Quantitative Association Rules Mining for Business Transactional Data

Dim En Nyaung, Wint Thida Zaw
University of Computer Studies, Yangon
dimennyaung@gmail.com, wintthida@gmail.com

Abstract

The explosive growth in data and database has generated a need for techniques and tools that can transform the processed data into useful information and knowledge that improves marketing strategy. Association rules mining is finding frequent patterns, associations, correlations, or causal structures among item sets in transaction databases, relational databases, and other information repositories. The relational tables that stored the transactions have richer attribute types such as quantitative and categorical attribute. Thus the development of tools that can extract useful information from this large database is greatly demand. This paper discusses the quantitative association rules mining from business transactional database that store the textile store. We introduce the quantitative association rules mining using with the direct application using on a real-life dataset.

Keywords: Association Rules Mining, quantitative and category attributes, quantitative association rules mining

1. Introduction

Data mining can be performed on data represented in quantitative, textual, or multimedia forms. Data mining applications can use a variety of parameters to examine the data. They include association (patterns where one event is connected to another event, such as purchasing a pen and purchasing paper), sequence or path analysis (patterns where one event leads to another event, such as the birth of a child and purchasing diapers), classification (identification of new patterns such as coincidences between duct tape purchases), clustering (finding and visually documenting group of previously unknown fact, such as geographic location and brand preferences), and forecasting (discovering patterns from which one can make reasonable predictions regarding future activities, such as the prediction that people who join an athletic club may take exercise classes) [5]. Data mining has become increasingly common in both the public and private sectors. A transaction database is a set of records representing transactions, each with a time stamp, an identifier and a set of items. Associated with the transaction files could also be descriptive data the items. Transactions are usually stored in flat or stored in two normalized tables, one for the transactions and one for the transaction items. According to the Market basket example: Basket 1: {bread, cheese, milk} and Basket 2: {apple, eggs, salt},…, Basket n: {biscuit, eggs, milk}. Definition for an item is that, an item: an article in a basket, or an attribute-value pair. And a transaction: items purchased in a basket; it may have TID (transaction ID). A transactional dataset: a set of transactions [6].

This paper discusses the usefulness of mining quantitative Association Rules by using with basic algorithm named Apriori and the guidelines of implementation for this Algorithm to the business oriented databases. The following sections show the algorithm and implementation of the association rules mining for business transactional data.

2. Related Works

Data mining, also known as knowledge discovery in databases, has been recognized as a new area for database research, The problem of discovering association rules was introduced in [3].

Mining frequent patterns or itemsets is a fundamental and essential problem in many data mining applications [1, 4]. These applications include the discovery of association rules, strong rules, correlations, sequential rules, episodes, multi-dimensional patterns, and many other important discovery tasks. Algorithms for extracting this basis and for reconstructing all association rules shows the results of experiments carried out on real datasets and it show the usefulness of each approach [2].

Most of the proposed pattern-mining algorithms are a variant of Apriori [4]. Apriori employs a bottom-up, breadth-first search that enumerates every single frequent itemset. Apriori

uses the downward closure property of itemset support to prune the search space — the property that all subsets of a frequent itemset must themselves be frequent. Thus only the frequent *k*-itemsets are used to construct candidate (*k* + 1)-itemsets. A pass over the database is made at each level to find the frequent itemsets among the candidates.

A typical example of Association rule mining is Market Basket Analysis. This process analyses customer buying habits by finding associations between the different items that customers place in their "shopping baskets".

Classical approaches for mining association rules with Market Basket Analysis operate in two phases. They are extracting frequent itemsets and generating association rules. In extracting frequent itemsets, each of these itemsets will occur at least as frequently as a pre-determined minimum support count. For generating association rules, it must satisfy minimum support and minimum confidence. Rules support and confidence are two measures of rule interestingness.

Association Rules are considered interestingness if they satisfy both a minimum support threshold and minimum confidence threshold [7]. The quantitative association rules problem can be mapped to the Boolean Association Rules problem [3].

