INTERNATIONAL JOURNAL OF ADVANCE RESEARCH AND DEVELOPMENT

(Volume 5, Issue 5)

Available online at: www.ijarnd.com

Vertical data format based association rule mining for market analysis

Thin Thin Win[1], Thin Thin Htwe[2]

[1]*University of Computer Studies, Mandalay, Myanmar*

[2]*Computer University, Panglong, Myanmar*

ABSTRACT

The association rule mining is one of the primary sub-areas in the field of data mining. The technique has been used in numerous practical applications, including customer market analysis. The discovery of interesting association relationships among huge amount of business transaction records can help in many business decisions making processes. With massive amount of data continuously being collected and stored in databases, many companies are becoming interested in mining association rules from their databases to increase their profits from large amount of transaction data. This system is intended to develop a system for market basket analysis on Store which will generate strong association rules among itemsets with use of vertical data format. The processing times of vertical data format and horizontal data format are also measured and compared in this paper.

Keywords— *Market Basket Analysis, Association Rule Mining, Vertical Data Format*

1. INTRODUCTION

The discovery of association relationships among huge amount of data is useful in selective marketing, decision analysis and business management. Association Rule is one of the major techniques or tasks in data mining, which can be simply defined as finding interesting rules from large collections of data. The main task of association rule discovery is to extract frequent itemsets from market basket data and to generate association rules from these frequent itemsets. Market basket data analysis has been addressed in mining association rules for discovering the set of large items. Large items refer to frequently purchased items among all transactions and each transaction is represented by a set of items purchased. Typical business decisions for the management of the supermarket has to make what to put on sales, how to design coupons, how to place merchandise on shelves in order to maximize the profit, etc. Analysis of pass transaction data is a commonly used approach in order to improve the quality of such decisions. A transaction typically consists of items bought together at the same point of time, but it may consist of items bought by a customer over a period of time.

Association rule mining has two advantages to the business organization after applying the basket analysis: (i) it helps customers to get all the related items from one place and save their time from visiting different places of the store, (ii) it helps organization in more selling of items by placing items closer that are sold together.

In this paper, this system is intended to mine association rule from frequent itemset by using vertical data format. Mining association rule using vertical format has shown to be very effective and usually outperform horizontal approaches because frequent pattern can be countered via transaction id (TID) set intersections in the vertical approach.

2. RELATED WORK

I. Dagan. Kdt and R. Feldman presented how to mine the association rules in temporal document databases and strategies for association rule in temporal document databases. Frequent itemset generation is the generation all sets of items that have support greater than or equal to a certain threshold, called min-support [1]. From the frequent itemsets, generate all association rules that have confidence greater than or equal to a certain threshold called min-confidence [6]. The main goal of above these papers [1, 6] is to find association between two sets of products in the transaction database such that the presence of products in one set implies the presence of the products from the other sets. The importance of discovered patterns depends on statistical measures like support and confidence which are usually computed by the mining application. M. J. Zaki presented how frequent itemsets can also be mined efficiently using vertical data format, which is the essence of the equivalence class transformation algorithm [5]. It is necessary to look at data from different angles to help in making the best decision. Specialized type of data analysis developed to enhance the business decision process [4]. Based on the works mentioned above, in this paper, mining association rules are induced by using vertical data format.

3. THEORETICAL BACKGROUND

3.1. Market basket analysis

Association rule mining searches for interesting relationships among items in a given data set. To learn more about the buying habits of the customers, the shop owner wonders which groups or sets of items are customers likely to purchase on a given trip from the store. To answer the question, market basket analysis may be performed on the retail data of customer transactions at the store. The results may be used to

plan marketing or advertising strategies, as well as catalog design. In one strategy, items that are frequently purchased together can be placed in close proximity in order to further encourage the sale of such items together. Market basket analysis can also help retailers plan which items put on sale at reduced prices. This process analyzes customer buying habits by finding associations between the different items that customers place in their "shopping baskets" [3].

