HOMEWORK 7

EXERCISE ON PAGERANK

Submitted by

Arun Rajagopalan

A20360689

For CS422 Data Mining

Illinois Institute of Technology

EXERCISE ON PAGERANK

Use the PageRank approach to find influential Twitter users.

Over this Twitter-User graph, apply the PageRank approach to rank the users. The main idea is that a user who is mentioned by other users is more influential.
Calculate the PageRank for a selection of four users based on the following four tweets:

user: Tim, tweet: "@Tom Howdy!"
user: Mike, tweet: "Welcome @Tom and @Anne!"
user: Tom, tweet: "Hi @Mike and @Anne!"
user: Anne, tweet: "Howdy!"

There are four short tweets generated by four users. The @mentions between users form a directed graph with four nodes and five edges. E.g., the "Tim" node has a directed edge to the "Tom" node.
Compute manually the first 3 iterations of the PageRank iterations over this 4 node graph. You should use 0.1 as the probability of teleporting. Show all steps of your calculation, provide details and explanations for them. Write down the rank order of the 4 users after on you compute 3 iterations.

Organizing the web has always been a challenge and many techniques are being explored and employed from the Internet inception till date to get good use of the web in a relatively shorter span of time. Had there been no organization, we would spend a whole lot of time searching for good websites but mostly ending up in bad, poor quality pages. The reason is that there has been a explosion of content on web, most of them being insignificant and hence intelligent organization becomes very important.

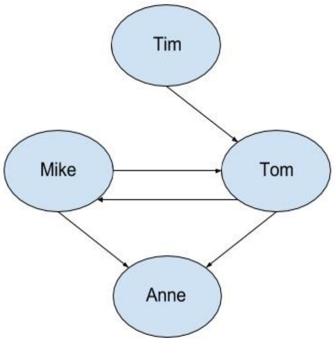
Using with text and human curated directories for organization is not possible keeping in mind the multitude of web sites that we have today. Also methods that use text in web pages to determine the content can be easily fooled. Hence, to intelligently organize web, Page rank was born.

Page Rank is an algorithm that ranks web pages based on their popularity. Popularity is measured by number of pages that link to a particular page and their popularity. If a web page has inlinks from all popular pages, then that page must be popular. This is the basic concept of

Page Rank. Technically put, it is the measure of where a random surfer using the web would end up at a particular time. The surfer will be hopping on to links found in a page. Thus there are two main things that need to be noted.

One is that there should not be dead ends (a node that has no outlinks) and there should be no traps (nodes that link to themselves) in the structure that Page Rank will evaluate or else we will not get the desired ranking. We cannot expect web to be so and hence we introduce modifications to Page Rank to tackle with traps and dead ends. We can understand the algorithm more by analyzing the given problem.

For the current problem, the connections between the nodes can be depicted as a directed graph as below.



The transaction matrix M for the given graph is as below.

This matrix M has n rows and columns, if there are n pages. The element Mij in row i and column j has value 1/k if page j has k arcs out, and one of them is to page i. Otherwise Mij is zero.

$$\mathit{Transaction}:\begin{bmatrix}Tim & Tom & Mike & Anne\end{bmatrix}$$

$$\begin{bmatrix}Tim\\Tom\\Mike\\Anne\end{bmatrix}=\begin{bmatrix}0&0&0&0\\1&0&0.5&0\\0&0.5&0&0\\0&0.5&0.5&0\end{bmatrix}$$

Initially, we give equal ranks to all pages.

Hence our initial page rank vector is,

$$v\,0=\begin{bmatrix}0.25\\0.25\\0.25\\0.25\end{bmatrix}$$

We know that, the rank vector after a specified number of iterations, v' is given by

$$v' = \beta*M*v + (1-\beta)*(E/N)$$

where β is the probability with which links in a page are followed and 1- β is the probability with which the surfer jumps to a random page. This jumping is called transportation and it helps in maintaining useful page ranks which might otherwise get completely leaked out or accumulated at a single page during the presence of dead ends and traps. E is a vector of all 1s and N is the number of nodes.

β*M*v is the usual page rank and the (1-β)*(E/N) part is the taxation that is used to save page ranks from dead ends and traps.

