Market Basket Insights

Introduction:

Market Basket analysis is a data mining method focusing on discovering purchase patterns of the customers by extracting association or co-occurrences from a store's transactional data. For example, when the person checkout items in a supermarket all the details about their purchase goes into the transaction database. Later, this huge data of many customers are analyzed to determine the purchasing pattern of customers. Also decisions like which item to stock more, cross selling, up selling, store shelf arrangement are determined.

Association rule mining (ARM) identifies the association or relationship between a large set of data items and forms the base for market basket analysis. Association rule mining has been widely used in various industries besides supermarkets, such as mail order, telemarketing production, fraud detection of credit card and e-commerce.

One of the challenges for companies that have invested heavily in customer data collection is how to extract important information from their vast customer databases and product feature databases, in order to gain competitive advantage. Market basket analysis has been intensively used in many companies as a means to discover product associations.

A retailer must know the needs of customers and adapt to them. Market basket analysis is one possible way to find out which items can be put together.

Problem Statement

Nowadays people buy daily goods from super market nearby. There are many supermarkets that provide goods to their customer. The problem many retailers face is the placement of the items. They are unaware of the purchasing habits of the customer so they don't know which

items should be placed together in their store. With the help of this application shop managers can determine the strong relationships between the items which ultimately helps them to put products that co-occur together close to one another. Also decisions like which item to stock more, cross selling, up selling, store shelf arrangement are determined

Objectives

a. To identify the frequent items from the transaction on the basis of support and confidence
b. To generate the association rule from the frequent item sets.

Scope

The scope of the application is limited to desktop application right now. The application is targeted towards a supermarket of Nepal.

Apriori algorithm

Association rule mining finds interesting associations and/or correlation relationships among large set of data items. Association rules shows attribute value conditions that occur frequently together in a given dataset. A typical and widely used example of association rule mining is Market Basket Analysis. For example, data are collected from the supermarkets. Such market basket databases consist of a large number of transaction records. Each record lists all items bought by a customer on a single purchase transaction.

Association rules provide information of this type in the form of “IF-THEN” statements. The rules are computed from the data, an association rule has two numbers that express the degree of uncertainty about the rule.

a. Support

b. Confidence

Support

The support of an item is the number of transaction containing the item. Those items that do not meet the minimum support are excluded from the further processing. Support determines how often a rule is applicable to a given data set.

Support (XUY) =min (Support(X), Support(Y))

Confidence

Confidence is defined as the conditional probability that a transaction containing the LHS will also contain the RHS.

Confidence (LHS->RHS->

P(RHS/LHS)=P(RHS∩LHS)/P(LHS)=support(RHS∩LHS)/support(LHS).

Confidence determines how frequently item in RHS appears in the transaction that Contain LHS. While determining the rules we must measure these two components as it is very important to us. A rule that has very low support may occur simply by chance.

Pseudocode

//Find all frequent itemset

Apriori(database D of transaction, min_support){

F1={frequent 1-itemset}

K=2

While Fk-1≠ Empty Set

Ck=AprioriGeneration (Fk-1)//Generate candidate item sets.

For each transaction in the database D {

Ct=subset (Ck, t)

For each candidate c in Ct{

Count c++

}

Fk={c in Ck such that countc>min_support}

K++

}

F=U K>Fk

}

//prune the candidate item sets

Apriori generation (Fk-1) {

//Insert into Ck all combination of elements in Fk-1 obtained by self-joining item sets in Fk-1

//Delete all item sets c in Ck such that some (K-1) subset of c is not in Lk-1 }

//find all subsets of candidate contained in t

Subset (Ck, t)

}

Dataset Description

- File name: Assignment-1_Data
- List name: retaildata
- File format: . xlsx
- Number of Row: 522065
- Number of Attributes: 7
    -
        - BillNo: 6-digit number assigned to each transaction. Nominal.
        - Itemname: Product name. Nominal.
        - Quantity: The quantities of each product per transaction. Numeric.
        - Date: The day and time when each transaction was generated. Numeric.
        - Price: Product price. Numeric.
        - CustomerID: 5-digit number assigned to each customer. Nominal.

- Country: Name of the country where each customer resides. Nominal.