3.2. Association rule mining

Association rule mining is a data mining task that discovers relationships among items in a transaction database. The goal of association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from a given database [7]. Association rule mining is commonly used in market basket analysis to find items frequently bought together by shoppers.

In general, the association rule mining can be viewed as a two-step process:

(a) Find all frequent itemsets: Each of the itemsets will occur at least as frequently as a pre-determined minimum support count.

(b) Generate strong association rules from the frequent itemsets: Rules must satisfy minimum support and minimum confidence [2].

3.2.1 Mining frequent itemsets using vertical data format

Data can be presented in item-TID_set format (that is, {item: TID_set}), where item is an item name, and TID_set is the set of transaction identifiers containing the item. This format is known as vertical data format.

In the process of mining frequent itemsets by exploring the vertical data format, the horizontally formatted data is first transformed to the vertical format by scanning the data set once. The support count of an itemset is simply the length of the *TID_set* of the itemset. Starting with *k=1*, the frequent *k*-itemsets can be used to construct the candidate *(k+1)* –itemsets which satisfy the minimum support. The computation is done by intersection of the *TID_sets* of the frequent *k*-itemsets to compute the *TID_sets* of the corresponding *(k+1)*-itmesets. This process repeats, with *k* incremented by 1 each time, until no frequent itemsets or no candidate itemsets can be found. In the generation of candidate *(k+1)*-itemsets from frequent *k*-itemsets, there is no need to scan the database to find the support of *(k+1)* itemsets (for $k \geq 1$). This is because the *TID_set* of each *k*-itemset carries the complete information required for counting such support [2].

3.2.2 Association rules

Let $I = \{I_1, I_2, \ldots, I_m\}$ be a set of items. Let D, the task-relevant data, be a set of database transactions where each transaction T is a set of items such that $T \in I$. Each transaction is associated with an identifier, called TID. Let A be a set of items. A transaction T is said to contain A if and only if $A \subseteq T$. An association rule is an implication of the form A => B, where $A \subset I$, $B \subset I$ and $A \cap B = \phi$.

The rule A => B holds in the transaction set D with support s, where s is the percentage of transactions in D that contain $A \cup B$. This is taken to be probability, P ($A \cup B$). The rule A => B has confidence c in the transaction set D if the c is the percentage of transactions in D containing A that also contain B. This is taken to be the conditional probability, P (B\A). That is,

Support (A => B) = P ($A \cup B$)
Confidence (A => B) = P (B\A)

The quality of association rules is evaluated by looking at their support and confidence [2].

3.2.3 Strong association rule

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

$$\text{Confidence } (A \Rightarrow B) = \frac{\text{Support\_count } (A \cup B)}{\text{Support\_count } (A)}$$

Where support_count ($A \cup B$) is the number of transactions containing the itemsets $A \cup B$, and support_count (A) is the number of transactions containing the item set A [3].

4. PROPOSED SYSTEM ARCHITECTURE

Firstly, the user can update transaction data. The user must enter the different minimum support thresholds to find frequently occurring itemsets. The system then compares each of them with minimum support count until no more frequent items can be found using vertical data format. Once all the frequent itemsets are found, then the system generates association rule. Figure 1 shows the process of the proposed system architecture.

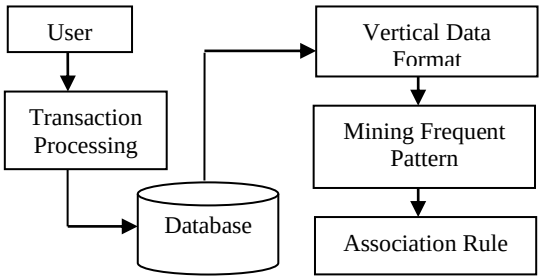

Fig. 1: Proposed system architecture

4.1. System flow diagram

The implemented system develops a method for analyzing buying pattern of the customers in the market. This system is used to enter the updated transaction data into the database. Thus, the system analyzes this transaction database and then generates vertical data format.

