the most visits. Let's also study other cases for different α. To facilitate computation, a Python code was developed to simulate the PageRank algorithm. A sample code can be found, for example, on Wikipedia: https://en.wikipedia.org/wiki/PageRank#Python.

In Figure 2 below there is a plot of the vector g versus different values of α. The recommended by Google $\alpha = 0.15$ is also shown. It is clearly seen that even for larger values of α, page B still

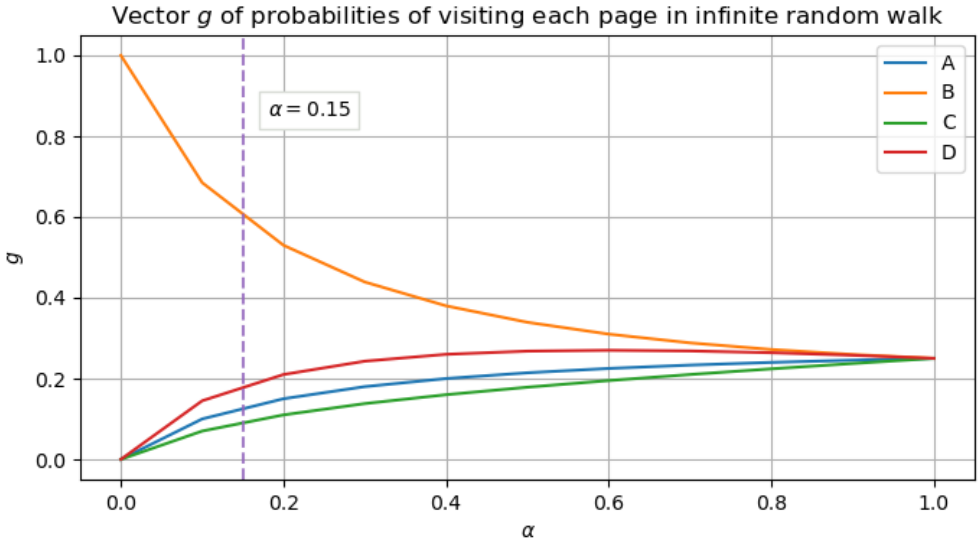

Figure 2: Vector g of probabilities of visiting each page in infinite random walk

attracts most visits. And only in the limit case of $\alpha = 1$ all four pages have the equal probability to be visited, which is expected, since in this case there is no Markov process in fact, and the vector g is determined by the matrix Q of equal probabilities of $1/4$, as per (3).

Answer

The most influenced vertex in the random walk directed graph is B. It means that this page should be ranked higher in the search results. However, this is dangerous, since page B has no outgoing links (edges), and eventually all users will end up on page B.

References

[1] Vsevolod Chernyshev. *Characteristic polynomial*. Faculty of Computer Science, Higher School of Economics. URL: https://smartedu.hse.ru/mod/page/0/701025.

[2] Vsevolod Chernyshev. *Eigenvectors and eigenvalues*. Faculty of Computer Science, Higher School of Economics. URL: https://smartedu.hse.ru/mod/page/0/701022.

[3] Vsevolod Chernyshev. *Matrix diagonalization*. Faculty of Computer Science, Higher School of Economics. URL: https://smartedu.hse.ru/mod/page/0/701028.

[4] Vsevolod Chernyshev. *Random walk on a graph*. Faculty of Computer Science, Higher School of Economics. URL: https://smartedu.hse.ru/mod/page/0/701030.