COMP9313 2017s2 Assignment

Question 1. MapReduce (5 pts)

Assume that in an online shopping system, a huge log file stores the information of each transaction. Each line of the log is in format of "userID\tproduct\t price\t time". Your task is to use MapReduce to find out the top-5 most expensive products purchased by each user in 2016.

You only need to write down the pseudo code (Mapper, Reducer, and optionally Combiner and Partitioner) to describe the algorithm (assume only one reducer). Note that the efficiency and scalability of your solution will be evaluated.

Answer:

```
class Mapper
        initialize an associate array H(integer UserID, priority queue Q of log record
based on price)
        method Map(key, log record R)
                if R.time == 2016
                        H(R.userID).add(R)
                        if(H(R.userID).size >5)
                                H(R.userID).remove(first element)
         method CleanUp()
                  foreach(entry E in H)
                          Emit(E.userID, E.Q)
class Reducer
        method Reduce(userID, list of queues[])
                P <- get top 5 products from the list of queues
                Emit(userID, P)
```

A combiner class can accelerate the computation, which is the same as the reducer.

Question 2. MinHash (5 pts)

We want to compute min-hash signature for two columns, C_1 and C_2 using two pseudo-random permutations of columns using the following function:

$$h_1(n) = 3n + 2 \bmod 7$$

$$h_2(n) = 2n - 1 \bmod 7$$

Here, n is the row number in original ordering. Instead of explicitly reordering the columns for each hash function, we use the implementation discussed in class, in which we read each data in a column once in a

sequential order, and update the min hash signatures as we pass through them.

Complete the steps of the algorithm and give the resulting signatures for C_1 and C_2.

Row	C_1	C_2
0	0	1
1	1	0
2	0	1
3	0	0
4	1	1
5	1	1
6	1	0

Answer:

Row	C1	C2	$h_1(n)$	$h_2(n)$
0	0	1	2	6
1	1	0	5	1
2	0	1	1	3
3	0	0	4	5
4	1	1	0	0
5	1	1	3	2
6	1	0	6	4

Initialize:

	SigC1	SigC2
h_1	∞	∞
h_2	∞	∞

Row 0:

	SigC1	SigC2
h_1	∞	2
h_2	∞	6

Row 1:

	SigC1	SigC2
h_1	5	2
h_2	1	6

Row 2:

	SigC1	SigC2
h_1	5	1
h_2	1	3

Row 3:

	SigC1	SigC2
h_1	5	1
h_2	1	3

Row 4 (since all 0 entries, you can early stop here):

	SigC1	SigC2
h_1	0	0
h_2	0	0

Row 5:

	SigC1	SigC2
h_1	0	0
h_2	0	0

Row 6 (the resulting signatures for C1 and C2):

	SigC1	SigC2
h_1	0	0
h_2	0	0

Question 3. Streaming Data (5 pts)

Suppose we are maintaining a count of 1s using the DGIM method. We represent a bucket by (i, t), where i is the number of 1s in the bucket and t is the bucket timestamp (time of the most recent 1).

Consider that the current time is 200, window size is 60, and the current list of buckets is: (16, 148) (8, 162) (8, 177) (4, 183) (2, 192) (1, 197) (1, 200). At the next ten clocks, 201 through 210, the stream has 0101010101. What will the sequence of buckets be at the end of these ten inputs?

Answer:

There are 5 1s in the stream. Each one will update to windows to be:

(1) (16, 148)(8, 162)(8, 177)(4, 183)(2, 192)(1, 197)(1, 200), (1, 202)

=> (16, 148)(8, 162)(8, 177)(4, 183)(2, 192)(2, 200), (1, 202)

(2) (16, 148)(8, 162)(8, 177)(4, 183)(2, 192)(2, 200), (1, 202), (1, 204)

(3) (16, 148)(8, 162)(8, 177)(4, 183)(2, 192)(2, 200), (1, 202), (1, 204), (1; 206)

=> (16, 148)(8, 162)(8, 177)(4, 183)(2, 192)(2, 200), (2, 204), (1, 206)

=> (16, 148)(8, 162)(8, 177)(4, 183)(4, 200), (2, 204), (1, 206)

(4) Windows Size is 60, so (16,148) should be dropped.

(16, 148)(8, 162)(8, 177)(4, 183)(4, 200), (2, 204), (1, 206), (1, 208) => (8, 162)(8, 177)(4, 183)(4, 200), (2, 204), (1, 206), (1, 208)

(5) (8, 162)(8, 177)(4, 183)(4, 200), (2, 204), (1, 206), (1, 208), (1, 210)

=> (8, 162)(8, 177)(4, 183)(4, 200), (2, 204), (2, 208), (1, 210)

Question 4. Collaborative Filtering (5 pts)

Consider four users u1, u2, u3 and u4, and three movies m1, m2, and m3. The ratings of movies from the users are as below:

user	item	rating
u1	m1	2
u1	m3	3
u2	m1	5
u2	m2	2
u3	m1	3
u3	m2	3
u3	m3	1
u4	m2	2
u4	m3	2

(a) Estimate the rating of u1 to m2 using the user-user collaborative filtering method (adopt the cosine similarity measure to compute the user similarities).

Answer:

The Similarity Metric is:

User	m_1	m_2	m_3
u_1	2		3
u_2	5	2	
u_3	3	3	1
u_4		2	2

a) We compute similarities between u1 and other users by cosine similarity formula

(Note that: *Cosine similarity:* $sim(x, y) = \frac{\sum_i r_{xi} \cdot r_{yi}}{\sqrt{\sum_i r_{xi}^2} \cdot \sqrt{\sum_i r_{yi}^2}}$)

$\text{Sim}(u_1, u_2) = \frac{2\times5}{\sqrt{2^2+3^2}\sqrt{2^2+3^2}} \approx 0.515$

$\text{Sim}(u_1, u_3) = \frac{2\times3+3\times1}{\sqrt{2^2+3^2}\sqrt{3^2+3^2+1^2}} \approx 0.5727$

$\text{Sim}(u_1, u_4) = \frac{3\times2}{\sqrt{2^2+3^2}\sqrt{2^2+2^2}} \approx 0.5883$

The rating of u_1 to m_2 can be estimated as:

(Note that: $r_{ix} = \frac{\sum_{j\in N(i;x)} s_{ij} \cdot r_{jx}}{\sum s_{ij}}$,

Where s_{ij}… similarity of items ***i*** and ***j***, r_{xj}…rating of user ***u*** on item ***j***, ***N(i;x)***… set items rated by ***x*** similar to ***i***.)

$r_{u_1,m_2} = \frac{2\times0.515+3\times0.5727+2\times0.5883}{0.515+0.5727+0.5883} \approx 2.34$

(b) Estimate the rating of u1 to m2 using the item-item collaborative filtering method (adopt the cosine similarity measure to compute the item similarities).

Answer:

b) We compute similarities between m2 and other movies by cosine similarity formula

$$\text{Sim}(m_2, m_1) = \frac{2\times5+3\times3}{\sqrt{2^2+3^2+2^2}\sqrt{2^2+5^2+3^2}} \approx 0.7475$$

$$\text{Sim}(m_2, m_3) = \frac{3\times1+2\times2}{\sqrt{2^2+3^2+2^2}+\sqrt{3^2+1^2+2^2}} \approx 0.4537$$

The rating of u_1 to m_2 can be estimated as:

$$r_{u_1,m_2} = \frac{2\times0.7475+3\times0.4537}{0.7475+0.4537} \approx 2.38$$