Author: **Khang Tran**

November 8, 2020

CSC 365 – Fall 2020

Project Report

**Association Rule Mining**

**REPORT**

**I. Theoretical Discussion**

1. Theoretical Calculation

\_ **Support Count(σ)** – Frequency of occurrence of a itemset.

\_ **Support(s)** –

The number of transactions that include items {X} of the rule as a percentage of the total number of transactions. It is a measure of how frequently the collection of items occur together as a percentage of all transactions.

\_ **Confidence(c)** –

It is the ratio of the no of transactions that includes all items in {B} as well as the no of transactions that includes all items in {A} to the no of transactions that includes all items in {A}. It measures how often each item in Y appears in transactions that contains items in X also.

2. Data Preprocessing Steps

\_ Convert the Groceries dataset into a transaction dataset where each row contains the following information: CustomerID, Date of purchase, no\_of\_items(k), item1, item2, ..., itemk.

\_ Convert the transaction dataset into an itemset list which has all information of purchased items.

\_ Convert the itemset list into support list which has:

+ All sub itemsets.

+ The frequency of all itemsets in the list by calculating the occurrence of these.

\_ Convert the frequency data into the list of association rule which has relationships among large sets of data items having more than 2 items.

**II.Results**

\_ support.txt

\_ associationRule.txt

**III. Discussion and Error Analysis**

|  |  |  |
| --- | --- | --- |
| **Support Threshhold for Itemsets** | | |
| Itemsets | Highest support | Lowest support |
| 1 | 2363/14963 | 1/14963 |
| 2 | 222/14963 | 1/14963 |
| 3 | 22/14963 | 1/14963 |
| 4 | 5/14963 | 1/14963 |
| 5 | 3/14963 | 1/14963 |
| 6 | 1/14963 | 1/14963 |
| 7 | 1/14963 | 1/14963 |
| 8 | 1/14963 | 1/14963 |

\_ Whole milk is the best seller in the grocery store with support: 2363/14963 0.15792287642.

\_ The relation has the highest confidence (0.024363233665559248) is [sausage] [yogurt, whole milk].

\_ The whole milk has the most relations, which means they often buy something else after taking the whole milk.

\_ When I set minimumSupport = 0.001 and minimumConfidence = 0.0001, there are just 54 relations between itemsets. As a result, we can conclude that the customers rarely buy more than 2 items in a grocery store.

\_ If there are no minimumSupport and minimumConfidence, the itemsets cases will be unlimited and too large. This will lead to the OutOfMemoryError, so the program can not run.