This system is to find out which items are commonly purchased together in order to make some selected frequent customers special bundle-offers which are likely to be in their interest. The concept of association rules can be used to detect relations between items.

Association rule searches for interesting relationships among items. These are step by step processing to generate association rule. Firstly, support count for each item is found. Then, it is compared with minimum support count. Items less than minimum support count is removed and others are going on processing. The user then again compares each of them with minimum support count and removes pairs which are less than minimum support count. After finishing these processing, the user can go on to generate association rules.

Finally, the rules are examined whether they are strong. The rules, having equal to or greater confidence than user specified one, are considered to be strong. Strong rule implications are given and other are discarded. Figure 2 shows system flow diagram.

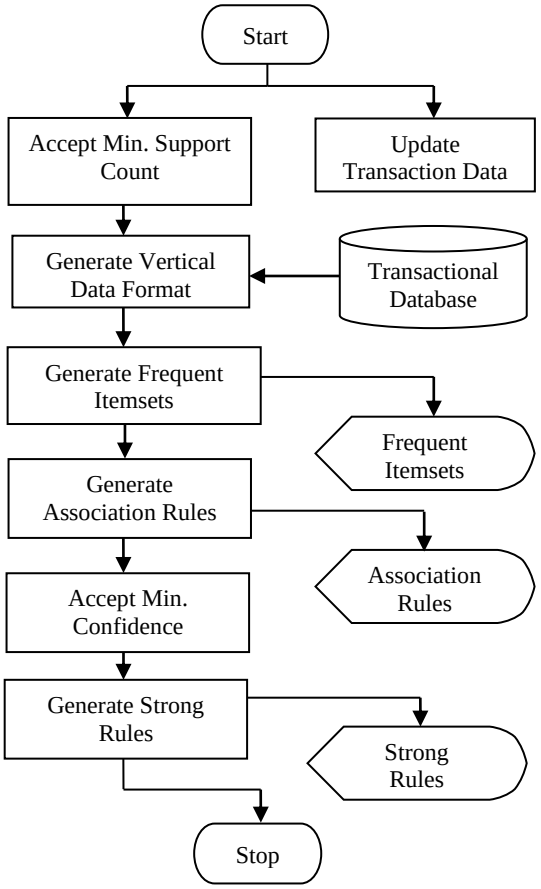

Fig. 2: System flow diagram

4.2. Explanation of the System

Let D be a database of transaction. Each transaction consists of a transaction identifier and a set of items $\{i_1, i_2,\ldots, i_n\}$ selected from the universe I of all possible descriptive items. Table 1 shows the transaction data of customer buying items.

Table 1: Transaction data

TID_set	Itemset
1	Biscuits, Bread, Cheese, Yogurt, Sugar
2	Bread, Cheese, Coffee, Sugar
3	Cheese, Chocolate, Donut, Milk, Sugar
4	Bread, Cheese, Coffee, Cereal, Juice
5	Chocolate, Donut, Juice

There are five transaction in this database, that is, D = 5. In the process of mining frequent itemsets by using vertical data format, the horizontally formatted data must be first transformed to the vertical format by scanning the data set. The support count of an itemset is the length of the TID_set of the itemset. Suppose that the minimum transaction support count is 3.

Table 2: The 1-Itemsets in vertical data format

Itemset	TID_set
{Bread}	{1, 2, 4}
{Cheese}	{1, 2, 3, 4}
{Sugar}	{1, 2, 3}

Table 2 shows the 1-itemsets in vertical data format. Initially, every item is considered as a candidate 1-itemset in above transaction. After counting their support, the candidate itemsets which appear in fewer than three minimum transaction support are discarded.

Table 3: The 2-Itemsets in vertical data format

Itemset	TID_set
{Bread, Cheese}	{1, 2, 4}
{Cheese, Sugar}	{1, 2, 3}

Table 3 shows the 2-itemsets in vertical data format. In the next iteration, candidate 2-itemsets are generated. There is no frequent itemsets at this process, iterated procedures are terminated. Thus, the frequent 2_itemsets {Bread, Cheese} and {Cheese, Sugar} have been received with four rules.