3. Itemsets and Association Rules Analysis

An itemset is a set of items. E.g., {milk, bread, cereal} is an itemset. A k-itemset is an itemset with k items. Given a dataset D, an itemset X has a (frequency) count in D. An association rule is about relationships between two disjoint itemsets X and Y is defined as X $\Rightarrow$ Y and it presents the pattern when X occurs, Y also occurs. Association analysis is the discovery of what are commonly called association rules. Association analysis is commonly used for market basket analysis. Association rules do not represent any sort of causality or correlation between the two itemsets.

X $\Rightarrow$ Y does not mean X causes Y, so no causality

$X \Rightarrow Y$ can be different from $Y \Rightarrow X$, unlike correlation

Association rules assist in marketing, targeted advertising, floor planning, inventory control, churning management, homeland security etc. The main idea of Support and Confidence of Association rules need support and confidence. Support of X in D is count(X)/|D|.

For an association rule $X \Rightarrow Y$, we can calculate

support ($X \Rightarrow Y$) = support (XY)

confidence ($X \Rightarrow Y$) = support (XY)/support (X)

Support: Support of a rule is a measure of how frequently the items involved in it occur together. Confidence: Confidence of a rule is the conditional probability of B given A. These statistical measures can be used to rank the rules and hence the predictions.

3.1 Apriori Algorithm

The real advantage for decision making relies on the add-on provided by comparing the extracted knowledge against the apriori domain knowledge [2]. Apriori is an influential algorithm for mining frequent itemsets for Boolean association rules. The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent itemset properties. The Pseudo-code of Apriori Algorithm is as shown in below:

```
Ck: Candidate itemset of size k
Lk : frequent itemset of size k
L1 = {frequent items};
for (k = 1; Lk !=∅; k++) do begin
    Ck+1 = candidates generated from Lk;
   for each transaction t in database do
     increment the count of all candidates in Ck+1
       that are contained in t
   Lk+1  = candidates in Ck+1 with min_support
   end
return ∪k Lk;
```

Once the frequent itemsets from transactions in a database D have been found, it is straight forward to generate strong association rules from them (where strong association rules satisfy both minimum support and minimum confidence). This can be done using the the following equations for confidence, where the conditional probability is expressed in terms of itemset support count.

$$\text{Confidence}(A \Rightarrow B) = P(B \backslash A) = \frac{\text{Support\_count (AUB)}}{\text{Support\_count(A)}}$$

Where support_count(AUB) is the number of transactions containing the itemsets AUB, and support_count(A) is the number of transactions containing the itemset A. Based on this equation, association rules can be generated as follows:

- For each frequent itemset *l*, generate all non-empty subsets of *l*.

- For every non empty subsets of *l*, output the rules.

"*s*=>(*l* - *s*)" if $\frac{\text{Support\_count (l)}}{\text{Support\_count (s)}} \geq \text{min\_conf},$

where min_conf, is the minimum confidence threshold. Since the rules are generated from frequent itemsets, each one automatically satisfies minimum support. Frequent itemsets can be stored ahead of time in has tables along with their counts so that they can be accessed.

3.2 Application Domain with Apriori Algorithm

The Textile data is stored in the transaction database and the transactions to be mined are extracted from database first. The transactional records include the textile Identifier with their age and salary level, their gender's choices. Then the multi-dimensional association rules can be generated by using Apriori Algorithm. They must satisfy minimum support and minimum confidence. Rules support and confidence are two measures of rule interestingness. Association Rules are considered interestingness if they satisfy both a minimum support threshold and minimum confidence threshold. In the first creation of the algorithm, each Textile item in the Transactions is a member of the set of candidate 1- itemset, C1 (Previous Session 3). The Apriori Algorithm simply scans all of the transactions in order to count the number of occurrences of each item. If minimum support count required is 2, that is min_sup=2. The Apriori property is that all subsets of a frequent itemsets must also be frequent.