In the given problem, Anne is a dead end since it does not give any outlink. To tackle this, we use a transportation probability of 0.1 as instructed. The iterations of page rank calculation are as follows.

First iteration:

v0 is the initial page rank vector with equal importance for all pages. The vector with all 0.025 is the taxation part, the product of 1- β which is 0.1 and the (E/N) vector which is a vector with all values ¼.

$$\beta * M * v0 = 0.9 * \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0.5 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0.5 & 0.5 & 0 \end{bmatrix} * \begin{bmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{bmatrix}$$

$$\beta * M * v0 = \begin{bmatrix} 0 \\ 0.3375 \\ 0.1125 \\ 0.225 \end{bmatrix}$$

$$v1 = \begin{bmatrix} 0 \\ 0.3375 \\ 0.1125 \\ 0.225 \end{bmatrix} + \begin{bmatrix} 0.025 \\ 0.025 \\ 0.025 \\ 0.025 \end{bmatrix} = \begin{bmatrix} 0.025 \\ 0.3625 \\ 0.1375 \\ 0.25 \end{bmatrix}$$

$$v1N = \begin{bmatrix} 0.025/0.775 \\ 0.3625/0.775 \\ 0.1375/0.775 \\ 0.25/0.775 \end{bmatrix} = \begin{bmatrix} 0.032 \\ 0.468 \\ 0.177 \\ 0.323 \end{bmatrix}$$

Here v1N is the normalized page vector that sums up to 1. Due to the presence of dead node, the columns of matrix M are not perfectly stochastic (column summing up to 1), due to which we introduce 1-beta taxation and do this normalization at the end. Normalization is done by dividing all elements in a vector by the sum of all elements in the vector.

Second Iteration:

$$\beta * M * v1 = 0.9 * \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0.5 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0.5 & 0.5 & 0 \end{bmatrix} * \begin{bmatrix} 0.032 \\ 0.468 \\ 0.177 \\ 0.323 \end{bmatrix}$$

$$\beta * M * v1 = \begin{bmatrix} 0 \\ 0.109 \\ 0.211 \\ 0.291 \end{bmatrix}$$

$$v2 = \begin{bmatrix} 0 \\ 0.109 \\ 0.211 \\ 0.291 \end{bmatrix} + \begin{bmatrix} 0.025 \\ 0.025 \\ 0.025 \\ 0.025 \end{bmatrix} = \begin{bmatrix} 0.025 \\ 0.134 \\ 0.236 \\ 0.316 \end{bmatrix}$$

$$v2N = \begin{bmatrix} 0.025/0.711 \\ 0.134/0.711 \\ 0.236/0.711 \\ 0.316/0.711 \end{bmatrix} = \begin{bmatrix} 0.035 \\ 0.188 \\ 0.332 \\ 0.445 \end{bmatrix}$$

Third Iteration:

$$\beta * M * v2 = 0.9 * \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0.5 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0.5 & 0.5 & 0 \end{bmatrix} * \begin{bmatrix} 0.035 \\ 0.188 \\ 0.332 \\ 0.445 \end{bmatrix}$$

$$\beta * M * v2 = \begin{bmatrix} 0 \\ 0.181 \\ 0.085 \\ 0.234 \end{bmatrix}$$

$$v3 = \begin{bmatrix} 0 \\ 0.181 \\ 0.085 \\ 0.234 \end{bmatrix} + \begin{bmatrix} 0.025 \\ 0.025 \\ 0.025 \\ 0.025 \end{bmatrix} = \begin{bmatrix} 0.025 \\ 0.206 \\ 0.11 \\ 0.259 \end{bmatrix}$$

$$v3N = \begin{bmatrix} 0.025/0.6 \\ 0.206/0.6 \\ 0.11/0.6 \\ 0.259/0.6 \end{bmatrix} = \begin{bmatrix} 0.042 \\ 0.343 \\ 0.183 \\ 0.432 \end{bmatrix}$$

Thus at the end of 3 iterations, the ranks of Anne is the highest with 0.432, followed by Tom with 0.343, then by Mike with 0.183 and finally ending with Tim who has a rank of 0.042.