	A	B	C	D	E	F	G
1	BillNo	Itemname	Quantity	Date	Price	CustomerID	Country
2	536365	WHITE HANGING HEART T-LIGHT HOLDER	6	01.12.2010 08:26	2,55	17850	United Kingdom
3	536365	WHITE METAL LANTERN	6	01.12.2010 08:26	3,39	17850	United Kingdom
4	536365	CREAM CUPID HEARTS COAT HANGER	8	01.12.2010 08:26	2,75	17850	United Kingdom
5	536365	KNITTED UNION FLAG HOT WATER BOTTLE	6	01.12.2010 08:26	3,39	17850	United Kingdom
6	536365	RED WOOLLY HOTTIE WHITE HEART.	6	01.12.2010 08:26	3,39	17850	United Kingdom

items			
WHITE HANGING HEART T-LIGHT HOLDER	WHITE METAL LANTERN	CREAM CUPID HEARTS COAT HANGER	KNITTED UNION FLAG HOT WATER BOTTLE
HAND WARMER UNION JACK	HAND WARMER RED POLKA DOT		
ASSORTED COLOUR BIRD ORNAMENT	POPPY'S PLAYHOUSE BEDROOM	POPPY'S PLAYHOUSE KITCHEN	FELTCRAFT PRINCESS CHARLOTTE DOLL
JAM MAKING SET WITH JARS	RED COAT RACK PARIS FASHION	YELLOW COAT RACK PARIS FASHION	BLUE COAT RACK PARIS FASHION
BATH BUILDING BLOCK WORD			
ALARM CLOCK BAKELIKE PINK	ALARM CLOCK BAKELIKE RED	ALARM CLOCK BAKELIKE GREEN	PANDA AND BUNNIES STICKER SHEET
PAPER CHAIN KIT 50'S CHRISTMAS			
HAND WARMER RED POLKA DOT	HAND WARMER UNION JACK		
WHITE HANGING HEART T-LIGHT HOLDER	WHITE METAL LANTERN	CREAM CUPID HEARTS COAT HANGER	EDWARDIAN PARASOL RED
VICTORIAN SEWING BOX LARGE			
WHITE HANGING HEART T-LIGHT HOLDER	WHITE METAL LANTERN	CREAM CUPID HEARTS COAT HANGER	EDWARDIAN PARASOL RED
HOT WATER BOTTLE TEA AND SYMPATHY	RED HANGING HEART T-LIGHT HOLDER		
HAND WARMER RED POLKA DOT	HAND WARMER UNION JACK		
JUMBO BAG PINK POLKADOT	JUMBO BAG BAROQUE BLACK WHITE	JUMBO BAG CHARLIE AND LOLA TOYS	STRAWBERRY CHARLOTTE BAG
JAM MAKING SET PRINTED			
RETROSPOT TEA SET CERAMIC 11 PC	GIRLY PINK TOOL SET	JUMBO SHOPPER VINTAGE RED PAISLEY	AIRLINE LOUNGE

At this step we already have our transaction dataset, and it shows the matrix of items which bought together. We can't see here any rules and how often it was purchase together. Now let's check how many transactions we have and what they are. We will have to have to load this transaction data into an object of the transaction class. This is done by using the R function read.transactions of the arules package. Our format of Data frame is basket.

```
34  transactions <- read.transactions('/Users/asik/Desktop/assigment1_itemslist.csv',
35                                    format = 'basket', sep=',')
```

Let's have a view our transaction object by summary(transaction)

```
36  summary(transactions)
```

We can see 18193 transactions (rows) and 7698 items (columns). 7698 is the product descriptions and 18193 transactions are collections of these items.

```
transactions as itemMatrix in sparse format with
 18193 rows (elements/itemsets/transactions) and
 7698 columns (items) and a density of 0.002291294

most frequent items:
WHITE HANGING HEART T-LIGHT HOLDER           REGENCY CAKESTAND 3 TIER            JUMBO BAG RED RETROSPOT
                              1718                               1468                               1395
                     PARTY BUNTING       ASSORTED COLOUR BIRD ORNAMENT                            (Other)
                              1245                               1226                             313843

element (itemset/transaction) length distribution:
sizes
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19   20   21   22   23   24   25   26   27
1546  860  744  743  743  696  642  633  632  566  598  517  494  520  533  508  460  428  468  406  385  307  306  267  232  246  226
  28   29   30   31   32   33   34   35   36   37   38   39   40   41   42   43   44   45   46   47   48   49   50   51   52   53   54
 210  213  209  164  153  135  140  131  108  109   88  108   90   86   84   84   63   58   67   59   58   57   48   60   39   39   47
  55   56   57   58   59   60   61   62   63   64   65   66   67   68   69   70   71   72   73   74   75   76   77   78   79   80   81
  41   35   27   37   29   26   27   16   24   25   20   27   24   23   13   20   19   13   16   15   11   15   12    6    7   14   13
  82   83   84   85   86   87   88   89   90   91   92   93   94   95   96   97   98   99  100  101  102  103  104  105  106  107  108
  10    8    8   11   10   13    8    6    5    5   11    5    4    4    3    5    5    2    4    1    4    4    2    2    2    6    3
 109  110  111  112  113  114  116  117  118  120  121  122  123  125  126  127  131  132  133  134  140  141  142  143  145  146  147
   4    3    2    1    3    1    3    3    3    1    2    2    1    3    2    2    1    1    2    1    1    2    2    1    1    2    1
 150  154  157  168  171  177  178  180  182  202  204  228  249  250  285  320  400  419
   1    3    2    2    2    1    1    1    1    1    1    1    1    1    1    1    1    1

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    5.00   13.00   17.64   23.00  419.00

includes extended item information - examples:
                  labels
1              1 HANGER
2     10 COLOUR SPACEBOY PEN
3 12 COLOURED PARTY BALLOONS
```

