CHAPTER 1

INTRODUCTION

1.1 Rationale

The food industry founded in 2002 has large volumes of data stored in the transaction processing system. The company sells various food products. There are 163 variations of products. According to the customers' choice regarding the product, when the consumers buy food A, they also buy food B, but on contrary they do not buy food A if they buy food B or food C and vice versa. The Company wants to build business actions to address the customers' taste patterns.

The goals can be achieved by selling products that meet the tastes and consumers' behaviors based on the way they select the food products. The ability to read consumers' tastes is obtained using data mining techniques with method called Market Basket Analysis. This method shows people's decisions to buy products [8][9][12].

However problems arise with the large number of data which will be used. There are 163 attributes with more than 90,000 records in 5 months. These will continue to expand along with the amount of data to be analyzed. Data preprocessing technique needs to determine specifically to handle large volumes of data to find the patterns of the food product consumers' tastes based on the sales transaction data in mining process using apriori algorithm[1][10][15]. After that, time series pattern on each group of goods are used as other way to see what kind of rules can be implemented on each group of goods[5].

1.2 Theoretical Framework

The knowledge used for designing business action is obtained by analyzing the relationship or correlation between the set of items in large amounts of transactional data called the basket data. The Market Basket Analysis will form relationship that becomes the rule: 'if consumers buy the X and Y then they also will buy Z'.

The presence of item X and Y is believed as a part of characteristics in Basket Data which reveal confidence while the correlation with the item set Z raises the ratio of transactions defined as a variable support. Thus the transactional data can be mapped into a pattern that

explains the consumers' taste patterns' for the products sold , then the information can be used as a reference to build business actions [1][5][6][10][11][12].

The association theory can be explained with the examples related to three products, such as: food A, food B and Food C. The samples of numbers association rules enlisted bellow with 'support' symbolized as (s) and confidence symbolized as (c):

1) if food_A = yes food_B = no then food_C = yes 4(s) 100%(c)

2) if food_A = no then food_C = no 4(s) 70%(c)

As mentioned above, the association rule can be explained:

- The association rule number one states: if customers buy food A and do not buy food B then the customers will buy food C.

  The rule number one happened four times in all data which is identified as support. The rule shows customers buy food A and do not buy food B always follow by the action to buy food C, the condition happened one hundred percent truth of the sequences regarding to the transactional data, by mean the customers who buy food A and do not buy food B , definitely will buy food C and no other which is identified as 100% of confidence.

- The second example of association rule states that if customers do not buy food A then the customers will not buy food C.

  This condition happened four times in all data which is defined with support value is four. The condition which shows customers do not buy food A then they will not buy food C, happened with the seventy percent of confidence, by mean the customers who do not buy food A then will not buy food B only happened 70% compare to the condition of the customers who do not buy food A. In this case, customers do not just choose not to buy food C when they decide not to buy food A, and some customers who do not buy food A, they might buy or do not buy other products except food C.

According to the example, not all attributes that develop rules are worth enough to explain the patterns of customers' buying decisions. Therefore, to determine the most potential rules

of customers' choice relationship, the percentage of confidence is involved as a measurement [10][15].

1.3 Statement of the Problem

Based on the rationale above, some problems are identified, namely:

1. Developing association rules base on the high volumes of transactional data contain 163 attributes with more than 90,000 records within 5 months from January 2012 until May 2012. This number will continuously increase along with the amount of data to be added in data preprocessing to be analyzed.

2. Regarding to the high volumes of transactional data, the company want to build the association rules to be uses as a basis of business actions

1.4 Conceptual Framework

The problems mentioned before are related to analyze data transaction. To create the procedure to handle massive data then needs process using mining techniques. These need some theories and concepts related to data preparation, reduction process, clustering method and association method.

The methods used for data preparation is normalization, attribute construction and aggregation, whereas the methods used for data reduction is feature selection. Clustering method using K-Means algorithm, while the association method using Apriori Algorithm.

The scope of framework is started with separate the transactional data from the transaction processing system which is archived as database file. Then the transformation is continued by forming atomic data. The atomic data will be reduced and mining stage will be held using apriori algorithm to build association rules.

Another data transformation techniques will be used to transform atomic data to build time series pattern and to fit the procedure in data clustered. The result of time series and clustering expected to strengthen the truth of the association rules which is been built.

Data transformation techniques which are used briefly enlisted bellow:

- Data transformations for building association rules are normalization and attributes construction.
- Data transformation for building time series patterns is aggregation.
- Data transformation for building clustering is aggregation.

Data reduction techniques which are used briefly enlisted bellow:

- Data reduction for building association rules is feature selection.
- Data transformation for building clustering patterns is feature selection.

1.5 Hypothesis

Implementation of the FSA-Red reduction techniques and apriori method on high volumes of transactional data will be resulting the rules with confidence no less than 70%.

1.6 Assumption

Based on the general and specific problems, some assumptions developed were:

1. Time series pattern on each group of goods are used as other way to see what kind of rules can be implemented on each group of goods.
2. K-Means algorithm will be used to cluster transactional data to reveal clustered customer behavior.

1.7 Scope and Delimitation

Some scopes and delimitations from the research described are below:

1. The focus of this study is to build business actions base on strong association rule using apriori algorithm, time series data and clustering from the high volumes transactional data.
2. The transactional data is limited from January until May 2012, extended to July 2012.
3. The data set had 163 products translated into 163 attributes to be analyzed with 92,555 instances from January 2012 to May 2012 and 23,381 instances in June 2012 until July 2012. All data set amounted to 115936 with 163 attributes.

1.8 The Objective

The objective of research is to implement apriori method for generating rules and analyze the rules performance regarding groups of goods.

1.9 Importance of the Study

This study had many contributions such as:

1. Minimizing the high volumes of data to be analyzed regarding data reduction procedure before building association rules which beneficial for resources consumption.

2. Building strong association rules related to the selling products, which can be use as a basis for business action, for example the way product would be displayed which is easier to reach or bundle with few products as a package to increase selling.
3. The reduction procedures can be implemented by user justification to build association rules related to the products which flexibility to the user.
4. The Company can take advantage regarding the transactional data, rather than stored in transaction processing systems. Implementation of data mining is able to map the customers' transaction behavior of the data generated by the association rule can serve as the basis of a business action.