CS 422

DATA MINING

HOME WORK – 6

SUBMITTED BY :

SACHIN KRISHNA MURTHY

CWID : A20354077

Question :

Use the PageRank approach to find influential Twitter users.

PageRank graph is constructed from web pages with hyperlinks. Pages are nodes, and hyperlinks are edges. For this problem, use the graph of Twitter users and their mentions of other Twitter users. Users are nodes, mention of another users are edges.
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.

Solution:

PageRank: PageRank is a function that assigns a real number to each page in the Web or at least to that portion of the Web that has been crawled and its links discovered. Also the higher the PageRank of a page, the more important it is. It uses the concept of in-links (the links where other pages points to a page) and out-links (the links where a page points toward other pages).

Teleporting: The solution for Spider Traps is that at each step the random surfer has two options :

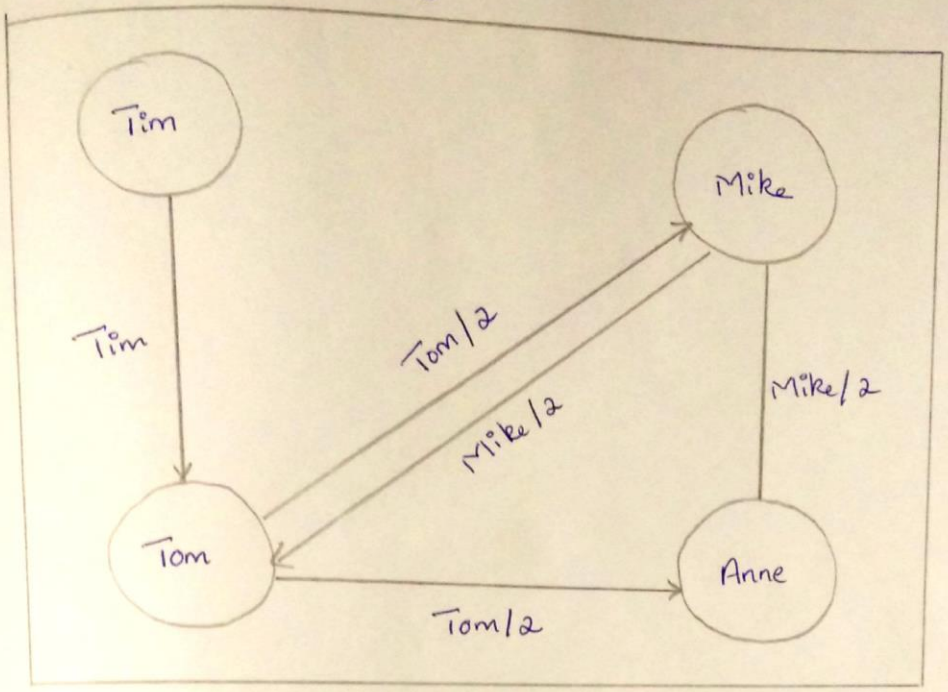
- With probability β, follow a link at random
- With probability 1-β, jump to some page uniformly at random

The surfer will teleport out of spider trap within a few steps.

Steps Accomplished:

- In the first step the matrix is obtained as shown in the calculations below.
- This matrix is used for the further computation of the PageRank.
- In Iteration 1, we are finding the product of the matrix obtained with 1/n vector (V_0).
- In Iteration 2, we are calculating the product of the matrix obtained with the value obtained in iteration 1.
- In Iteration 3, we are calculating the product of the matrix obtained with the value of iteration 2.

consider the below figure:



Rank of a page is given by: $r_j = \sum_{i \to j} \frac{r_i}{d_i}$

d_i = out degree of node i

$R(Tim) = 0$

$R(Mike) = R(Tom)/2$

$R(Tom) = R(Tim) + R(Mike)/2$

$R(Anne) = R(Tom)/2 + R(Mike)/2$

Matrix can be written as:

	Tim	Mike	Tom	Anne
Tim	0	0	0	0
Mike	0	0	1/2	0
Tom	1	1/2	0	0
Anne	0	1/2	1/2	0

Probability of Teleporting = 0.1 = $(1-\beta)$

Therefore, the Page Rank equation is :

$$r_j = \sum_{i \to j} \beta \frac{r_i}{d_i} + (1-\beta)\frac{1}{n}$$

d_i = out degree of node i

$1-\beta = 0.1$

$\therefore \beta = 1-0.1$

$\beta = 0.9$

Step 1 : obtaining the Matrix

$$= 0.9 \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 0 \\ 1 & 1/2 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 \end{bmatrix} + 0.1 \begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 9/20 & 0 \\ 9/10 & 9/20 & 0 & 0 \\ 0 & 9/20 & 9/20 & 0 \end{bmatrix} + \begin{bmatrix} 1/40 & 1/40 & 1/40 & 1/40 \\ 1/40 & 1/40 & 1/40 & 1/40 \\ 1/40 & 1/40 & 1/40 & 1/40 \\ 1/40 & 1/40 & 1/40 & 1/40 \end{bmatrix}$$

$$= \begin{bmatrix} 1/40 & 1/40 & 1/40 & 1/40 \\ 1/40 & 1/40 & 19/40 & 1/40 \\ 37/40 & 19/40 & 1/40 & 1/40 \\ 1/40 & 19/40 & 19/40 & 1/40 \end{bmatrix}$$

Iteration 1 :-

$$\begin{bmatrix} 1/40 & 1/40 & 1/40 & 1/40 \\ 1/40 & 1/40 & 19/40 & 1/40 \\ 37/40 & 19/40 & 1/40 & 1/40 \\ 1/40 & 19/40 & 19/40 & 1/40 \end{bmatrix} \times \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}$$

$$= \begin{bmatrix} 1/160 + 1/160 + 1/160 + 1/160 \\ 1/160 + 1/160 + 19/160 + 1/160 \\ 37/160 + 19/160 + 1/160 + 1/160 \\ 1/160 + 19/160 + 19/160 + 1/160 \end{bmatrix}$$

$$= \begin{bmatrix} 4/160 \\ 22/160 \\ 58/160 \\ 40/160 \end{bmatrix} = \begin{bmatrix} 0.025 \\ 0.1375 \\ 0.3625 \\ 0.25 \end{bmatrix}$$

Iteration 2 :-

$$\begin{bmatrix} 1/40 & 1/40 & 1/40 & 1/40 \\ 1/40 & 1/40 & 19/40 & 1/40 \\ 37/40 & 19/40 & 1/40 & 1/40 \\ 1/40 & 19/40 & 19/40 & 1/40 \end{bmatrix} \times \begin{bmatrix} 4/160 \\ 22/160 \\ 58/160 \\ 40/160 \end{bmatrix}$$

$$= \begin{bmatrix} 1/1600 + 11/3200 + 29/3200 + 1/160 \\ 1/1600 + 11/3200 + 551/3200 + 1/160 \\ 37/1600 + 209/3200 + 29/3200 + 1/160 \\ 1/1600 + 209/3200 + 551/3200 + 1/160 \end{bmatrix}$$

$$= \begin{bmatrix} 62/3200 \\ 584/3200 \\ 332/3200 \\ 782/3200 \end{bmatrix} = \begin{bmatrix} 0.0193 \\ 0.1825 \\ 0.1037 \\ 0.2443 \end{bmatrix}$$

Iteration 3 :-

$$\begin{bmatrix} 1/40 & 1/40 & 1/40 & 1/40 \\ 1/40 & 1/40 & 19/40 & 1/40 \\ 37/40 & 19/40 & 1/40 & 1/40 \\ 1/40 & 19/40 & 19/40 & 1/40 \end{bmatrix} \times \begin{bmatrix} 62/3200 \\ 584/3200 \\ 332/3200 \\ 782/3200 \end{bmatrix}$$

$$\begin{bmatrix} 31/64000 + 292/64000 + 166/64000 + 391/64000 \\ 31/64000 + 292/64000 + 3154/64000 + 391/64000 \\ 1147/64000 + 5548/64000 + 166/64000 + 391/64000 \\ 31/64000 + 5548/64000 + 3154/64000 + 391/64000 \end{bmatrix}$$

$$= \begin{bmatrix} 880/64000 \\ 3868/64000 \\ 72252/64000 \\ 9124/64000 \end{bmatrix} = \begin{bmatrix} 0.0137 \\ 0.0604 \\ 0.1133 \\ 0.1425 \end{bmatrix}$$

Table showing the values of Iteration 1,2 & 3 for all the users :

Users	Iteration 1	Iteration 2	Iteration 3
Tim	0.025	0.0193	0.0137
Mike	0.1375	0.1825	0.0604
Tom	0.3625	0.1037	0.1133
Anne	0.25	0.2443	0.1425

Conclusion :

The higher the PageRank of a page, the more important it is.

After 3rd iteration from the above table we can notice that, the order of Ranking is :

Anne > Tom > Mike > Tim

Therefore,

Rank	User
1	Anne
2	Tom
3	Mike
4	Tim