Table 4: Association rule

Subset (A)	Subset (B)	Sup(AUB) / Sup (A)	Confidence (%)
{Bread}	{Cheese}	3/3	100
{Cheese}	{Bread}	3/4	75
{Cheese}	{Sugar}	3/4	75
{Sugar}	{Cheese}	3/3	100

Table 4 shows association rule. After finishing these processing, association rules come out as output analytical result.

Table 5: Strong rule

Subset (A)	Subset (B)	Sup(AUB)/ Sup (A)	Confidence (%)
{Bread}	{Cheese}	3/3	100
{Sugar}	{Cheese}	3/3	100

Table 5 shows strong rule. Suppose that the minimum confidence is 100%. As a result, each user is allowed to enter minimum confidence to produce strong rule. Since, the rules, having equal to or greater confidence than use specified one, are considered to be strong. Finally, reasonably strong rule implications are given and others are discarded.

4.3. Experimental Result of the System

This system is implemented for the analysis of transaction using association rule mining by analyzing the itemsets pairs that likely to happen for future sales transactions. According to support and confidence, this system generates association rules that are used to produce the results of analysis report by using vertical data format. Mining frequent itemsets using vertical data format is better than horizontal data format in processing time because this format does not need to scan the database to find the support. Table 6 shows processing time using vertical and horizontal data format.

Table 6: Processing time using vertical and horizontal data format

Minimum Support Count	Vertical Data Format (milliseconds)	Horizontal Data Format (milliseconds)
2	11	35
3	25	60
4	34	90
5	45	125
6	55	150

Figure 3 shows comparison of processing time between vertical and horizontal data format of the dataset.

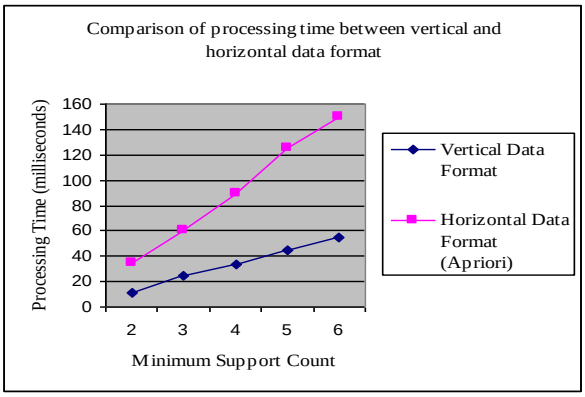

Fig. 3: Comparison of processing time between vertical and horizontal data format

5. CONCLUSION

In this paper, a system for frequent itemset mining and association rule mining is implemented on the basis of vertical data format. This system is intended to use in finding out which items are commonly purchased together in a store. The system can also evaluate support and confidence (%) to produce strong rules. Moreover, the processing times of vertical data format and horizontal data format are also measured and compared in this paper. According to the experimental results, it is obviously seen that the processing time of vertical data format is always faster than the processing time of horizontal data format. As a result, it can be applied to the application with theoretical data mining of Market based analysis and any real-world transactional databases.

6. REFERENCES

[1] I. Dagan. Kdt and R. Feldman, "Knowledge Discovery in Texts", *In Proceedings of the First International Conference on Knowledge Discovery*, 1995.

[2] J. Han and M. Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann, 2001.

[3] J. Han and M. Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann, 2000.

[4] L. Barry, "Using the Data Warehouse for Decision Support", 1998.

[5] M. J. Zaki, "Knowledge and Data Engineering", 2000.

[6] T.M. Mitchell, "Machine learning", MC Graw Hill, New York, 1997.

[7] U. Fayyad, G. Piatesky-Shapiro and P. Smyth, "From data mining to knowledge discovery in databases", *In: AAAI Magazine*, 1996.