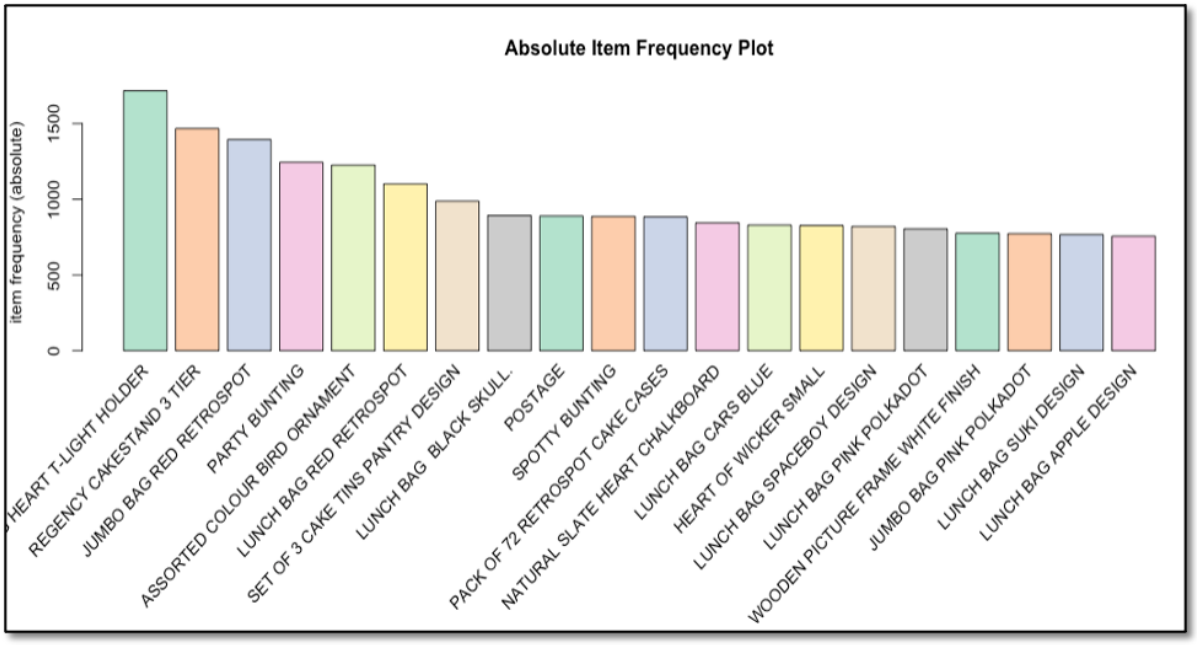
The summary gives us some useful information:

- Density tells the percentage of non-zero cells in a sparse matrix. In other words, total number of items that are purchased divided by a possible number of items in that matrix. You can calculate how many items were purchased by using density: 18193x7698x0.002291294=337445
- Summary will show us most frequent items.
- Element (itemset/transaction) length distribution: It will gave us how many transactions are there for 1-itemset, 2-itemset and so on. The first row is telling you a number of items and the second row is telling you the number of transactions.
  For example, there is only 1546 transaction for one item, 860 transactions for 2 items, and there are 419 items in one transaction which is the longest.

Let's check item frequency plot, we will generate an itemFrequencyPlot to create an item Frequency Bar Plot to view the distribution of objects based on itemMatrix (e.g., >transactions or items in >itemsets and >rules) which is our case.

```
itemFrequencyPlot(transactions,topN=20,type="absolute",
                  col=brewer.pal(8,'Pastel2'), main="Absolute Item Frequency Plot")
```

```
if (!require("RColorBrewer")) {install.packages("RColorBrewer")
  library(RColorBrewer)
```



In itemFrequencyPlot(transaction,topN=20,type="absolute") first argument - our transaction object to be plotted that is tr. topN is allows us to plot top N highest frequency items. type can be as type="absolute" or type="relative". If we will chouse absolute it will plot numeric frequencies of each item independently. If relative it will plot how many times these items have appeared as compared to others. As well I made it in colure for better visualization.

Conclusion

The Apriori algorithm effectively generates highly informative frequent itemsets and association rules for the data of the supermarket. The frequent data items are generated from the given input data and based on the frequent item stets strong association rules were generated.