3.3 Our Approach for Quantitative Association Rules Mining

Association rules that involve two or more dimensions or predicates can be referred to as multidimensional association rules. Multidimensional association rules with no repeated predicates are called inter-dimensional association rules. For example,
age (X, "21 ..25") ^ gender (X, "အမျိုးသမီး") ^ income (X, "21000 ..40000") → buys (X, "ပိုးခြဲစ်")
Multidimensional association rules with repeated predicates are called hybrid-dimension association rules.
age (X, "21 ..25") ^ gender (X, "အမျိုးသမီး") ^ income (X, "21000 ..40000") ^ buys (X, "ပိုးခြဲစ်") → buys (X, "လိနင်ချည်") .

Quantitative association rules are multidimensional association rules. If a rule describes association between quantitative items or attributes, then it is quantitative association rules.

In these rules, quantitative values for items or attributes are partitioned into intervals. The following rule is an example of a quantitative association rule, where X is a variable representing a customer-
age (X, "21 ..25") ^ gender (X, "အမျိုးသမီး") ^ income (X, "21000 ..40000") ^ buys (X, "ပိုးခြဲစ်") → buys (X, "လိနင်ချည်").

Attributes used for quantitative association rules mining of textile store are –
Quantitative Attributes
- Age
- Income

Categorical Attributes
-Gender

Association rules for objects with quantitative require the discrimination of these attributes to limit the size of the search space. Equi-depth partitioning gives the minimum loss of information. Equi-depth partitioning is used for quantitative association mining of textile store data.

3.4 The overall process of Quantitative Association Rules Mining for Textile store data

Step 1: Get transactions from database.
Step 2: Determine the number of partitions for each quantitative attribute.
Step 3: For categorical attributes, map the values of the attribute to a set of consecutive identifiers. For quantitative attributes that are not partitioned into intervals, the values are mapped to consecutive identifiers such that the order of the values is preserved. If a quantitative attribute is partitioned into intervals, the intervals are mapped to consecutive identifiers, such that the order of the intervals is preserved. From this point, the algorithm only sees values for quantitative attributes. That these values may represent intervals is transparent to the algorithm.
Step 4: Accept minimum support count.
Step 5: Find frequent itemsets using Apriori algorithm.
Step 6: Generate quantitative association rules.

Equi-depth partition for income with depth 20. They are grouped into identifier in salary dimensional table.

- ☐ 21000 - 40000
- ☐ 41000 – 60000
- ☐ 61000 – 80000
- ☐ 81000 – 100000

Equi-depth partition for age with depth 5. They are grouped into age dimensional table.

