COMP9313 Assignment

Chenxuan Rong(z5121987)

November 5, 2017

Question1. MapReduce

Algorithm 1: top-5 most expensive products purchased by each user in 2016

Input: log file
Output: userID+productID list

```
1  class MAPPER:
2      initialization
3      method mapper(object,text):
4          userID, productID, price, time = input.split(" ")
5          if time="2016" then
6              t ← userID
7              value ← price + " " + productID
8          end
9          EMIT(string t, string value)
10
11 class REDUCER:
12     method reducer(text,text):
13         userID ← key
14         Hashmap priceMap = new Hashmap()
15         for value in values: do
16             value.split(" ")
17             priceMap.put (value[0],value[1])
18         end
19         priceMap inverse key value
20         priceMap.sortByKey()
21         for i=0,i <5,i++ do
22             produdctIDList += priceMap [i]
23         end
24         EMIT(string userID, string produdctIDList)
```

Question2. MinHash

Row	C_1	C_2	$h_1(n) = 3n + 2 \ mod \ 7$	$h_2(n) = 2n - 1 \ mod \ 7$
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

The initialize step ,we set all value to ∞

	C_1	C_2
h_1	∞	∞
h_2	∞	∞

Table 1: Initialize step

	C_1	C_2
h_1	∞	2
h_2	∞	6

Table 2: processing row 0

	C_1	C_2
h_1	5	2
h_2	1	6

Table 3: processing row 1

When row2 are processed, the value of C_2 in signature matrix should be updated with 1,3 (since 1<2 and 3<6)

	C_1	C_2
h_1	5	1
h_2	1	3

Table 4: processing row 2

Since in row 3, both C_1 and C_2 are 0, we don't need to anything this time. Then row 4 are processed, as I explained above, after update, the matrix should be like

	C_1	C_2
h_1	0	0
h_2	0	0

Table 5: processing row 4

Now all the values in matrix are 0 , of course we can keep processing the remaining two rows, but the consequence should stay the same as above.

Question3. Streaming Data

The input strem is 0101010101 and window size is 60, the bucket will be updated only when timestampe is even.

1. **t=202**
   (16,148)(8,162)(8,177)(4,183)(2,192)(1,197)(1,200)(1,202)
   we combine first two elements:
   (16,148)(8,162)(8,177)(4,183)(2,192)(2,200)(1,202)

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

3. **t=206**
   (16,148)(8,162)(8,177)(4,183)(2,192)(2,200)(1,202)(1,204)(1,206)
   same as shown above:
   (16,148)(8,162)(8,177)(4,183)(2,192)(2,200)(2,204)(1,206)
   Again, we combine the first two elements:
   (16,148)(8,162)(8,177)(4,183)(4,200)(2,204),(1,206)

4. **t=208**
   (16,148)(8,162)(8,177)(4,183)(4,200)(2,204)(1,206)(1,208)
   Now, recall the **window size** is 60, so the oldest record (16,148) shoule be dropped:
   (8,162)(8,177)(4,183)(4,200)(2,204)(1,206)(1,208)

5. **t=210**
   (8,162)(8,177)(4,183)(4,200)(2,204)(1,206)(1,208)(1,210)
   After combining, we have the final output as:
   (8,162)(8,177)(4,183)(4,200)(2,204)(2,208)(1,210)

Question4. Collaborative Filtering

a) user-user collaborative filtering

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

Table 6: Similarity Metric

We can compute similarity between u_1 and other users by cosine similarity formula.

- $sim(u_1, u_2) = \dfrac{2 \times 5}{\sqrt{2^2 + 3^2}\sqrt{5^2 + 2^2}} \approx 0.5150$
- $sim(u_1, u_3) = \dfrac{2 \times 3 + 3 \times 1}{\sqrt{2^2 + 3^2}\sqrt{3^2 + 3^2 + 1^2}} \approx 0.5727$
- $sim(u_1, u_4) = \dfrac{3 \times 2}{\sqrt{2^2 + 3^2}\sqrt{2^2 + 2^2}} \approx 0.5883$

Hence, the predication of u_1 's rating in terms of m_2 can be calculated as:

$$\begin{aligned} r_{u1,m2} &= \frac{r_{u_2,m_2} \times sim(u_1, u_2) + r_{u_3,m_2} \times sim(u_1, u_3) + r_{u_4,m_2} \times sim(u_1, u_4)}{sim(u_1, u_2) + sim(u_1, u_3) + sim(u_1, u_4)} \\ &= \frac{2 \times 0.5150 + 3 \times 0.5725 + 2 \times 0.5883}{0.5150 + 0.5727 + 0.5883} \\ &\approx 2.34 \end{aligned}$$

b) item-item collaborative filtering This time, we need to compute the similarity between items

- $sim(m_2, m_1) = \dfrac{2 \times 5 + 3 \times 3}{\sqrt{2^2 + 3^2 + 2^2}\sqrt{2^2 + 5^2 + 3^2}} \approx 0.7475$
- $sim(m_2, m_3) = \dfrac{2 \times 2 + 3 \times 1}{\sqrt{2^2 + 3^2 + 2^2}\sqrt{3^2 + 1^2 + 2^2}} \approx 0.4537$

the predication of u_1's rating to m_2 is:

$$\begin{aligned} r_{u_1,m_2} &= \frac{r_{u_1,m_1} \times sim(m_2, m_1) + r_{u_1,m_3} \times sim(m_2, m_3)}{sim(m_2, m_1) + sim(m_2, m_3)} \\ &= \frac{2 \times 0.7475 + 3 \times 0.4537}{0.7475 + 0.4573} \\ &\approx 2.38 \end{aligned}$$