CHAPTER 1: THE PROBLEM

This chapter discusses the background of this study; it includes the following subtopics: rationale, theoretical framework, conceptual framework/paradigm, statement of the problem, hypothesis, assumption, scope and delimitation, and importance of the study.

1.1 Rationale

Association rules are rules that relate one thing to another thing in which two things are closely interrelated. Association rules categorized as unsupervised learning because it can generate rules dynamically in terms of number and variety [1]. There are two reference values are calculated on association rules which are support values and confidence values. Support value is the ratio of one variable against frequency of the entire transaction. Support value will be compared with the value of the specified minimum support, only support values equal to or greater that will be processed further. Confidence value is a value that indicates the relationship of one variable with another ones, it compares the value of two or more variables combination. Result that will be taken from association rules are values that meet the minimum support and minimum confidence [2] [3]. These results can be used to determine managerial business actions.

There are two common forms of association rules, one is by forming candidate itemset and other by forming tree structures. Both forms of the association rules require memory resources and storage resources allocations depends on the size of dataset. When large datasets are used then it will require a quite long processing times (depending on the ability of the hardware used). One solution to solve this problem is the use of DBMS as datasets media processing, where data transfer from secondary storage can be minimized and features of dynamic memory management also available. There are already exists previous research that use Oracle RDBMS as media processing datasets with apriori algorithm to produce association rules. [4]

FP-Growth is an association rules algorithm which does not generate candidate itemset. FP-Growth forming tree structures by going through the stages of generation FP-Tree, Conditional Pattern Bases, Conditional FP-Trees and ends with the formation of Frequent Pattern [5]. Frequent pattern will generate rules that will generate confidence values. Stages of sufficient length requires a relatively large resource that is used depends on the size of the dataset. Many previous studies which seek to minimize the time to make the FP-Growth algorithm process more effective by create new and better methods [6] [7] [8] [9]. This research will apply a new method

to improve completion time of FP-Growth algorithm by reducing the size of the dataset (preprocessing phase) but results of confidence values are relatively unchanged.

1.2 Theoritical Framework

The dataset is a collection of transaction data consisting of multiple variables, each variable has a link that refers to a key variable. A dataset can be modeled as a set $D_i = \{r_i \mid r_i$ is a record of the transaction$\}$. Record the transaction itself consists of variables that can be written as $r_i = \{v_{ik} \mid v_{ik}$ is the value of a variable). There are transitive dependencies between variables with a dataset that can be written as $v_{ik} \subseteq r_i$ and $r_i \subseteq D_i$ so $v_{ik} \subseteq D_i$. Frequent itemset is a set of variables from the dataset that have support values and confidence values. The formation of a frequent itemset is assumed the same as the formation of a dataset that is comprised of the variables in the set.

Association rules perform two calculations which are the calculation of the support value and the calculation of confidence value. The formula for calculating the support value occure on several stages (k-itemset). For the 1-itemset the calculation is Sup (X) = Σ(X) / ΣT, X is the variable / item and T is the transansion recorded by the dataset. The calculation of the 2-itemset support value is Sup (X, Y) = $\Sigma(X \cap Y) / \Sigma T$. The calculation for writing confidence value is Conf ($X \Rightarrow Y$) = (Sup (X, Y)) / (Sup (X)) or Conf ($X \Rightarrow Y$) = $\Sigma(X \cap Y) / \Sigma X$. [3]

In the paper discusses the association rule, the notation used for representing the dataset is the notation of set theory. I = {i1, i2, i3, ... i_m} denote the set of I, which consists of the attribute i_m. For $t \in D$, t declare transactions that occurred on the set of transactions D, where $t \subseteq I$.

In set theory there are several possible relationships that can occur in the sets that exist in a universe set S. The relationship could be combination, slice or set part. Combined is 2 or more sets relate to each other with the kind of $A \cup B$ where $A \in S$ and $B \in S$, next this kind of relationship will be called Union. The slices are 2 sets or more relate to each other with the kind of $A \cap B$ where $A \in S$ and $B \in S$, next it will be called Intersection. The set part is a set of completely into the other part of the larger set, $A \subseteq B$ where $A \in S$, $B \in S$ and $A \leq B$ [10]. Record dataset can be thought of as subset and variables can be thought of as elements of the set.

Oracle database is one of the existing RDBMS. Oracle RDBMS has its own association rule algorithm in the form of frequent itemset table function that dynamically calculating the value of support from a dataset [4] [11]. Frequent itemset table function which is on Oracle RDBMS acquired Apriori algorithm of data mining techniques. Apriori algorithm is able to generate the frequent itemset by previously generate candidate itemset. Candidate itemset is all combination of

variables that exist in a dataset (k-itemsets). Apriori algorithm has two drawbacks, namely the establishment candidate itemset might be very large and the repeated scans on the dataset when implemented [3] [9].

There are other algorithms that do not need to generate candidate itemset first, one is FP-Growth algorithm in which the performance of the FP-Growth algorithm is better than the Apriori algorithm. FP-Growth algorithm working principle is to generate FP-Tree. On FP-Tree, pruning / removal process occurs on items that do not meet minimum support to obtain the optimum frequent itemset. The formation process of FP-Tree that is used by FP-Growth might require large memory allocation (depend on dataset size), it will burden the work of computer processing. The workload can be handled by the DBMS workloads where memory can be processed with a buffering mechanism.