- ☐ 16 – 20

- ☐ 21 – 25
- ☐ 26 – 30
- ☐ 31 – 35
- ☐ 36 – 40

3.5 System Overview

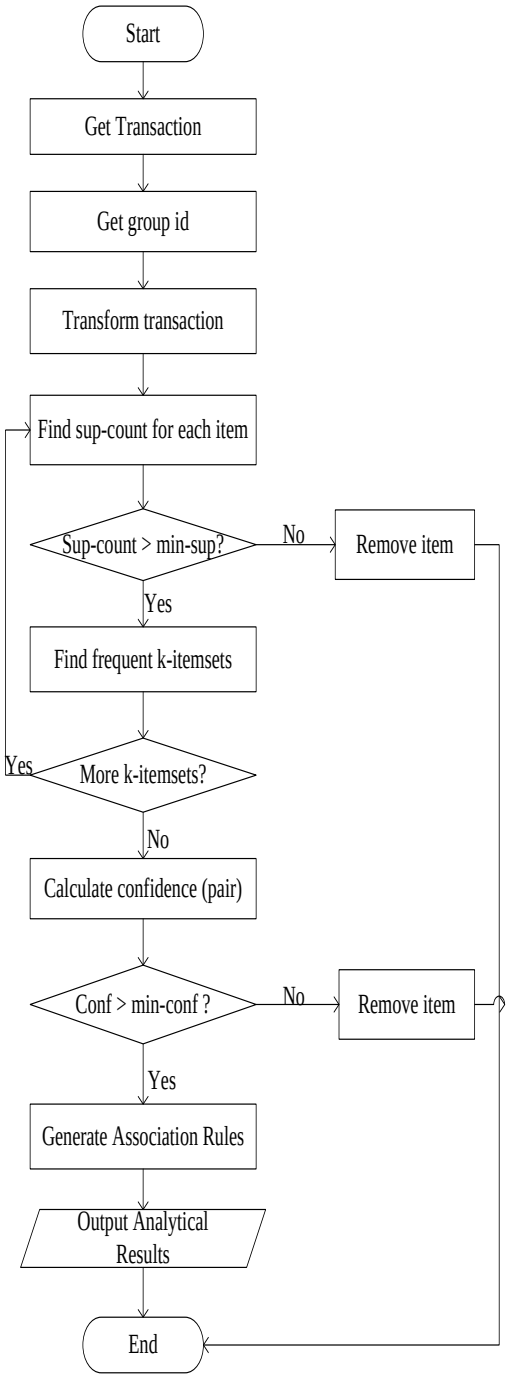

Figure 1. System Overview

The overview of the system can be seen in figure 1. The implementation of this system collects the quantitative information from the customers' age, salary and gender by their vouchers in the transaction database. Their information is grouped into the dimensional groups. Get Group Id takes the dimensional group id discussed in previous session. The minimum support count and minimum confidence threshold values can be set up by the market analyst of the Textile Store according to their requirements analysis to examine the which groups of customer are likely to purchase which items together on a given trip to the store.

The results may be used to plan marketing or advertising strategies, as well as catalog design. For instance, market basket analysis may help managers design different store layouts. In one strategy, items that are frequently purchased together can be placed in close proximately in order to further encourage the sale of such items together.

4. Implementation and Testing

The implementation of the testing is presented in this session. This system accepts the input quantitative values as shown in figure 2.

For creating voucher number, this system lets staff to generate auto voucher number by clicking Auto Voucher No. CheckBox. Customer's Name, Age range, Gender, Income Range, Textile data and quantity can be chosen from corresponding list boxes. The total amount will be automatically can be seen in Total Amount Text Box. Staff can insert, update and delete the sales transactions data by clicking Control Buttons. For searching previous, next, first and last records, staff can click corresponding records' control buttons.

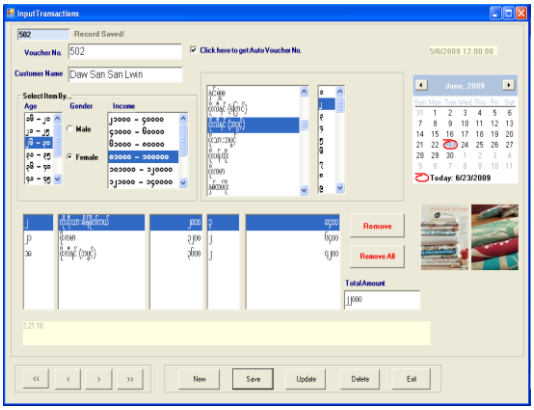

Figure 2. Input Transactions

In the Textile Store association rules generation dialog, users can choose the start date and end date of the transactions according to the decision making requirements. These filtered transactional records can be generated. The association can be generated by entering minimum support count and confidence

threshold. The process of checking and analyzing for mining frequent pattern is as shown in the following figure 3. Figure 4 shows the association rules result from textile store database.

