CIS 833 – Information Retrieval and Text Mining Name:_____

Homework assignment 4 – due November 11th

Please remember that you are allowed to discuss the assigned exercises, but you should write your own solution. Identical solutions will receive 0 points.

Exercise 1. (Breadth-first spider)

Consider the following web pages and the set of web pages that they link to:

Page A points to pages B, C, and D.
Page B points to pages E and F.
Page C points to page G.
Page D points to page B.
Page E points to pages C and D.
Page G points to page E.

Show the order in which the pages are indexed when starting at page A and using a breadth-first spider (with duplicate page detection) as discussed in class. Assume links on a page are examined in the orders given above.

Solution:

Web page visit order	Status
A (initial)	Indexed
B (from A)	Indexed
C (from A)	Indexed
D (from A)	Indexed
E (from B)	Indexed
F (from B)	Indexed
G (from C)	Indexed
B (from D)	Already visited
C (from E)	Already visited
D (from E)	Already visited
E (from G)	Already visited

Indexing order in BFS is: A, B, C, D, E, F, G

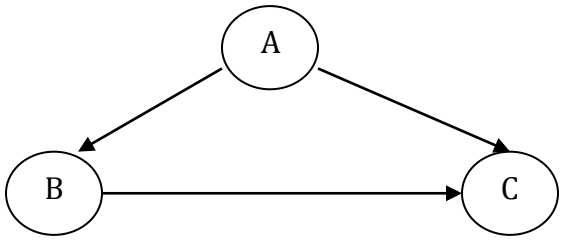
Exercise 2. (PageRank algorithm)

Consider the following web pages and the set of web pages that they link to:

Page A points to pages B and C.
Page B points to page C.

Consider running the MapReduce PageRank algorithm on this subgraph of pages. Assume α = 0.10.

Simulate the algorithm for three iterations. Show specifically what happens inside the map and reduce functions for each record processed.



Iteration 1:

Map input:

(A, [1/3, {B,C}]);

(B, [1/3, {C}]);

(C, [1/3, {}])

Map output:

(B, 1/6), (C, 1/6), (A, {B,C})

(C, 1/3), (B, {C})

(C, {})

Reduce input:

(A, [{B,C}])

(B, [1/6, {C}])

(C, [1/6, 1/3, { }])

In reduce, calculate:

R(A) = 0.1/3 = 1/30 = 0.033

R(B) = 0.1/3 + (1 – 0.1)* (1/6)= 11/60 = 0.183

R(C) = 0.1/3 + (1 – 0.1)*(1/6 + 1/3) = 29/60 = 0.483

Reduce output:

(A, [0.033, {B,C}])

(B, [0.183, {C}])

(C, [0.483, {}])

Next, normalize

c = 1/(0.033+0.183+0.483) ~ 1.42

After normalization: R(A) = 0.047; R(B) = 0.262; R(C) = 0.691

Iteration 2

Map input:

(A, [0.047, {B,C}])

(B, [0.262, {C}])

(C, [0.691, { }])

Map output:

(B, 0.047/2), (C, 0.047/2), (A, {B,C})

(C, 0.262/1), (B, {C})

(C, {})

Reduce input:

(A, [{B,C}])

(B, [0.047/2, {C}])

(C, [0.047/2, 0.262/1, { }])

In reduce, calculate:

R(A) = 0.1*/3 = 0.033

R(B) = 0.1*/3 + (1 – 0.1)* 0.047/2 = 0.054

R(C) = 0.1*/3 + (1 – 0.1)*(0.047/2 + 0.262/1) = 0.3

Reduce output:

(A, [0.033, {B,C}])

(B, [0.054, {C}])

(C, [0.3, { }])

Next, normalize

c = 1/(0.033+0.054+0.3) ~ 2.65

After normalization: R(A) = 0.089; R(B) = 0.143; R(C) = 0.769

Iteration 3:

(A, [0.089, {B,C}])

(B, [0.143, {C}])