Apriori algorithm and FP-Growth applying machine learning concepts. Type used machine learning is unsupervised machine learning because it can generate rules dynamically in terms of number or combination of items in the rule [12].

There are some previous studies have been conducted to analyze the FP-Growth algorithm is applied to the DBMS [13] [14] [15] [16]. Most of the research results are obtained performance comparison between Apriori algorithm with FP-Growth algorithm where performance of FP-Growth algorithm is better than Apriori algorithm [2]. The use of FP-Growth algorithms in a DBMS obtained good performance on time, but the process of its formation requires several operations through the establishment of selection attribute table of the previous table using SQL. The attribute selection process is basically applying the concept of relation in set theory.

1.3 Conceptual Framework/Paradigm

The end result of the implementation of the FP-Growth algorithm on the Oracle RDBMS is a frequent pattern which will then result in the value of the rule and its confidence value. Rule that counts is the rule that has a confidence value above the specified minimum confidence. This study will analyze whether the value of confidence can be sustained if dataset decreased into a smaller dataset by analyzing the set theory. Analysis of set theory will be done on a frequent pattern is generated. Analysis results obtained will be compared with the items in the original dataset, the relationship can be used as a reference when pruning the items on the original dataset.

First will be generated the value of frequent itemset generated by Apriori algorithm and frequent pattern generated by the FP-Growth algorithm. Both of these results will then be compared, both

should have the same identical results. Frequent pattern will resulting rules after do some process of the formation of association rule. Based on the obtained rules, the confidence values will be obtained as well. The results of the dataset will be processed by FP-Growth algorithm and will resulting Frequent Pattern. Frequent Pattern will generate rules that can be calculated on confidence value. High value can indicate confidence linkages items contained in a rule.

The next stage is to analyze the frequent pattern obtained to generate a new dataset. This new dataset should be smaller in size than the initial dataset because it has been through the process of pruning on certain items. Furthermore, we will perform an analysis of the Rule and its confidence values. Confidence values obtained will be analyzed and compared with the confidence values at the early process, should not be any difference between the two results of the confidence values.

1.4 Statement of the Problems

Association rule mining is the process of finding patterns among the items that correlate in a dataset. Frequent itemset / frequent pattern must be obtained prior to the formation of association rule. Rules will consist of attributes that related, the establishment process of the rule is done in unsupervised. Unsupervised means an attribute doesn't need to be specified on which rule but would automatically become part of a particular rule [1]. Confidence value indicates rule faiths level of the resulting rule, greater the confidence value will make greater the rule faith. The resulting rules will varies on confidence values, there will be a high value (approaching 100%) and there will be also a low value (close to 0%). When processing datasets using association rule algorithm, not known yet whether the rules obtained will be on high or low confidence, it will be known after association rule execution is completed. The larger the dataset used, obtained rules will more varied and will require longer on the processing time.

1.5 Hypothesis

Confidence value is a value that indicates the relationship between an item with other items, higher of confidence value is obtained, the formed of association rules will be more valid. The time needed to complete the process of formation of association rule is to be greatly influenced by the size of the dataset, the larger the dataset then the time required will be longer. If a dataset reduced on size / dimensions, assumptions will result in faster execution time, but it should be noted that frequent pattern and the confidence value should not be changed.

Basically frequent pattern is a set of items that have relevance to transactions on the FP-Tree. Set theory is able to analyze related items in a set, which are intersection, union or subsets. A rule is a relationship between the correlated items. In RDBMS, relations between tables is the application

of set theory intersection. For example found table X = {a, b, c) is related to the table Y = {z, a, b} on attributes a and b, in theory, the set can be written as X ∩ Y = {a, b}. The results of several sets of relationships will generate a new set with a smaller number of attributes. The dataset will consist of one or a few records and a record will be made up of several items. Record is a representation of the set and the item is a representation of the attribute. By using the intersection of set theory and unsupervised learning of FP-Growth algorithm would be produces new dataset with smaller on dimensions / sizes.

1.6 Assumption

1. Datasets already in the Oracle RDBMS, do not need to do data migration process.
2. The records in the dataset will contain more than one variation of an item.
3. The test dataset that will be more than one with different sizes but same on table structure.

1.7 Scope and Limitation

1. The dataset will be used is two of real transaction datasets which are minimarket sales dataset and bread company sales dataset.
2. This research only processing rules with pattern X⇒Y.
3. This study will compare the sizes / dimensions of dataset before and after the model is applied.
4. Study will compare the execution time of FP-Growth algorithm before and after the model is applied.
5. This study will provide an interpretation of some rules that are formed and will recommend business actions based on the interpretations obtained.

1.8 Significant of the Study

1. The proposed model introduce a new algorithm called IST-EFP (Intersection Set Theory – Expand FP Growth) that is able to decrease the size of datasets.
2. Using the experiment results will found that smaller dataset size will improve the execution times.