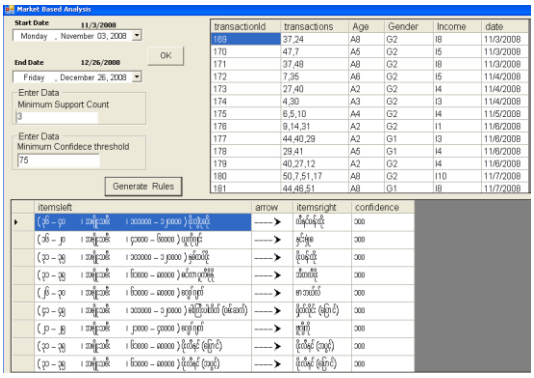

Figure 3. Generating Rules

No.	Left Rules		Right Rules	Confidence	Date
၁	(၂၆ – ၃၀ ။ အမျိုးသမီး ။ ၆၁၀၀၀– ၈၀၀၀၀) ကျော်ဂျက်	---➤	မာဘယ်လ်	၁၀၀.၀၀	၅။၁၈။၂၀၀၉
၂	(၁၆ – ၂၀ ။ အမျိုးသမီး ။ ၄၁၀၀၀– ၈၀၀၀၀) ယူကိုဂျင်း	---➤	နှင်းမုံစု	၁၀၀.၀၀	၅။၁၈။၂၀၀၉
၃	(၃၁ – ၃၅ ။ အမျိုးသမီး ။ ၁၀၁၀၀၀ – ၁၂၀၀၀၀) ပိုးပန်းထိုး	---➤	နှစ်ထပ်ပိုး	၇၅.၀၀	၅။၁၈။၂၀၀၉
၄	(၃၁ – ၃၅ ။ အမျိုးသမီး ။ ၁၀၁၀၀၀ – ၁၂၀၀၀၀) နှစ်ထပ်ပိုး	---➤	ပိုးပန်းထိုး	၁၀၀.၀၀	၅။၁၈။၂၀၀၉
၅	(၃၁ – ၃၅ ။ အမျိုးသမီး ။ ၆၁၀၀၀– ၈၀၀၀၀) စင်ကာပူကီမိုနို	---➤	အီတလီပိုး	၁၀၀.၀၀	၅။၁၈။၂၀၀၉
၆	(၂၆ – ၃၀ ။ အမျိုးသမီး ။ ၆၁၀၀၀– ၈၀၀၀၀) ချည်ပိုး	---➤	ဗီယက်နမ်ဂျက်ကပ်	၇၅.၀၀	၅။၁၈။၂၀၀၉
၇	(၂၆ – ၃၀ ။ အမျိုးသမီး ။ ၆၁၀၀၀– ၈၀၀၀၀) ဗီယက်နမ်ဂျက်ကပ်	---➤	ချည်ပိုး	၁၀၀.၀၀	၅။၁၈။၂၀၀၉
…	…	…	…	…	…

Figure 4. Quantitative Association Rules

5. Conclusion

Data mining deals with the processing of large and complex data. This thesis intends to implement the quantitative association rules mining for effective decision making process. The advantage of the system is that it gives the decision maker to get the more knowledge about market condition and customer group characteristics and association rules. This system applied the datasets of relational database containing both quantitative and categorical attributes. This system dealt with quantitative attributes by fine-partitioning values but system doesn't allow the adding of dimensional attributes dynamically.

References

[1] B. Goethals and M.J. Zaki, "Advances in frequent itemset mining implementations": IEEE ICDM Workshop on Frequent Itemset Mining Implementations, Nov 2003.

[2] H. Liu, "Market Basket Analysis and Itemsets APRIORI Efficient Association Rules, Multilevel Association Rule, Post-processing", CSE 572, CBS 598.

[3] J. Han, M. Kamber, "Data Mining Concepts and Techniques", Acadamic Press, USA, 2001.

[4] M.H. Margahny and A.A. Shakour, "Scalable Algorithm for Mining Association Rules", Dept. of Computer Science, Faculty of Computers and Information, Assuit University, Egypt.

[5] R. Agrawal, T. Imielinski, and A. Sawmi, "Mining association rules between sets of items in large databases." In proc. of the ACM SIGMOD Conference on Management of Data, May 1993, pages 207-216.

[6] R. Agrawal, R. Srikant, "Fast Algorithms for Association Rule Mining" In VLDE 94 , Chile Sept, 1994.

[7] R. Srikant, R. Agrawal, "Mining Quantitative Association Rules in Large Relational Tables", IBM Almaden Research Center 650 Harry Road, San Joe, CA 95120.

[8] Shi-Guang Ju "Mining conditional hybrid-dimension association rules on the basis of multi-dimensional transaction database, an Machine Learning and Cybernetics", 2003 International Conference on Volume 1, Issue 2-5 Nov 2003 Page(s): 216-221 Vol.1.