Question 2 : Data Analytic

ธีรภัทร ดีประเสริฐ (Theerapat Deeprasert)
theerapat.hybrid@gmail.com

2.1 Please find the correlation between products in the dataset (sheet data1) and explain in detail how to implement a solution to find them. You can use any tool and lib.

2.1.2 Apriori Algorithm

To find the correlation between products, an Apriori Algorithm is used. Apriori Algorithm is an algorithm that is widely used in Frequent pattern discovery today that is proved to be effective to find relationships between items for recommending related items in online shopping systems.

The Apriori Algorithm needs 2 values to be set before the algorithm can be used: Support and Confidence values(%). Support value is a threshold percent of the number of items so items that appear less than the threshold compared to other items will not be used in calculation. Confidence values is a threshold percent that the algorithm uses to determine which association of items can be used to have percent of possibility that are more than the confidence value.

The Algorithm consists of two parts of operation: Create Frequent Itemsets and Create association rule. The first part counts the frequency of items that appear in a dataset and then compares with support value if any item has frequency lower than the support value will be cut out of the operation. A simple flow chart of the first part of the operation is shown as follows.

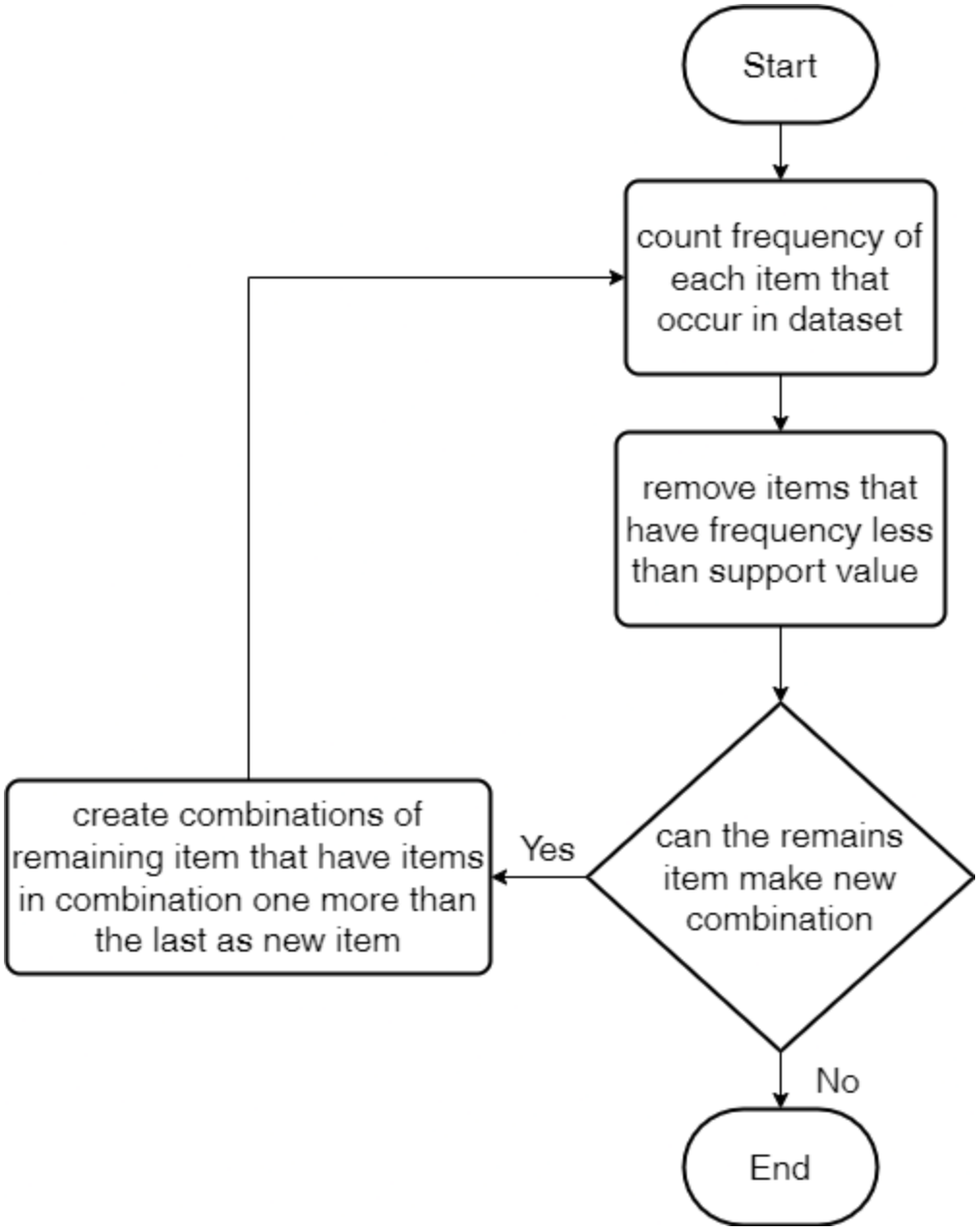

Figure 1: Flowchart to make Frequent Itemsets

The second part creates association between items by permuting all possibilities of items that are left in the last part with the most combination along with possibility percent and algorithm will cut out one that has possibility less than Confidence values.

