2009 Final Answers

Charlie Zhang

Mar 2013

1 Question 1

2 Question 2

2.1 a)

(0.1, 0.9. 0.36, 0.16)-sensitive

2.2 b)

(0.1, 0.9. 0.936, 0.784)-sensitive

3 Question 3

3.1 a)

$sim(X, Y) = \frac{1}{4}$

$sim(X, Z) = \frac{1}{5}$

$sim(Y, Z) = \frac{2}{4}$

3.2 b)

Perm. X Y Z

1 b b a

2 b b b

3 d c c

4 d e a

5 b e a

3.3 c)

$est - sim(X, Y) = \frac{2}{5}$

$est - sim(X, Z) = \frac{1}{5}$

$est - sim(Y, Z) = \frac{4}{5}$

4 Question 4

4.1 a)

$s \leq n(n-1)$

TODO

4.2 b)

$s \leq n(n-1)/m$

TODO

5 Question 5

5.1 a)

{A}, {B}, {C}, {D}, {E}, {A, B}, {B, C}, {A, C},

5.2 b)

{A, D}, {A, E}, {B, D}, {B, E}, {C, D}, {C, E}

6 Question 6

6.1 a)

8

6.2 b)

4

6.3 c)

i=2, estimated distinct elements: 2

6.4 d)

$(3+i)\%2 \neq (8+i)\%2$, so whatever i is, at least one of $h_i(3), h_i(8)$ has last binary digit equals 0. So it is impossible to find i such that estimated distinct element is 1. Therefore, 2 is the minimum possible estimation.

7 Question 7

$X1 = \{'element' :' a','value' : 3\}$

$X2 = \{'element' :' a','value' : 2\}$

$X3 = \{'element' :' c','value' : 1\}$

8 Question 8

8.1 a)

Define clustroid of S as: $arg\min_{c \in S} \sum_{i \in S} dist(c, i)$.

Therefore, a is the clustroid as it's cost is $3+10+4 = 17$, which is the minimum among {a, b, c, d}.

8.2 b)

Define clustroid of S as: $arg\min_{c \in S} \max_{i \in S} dist(c, i)$.

Therefore, b is the clustroid as it's cost is $max(3, 7, 8) = 8$, which is the minimum among {a, b, c, d}. (d could also be clustroid in this case)

9 Question 9

TODO

10 Question 10

10.1 a)

TODO: home many chunks?

Disk I/O: Mapper Input(Read): 100 million records * (100 + 4) bytes per record = 10.4 GB

Mapper Output(Write): 100 million records * (20 + 4) bytes per record = 2.4 GB

Reducer Input(Read): Same as mapper output, 2.4 GB

Reducer Output(Write): 10 million unique hosts * (20 + 4) bytes per host * 2 replica = 480 MG

Total Dist I/O: 15.68 GB

Network I/O: Equals to Mapper output/ Reducer input, 2.4GB

10.2 b)

Disk I/O: Mapper Input(Read): 100 million records * (100 + 4) bytes per record = 10.4 GB

Mapper Output(Write): 100 million records * (20 + 4) bytes per record = 2.4 GB

Combiner Input(Read): Same as mapper output, 2.4 GB

Combiner Output(Write): 10 Chunk * 3 million unique host per chunk * (20 + 4) byte per host record = 720 MB

Reducer Input(Read): Same as combiner output, 720 MB

Reducer Output(Write): 10 million unique hosts * (20 + 4) bytes per host * 2 replica = 480 MB

Total Dist I/O: 17.12 GB

Network I/O: Equals to Combiner output/ Reducer input, 720 MB

11 Question 11

11.1 a)

For each node, the record is {nodeId, degree, [outlinks]} where each element takes 4 bytes.

Therefore, total size on disk = 100 million pages * (1 + 1 + 10) elements per page * 4 bytes per element = 4.8 GB

11.2 b)

Pagerank score of 100 million pages takes 400MB, so we can put both old PR score and new PR score into memory, it takes 800M in total.

Computation method:

1. initialize in memory all old PR scores as $1/N$

2. For each iteration: initialize all new PR scores

Read graph from disk one node at a time, and update new PR score vector in memory. Swap old PR score with the new ones. 3. After 50 iterations, write PR scores vector to disk

11.3 c)

Disk read: 4.8G * 50 iterations = 240 GB

Disk write: 400 MB of PR scores.

Total: 240.4 GB

12 Question 12

TODO

13 Question 13

Let x be the page rank of spam page, y be the page rank of target page

We have: $x = \beta x + \beta \frac{y}{M} + (1-\beta)\frac{1}{N}$

$x = \frac{\beta}{(1-\beta)M} y + \frac{1}{N}$

We also have: $y = \beta(M * x + \alpha) + (1-\beta)\frac{1}{N}$ Solve the equation and we get

$y = \frac{1-\beta}{1-\beta-\beta^2}\alpha + c * \frac{M}{N}$, where $c = \frac{\beta-\beta^2}{1-\beta-\beta^2}$

14 Question 14

Generalized Balance:

Bid = x_i

Budget = b_i

Amount spent = m_i

Fraction of budget left over $f_i = 1 - m_i/b_i$

Define $\psi(q) = x_i(1 - e^{-f_i})$

Here we use eCPM to replace bid, where eCPM = CTR * Bid

For A, $x_i = 0.1, f_i = 0.9, \psi(a) = 0.1 * (1 - e^{-}0.9) = 0.05934$

For B, $x_i = 0.16, f_i = 0.5, \psi(b) = 0.16 * (1 - e^{-}0.5) = 0.06296$

For B, $x_i = 0.12, f_i = 0.6, \psi(c) = 0.12 * (1 - e^{-}0.6) = 0.05414$

Thus, winner of A, B is B

Thus, winner of A, C is A

Thus, winner of A, B and C is B