Seoul National University

M1522.001400 Introduction to Data Mining

Spring 2016, Kang

Homework 4: Link Analysis (Chapter 5)

Due: May 2, 09:30 AM

Reminders

- The points of this homework add up to 100.

- Like all homeworks, this has to be done individually.

- Lead T.A.: Minsoo Jung (qtyp456987@gmail.com)

- Please type your answers *in English*. Illegible handwriting may get no points, at the discretion of the graders.

- If you have a question about assignments, please upload your question in eTL.

- If you want to use slipdays or consider late submission with penalties, please note that you are allowed one week to submit your assignment after the due date. That is, after May 9, we will NOT receive your submission for this assignment.

Remember that:

- Whenever you are making an assumption, please state it clearly.

Question 1

Compute the PageRank of each page in Figure 1, assuming parameter $\beta = 1$ and error tolerance $\varepsilon = 10^{-6}$. [15 points]
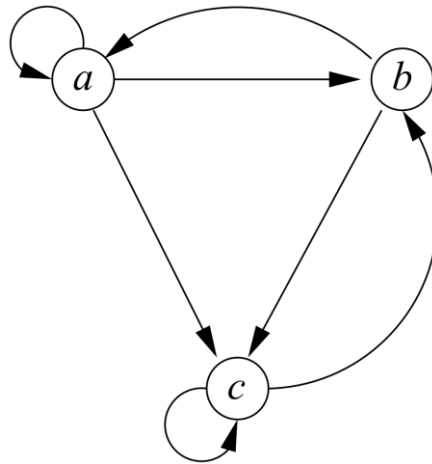

Figure 1: An example Web graph for Question 1 and 2

$$\text{PageRank } \mathbf{r} = \begin{bmatrix} 0.230769 \\ 0.307692 \\ 0.461538 \end{bmatrix}$$

Question 2

Compute the PageRank of each page in Figure 1, assuming parameter $\beta = 0.8$ and error tolerance $\varepsilon = 10^{-6}$. [15 points]

$$\text{PageRank } \mathbf{r} = \begin{bmatrix} 0.259259 \\ 0.308641 \\ 0.432098 \end{bmatrix}$$

Question 3

Compute Topic-specific PageRank for the graph of Figure 2. Assume parameter $\beta = 0.8$ and error tolerance $\varepsilon = 10^{-6}$. The teleport set is:
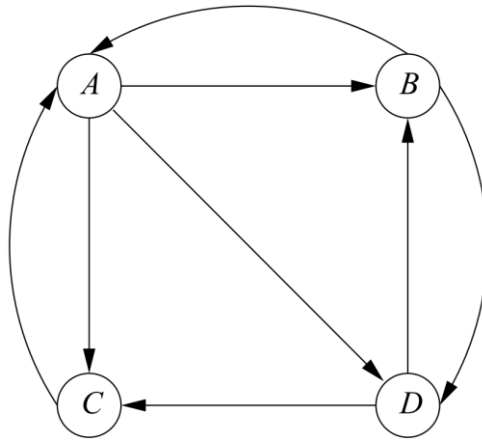

Figure 2: An example Web graph for Question 3

(a) A only. [10 points]

$$\text{PageRank } \mathbf{r} = \begin{bmatrix} 0.428571 \\ 0.190476 \\ 0.190476 \\ 0.190476 \end{bmatrix}$$

(b) A and C. [10 points]

$$\text{PageRank } \mathbf{r} = \begin{bmatrix} 0.385714 \\ 0.171428 \\ 0.271428 \\ 0.171428 \end{bmatrix}$$

Question 4

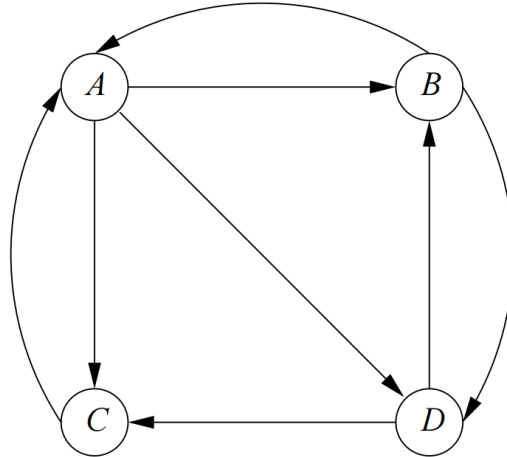For the Web graph of Figure 3, assuming only B is a trusted page:

Figure 3: An example Web graph for Question 4 and 5

(a) Compute the TrustRank of each page. Suppose parameter $\beta = 0.8$ and error tolerance $\varepsilon = 10^{-6}$. [15 points]

$$\text{TrustRank } \mathbf{r} = \begin{bmatrix} 0.269387 \\ 0.357823 \\ 0.157823 \\ 0.214966 \end{bmatrix}$$

(b) Compute the spam mass of each page. [15 points]

You can get 15 points if you state one of the 2 solutions as follows:

1) PageRank without teleports is used.

$$\text{Spam mass} = \begin{bmatrix} 0.191836 \\ -0.61204 \\ 0.289795 \\ 0.032653 \end{bmatrix}$$

2) PageRank with teleports ($\beta = 0.8$) is used.

$$\text{Spam mass} = \begin{bmatrix} 0.161906 \\ -0.581955 \\ 0.302258 \\ 0.049626 \end{bmatrix}$$

Question 5

Compute the hub score and authority score of each node in the Web graph of Figure 3, using HITS algorithm where error tolerance $\varepsilon = 10^{-6}$. [20 points]

You can get 20 points if you state one of the 3 solutions as follows:

    1) Scale the values of the vectors **h** and **a** in order that the largest component is 1.

Hub score $\mathbf{h} = \begin{bmatrix} 1.0 \\ 0.391944 \\ 0.102775 \\ 0.710831 \end{bmatrix}$

Authority score $\mathbf{a} = \begin{bmatrix} 0.289169 \\ 1.0 \\ 1.0 \\ 0.813607 \end{bmatrix}$

    2) Scale in order that the sum of components is 1.

Hub score $\mathbf{h} = \begin{bmatrix} 0.453401 \\ 0.177708 \\ 0.046598 \\ 0.322291 \end{bmatrix}$

Authority score $\mathbf{a} = \begin{bmatrix} 0.093197 \\ 0.322291 \\ 0.322291 \\ 0.262219 \end{bmatrix}$

    3) Scale in order that the square sum of components is 1.

Hub score $\mathbf{h} = \begin{bmatrix} 0.773947 \\ 0.303343 \\ 0.079542 \\ 0.550146 \end{bmatrix}$

Authority score $\mathbf{a} = \begin{bmatrix} 0.174515 \\ 0.603508 \\ 0.603508 \\ 0.491018 \end{bmatrix}$