CHAPTER 3: SYSTEM DESIGN

3.1 Methodology

3.1.1 Data collection

The data was collected from http://www.salemmarafi.com/wp content/uploads/2014/03/groceries.csv due to the unavailability of data from the supermarkets.

Table 4- Sample data

Transaction	Items				
1	Fruit	Bread	Butter	Soups	
2	Fruit	Yogurt	Coffee		
3	Milk				
4	Fruit	Yogurt	Cheese	Meat	
5	Vegetables	Milk	Bakery		
6	Milk	Butter	Yogurt	Rice	cleaner
7	Rolls/bun				

3.1.2 Data preprocessing

The data collected was mapped manually as integer values as shown in Figure 4. For example the "Fruit" was labeled as 1, "Bread" as 2 "Soups" as 4 and so on.

```
1,fruit
2, bread
4,soups
6,yogurt
7,coffee
10,cheese
108,meat
12,vegetables
13,milk
14,bakeryproduct
15,butter
16,rice
17,cleaner
19,buns
21,beer
22,appetizer
23,potplants
24,cereals
26,bottledwater
27,chocolate
18,curd
28,flour
29,dishes
30,beef
31,frankfurter
```

Figure 4- Mapped to integers

The mapped integer's values were then saved in a text file and given as the input to the system. Figure 5 shows the input file that is given to the system.

```
1,2,3,4,5
2,10,9,11,5
1,5
1,3,8,10,100
2,4,6,8,10,11,12,14
102,3,4,5,6,8
34,456,123,67
45,34,56,7,8,9,100
67,11,123,11,23,45
89,4,5
```

Figure 5- Input file to system

The Apriori algorithm was used for processing the input data and result was produced as the list of rules that are strongly associated with each other.

3.1.3 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&gt;min_support}
K++
}
F=U K&gt;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)
}
```

3.2 System Design

3.2.1 Class diagram

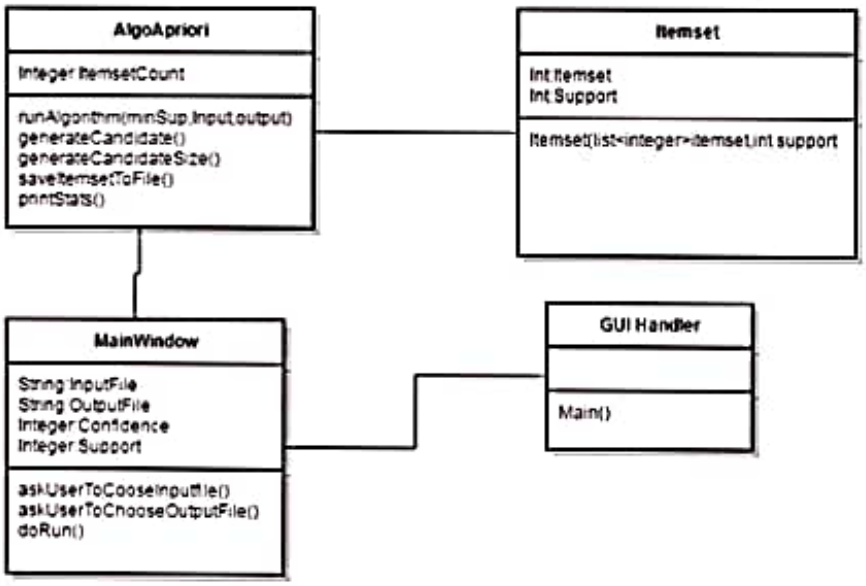

Figure 6- Class diagram

As shown in Figure 6, there are three main classes used in the application

The mainWindow class is used to present the user interface for choosing the input file and output file as desired by the user.

The AlgoApriori is the class that performs all the calculations once the data is provided by the user.It generates the candidate item sets and determines the size of the item sets. Finally the statistics are provide to the user in the same GUI and output is written to the desired file.

The item set class stores the items as the array of integer and provides the support of the respective item from the given input data

3.2.2 Sequence diagram

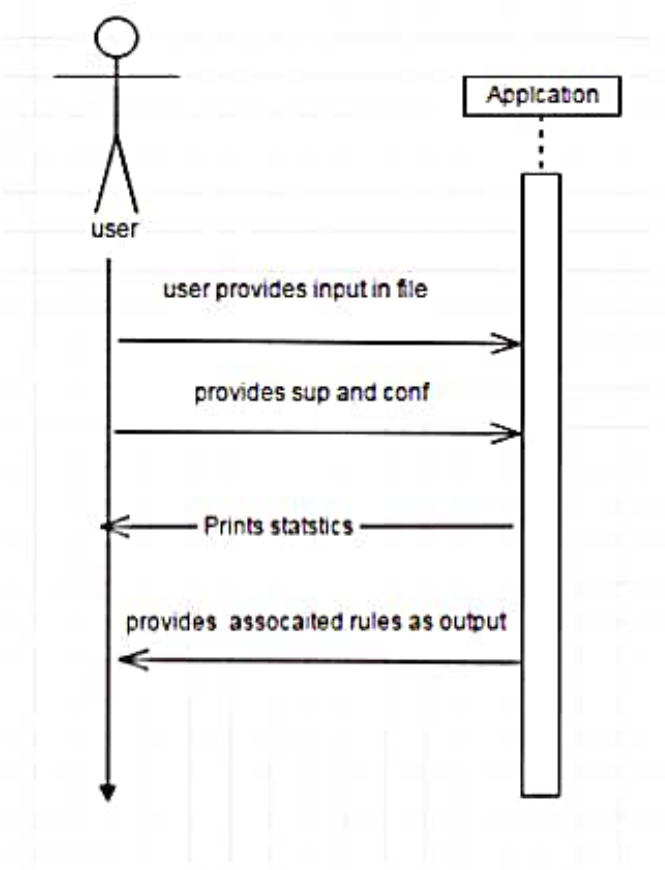

Figure 7- Sequence diagram

Figure 7 shows the sequence diagram of this application.

The user needs to choose the input file that is going to be processed. The file should contain the data in integer where the row represents the items and column represents transactions.

Confidence and support should be provided by the user. After all the input is given the application process the data and provide the output to the user.

The output will be a text file containing the association rules.