(C, [0.769, { }])

Map output:

(B, 0.089/2), (C, 0.089/2), (A, {B,C})

(C, 0.143/1), (B, {C})

(C, {})

Reduce input:

(A, [{B,C}])

(B, [0.089/2, {C}])

(C, [0.089/2, 0.143/1, { }])

In reduce, calculate:

R(A) = 0.1*/3 = 0.033

R(B) = 0.1*/3 + (1 – 0.1)* (0.089/2) = 0.072

R(C) = 0.1*/3 + (1 – 0.1)*(0.089/2 + R(0.143/1) = 0.201

Reduce output:

(A, [0.033, {B,C}])

(B, [0.072, {C}])

(C, [0.201, { }])

Next, normalize:

c = 1/(0.033+0.072+0.201) = 1/0.3 = 3.26

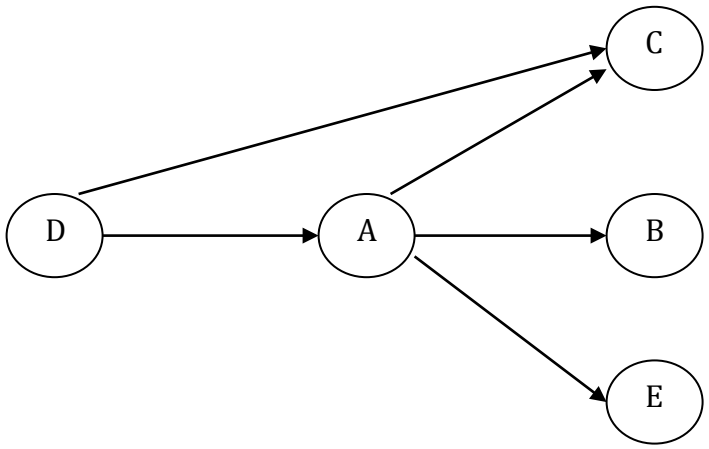
After normalization: R(A) = 0.107; R(B) = 0.235; R(C) = 0.656

Exercise 3. (HITS algorithm)

Consider the following web pages and the set of web pages that they link to:

Page A points to pages B, C, and E.
Page D points to pages A and C.

Consider running the HITS (Hubs and Authorities) algorithm on this subgraph of pages. Simulate the algorithm for three iterations. Show the authority and hub scores for each page twice for each iteration, both before and after normalization.



Solution:

A B C D E

Iteration 1:

A = [1, 1, 2, 0, 1]

H = [4, 0, 0, 3, 0]

$c_a = \text{sqrt}\,(1+1+4+0+1) = 2.65$

$c_h = \text{sqrt}\,(16+0+0+9+0) = 5$

Norm A = [0.38, 0.38, 0.76, 0.0, 0.38]

Norm H = [0.80, 0.0, 0.0, 0.60, 0.0]

Iteration 2:

A = [0.60, 0.80, 1.40, 0.0, 0.80]

H = [3.0, 0.0, 0.0, 2.0, 0.0]

$c_a = \text{sqrt}(0.6^2 + 0.8^2 + 1.4^2 + 0 + 0.8^2) = 1.9$

$c_h = \text{sqrt}\,(3^2+0+0+2^2+0) = 3.61$

Norm A = [0.32, 0.42, 0.74, 0.0, 0.42]

Norm H = [0.83, 0.0, 0.0, 0.55, 0.0]

Iteration 3:

A = [0.55, 0.83, 1.39, 0, 0.83]

H = [3.05, 0.0, 0.0, 1.94, 0.0]

$c_a = \text{sqrt}\,(0.56^2 + 0.83^2 + 1.39^2 + 0 + 0.83^2) = 1.9$

$c_h = \text{sqrt}\,(3.05^2+0+0+1.95^2+0) = 3.62$

Norm A = [0.29, 0.44, 0.73, 0.0, 0.44]

Norm H = [0.84, 0.0, 0.0, 0.54, 0.0]