2.1.2 Method

To apply the apriori algorithm with the dataset given, a local database was set up with the dataset inside to be called by the algorithm. Start by using SQL to query items with frequency in selected transactions and then enter it into an array. Next remove items in the array that have frequency less than threshold that can be calculated using the equation 1.

$$Threashold = \frac{S}{S+C} \cdot T \qquad (1)$$

S Support value
C Confidence value
T Number of Transaction

Then, create another array with all permutation combinations of two from items that were left in the last array. Compare the frequency of item combinations in the new array appearing in transactions, if less than the threshold, remove from the new array. Iterate the method until no new array of higher number of item combinations can be created.

Finally, creates association rules from the array of items that has the most combination number. Possibility value of each association can be calculated by finding the frequency of item combination that contains all items in the association divided by frequency of item combination that happened before the association. A simple example of possibility value calculation can be shown as follows.

Association of item I3 if item I1 and I2 was already in the transaction. possibility of item I3 is in transaction with item I1 and I2 can be calculated by dividing frequency of transaction with item I1, I2, and I3 with frequency of transaction with item I1 and I2. If the probability value calculated was less than the confidence value, the algorithm will determine that the association is not good enough.

2.2 From question 2.1, Please find the correlation between products for Customer R44.

Material used to answer this question is “material_000010”, because Customer R44 orders this material the most across all his/her transactions at 250 transactions. The information can be known by using the following SQL to query from the local database.

```
SELECT Material,COUNT(*) as num FROM sale WHERE CustomerID = 'R44' GROUP BY Material ORDER BY num DESC;
```

Material	num 1
material_000010	250
material_000011	242
material_000012	235
material_000028	235
material_000032	226
material_000030	221

To use apriori algorithm support and confidence values have to be set, so in this answer set support value at 60% and confidence value at 80%. With both values set, threshold frequency can be calculated.

$$Threshold = \frac{60}{(60+80)} \cdot 250$$

From the equation above threshold is 107.143 which can be interpreted as any frequency below 108 will not be included. Material that shares transactions with “material_000010” and has frequency above 107 can be shown by using the following SQL.

```
SELECT Material,
count(*) as num
FROM sale
WHERE Saleorder
IN (
      SELECT DISTINCT(Saleorder)
      FROM sale
      WHERE CustomerID = 'R44'
      AND Material = 'material_000010'
)
GROUP BY Material
HAVING num > 107;
```

Material	num
material_000010	250
material_000011	211
material_000012	193
material_000013	182
material_000028	199
material_000030	115
material_000155	139

Then the frequency of combination of two materials is shown by using SQL as follows(remark: from now on material name is reduced to last 3 characters to be easier to describe).

item set	frequency
{[010],[011]}	211
{[010],[012]}	192
{[010],[013]}	182
{[010],[028]}	199
{[010],[030]}	115
{[010],[155]}	139
{[011],[012]}	196
{[011],[013]}	165
{[011],[028]}	179
{[011],[030]}	116
{[011],[155]}	143
{[012],[013]}	151
{[012],[028]}	156
{[012],[030]}	132
{[012],[155]}	142
{[013],[028]}	180
{[013],[030]}	81
{[013],[155]}	110
{[028],[030]}	86
{[028],[155]}	112
{[030],[155]}	154

Then the frequency of combination of two materials is shown by using SQL as follows

item set	frequency
{[010],[011],[012]}	179
{[010],[011],[013]}	158
{[010],[011],[028]}	175
{[010],[011],[030]}	99
{[010],[011],[155]}	125
{[010],[012],[013]}	143
{[010],[012],[028]}	154
{[010],[012],[030]}	101
{[010],[012],[155]}	118
{[010],[013],[028]}	161
{[010],[013],[155]}	102
{[010],[028],[155]}	109
{[010],[030],[155]}	98
{[011],[012],[013]}	137
{[011],[012],[028]}	148
{[011],[012],[030]}	104
{[011],[012],[155]}	123
{[011],[013],[028]}	145
{[011],[013],[155]}	96
{[011],[028],[155]}	104

{[011],[030],[155]}	102
{[012],[013],[028]}	127
{[012],[013],[155]}	91
{[012],[028],[155]}	95
{[012],[030],[155]}	109
{[013],[028],[155]}	90

Iteration 4 with frequency > 107

item set	frequency
{[010],[011],[012],[013]}	134
{[010],[011],[012],[028]}	147
{[010],[011],[012],[155]}	112
{[010],[011],[013],[028]}	142
{[010],[012],[013],[028]}	125
{[011],[012],[013],[028]}	120

Iteration 5 with frequency > 107

item set	frequency
{[010],[011],[012],[013],[028]}	119

With Frequent Item sets at fifth iteration and cannot go further, association can be made to measure association between items. Example associations can be listed as follows.

Association	Confidence values
if transactions have [010],[011],[012],[013] customer will by [028]	119/134*100 = 88%
if transactions have [010],[011],[012],[028] customer will by [013]	119/147*100 = 80.9%
if transactions have [010] customer will by [011],[012],[013],[028]	119/250*100 = 47.6%

From examples, only the first two are considered valid because the possibility of the association exceeds the confidence value threshold which is 80%.