Performance Comparison of Data Mining Algorithms

THIDA AUNG¹, MAY ZIN OO²

¹Dept of Information Technology, Mandalay Technological University, Mandalay, Myanmar, Email: thidaung22@gmail.com.
²Dept of Information Technology, Mandalay Technological University, Mandalay, Myanmar.

Abstract: Nowadays, association rule mining 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 decision 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. So, this system is intended to develop a system for market basket analysis on Electronic Shop which will generate association rules among itemsets with the use of ECLAT (Equivalence CLASS Transformation) and Apriori algorithms. The system is also intended to display the relation between items by finding frequent itemsets of the database. According to the interestingness measures, such as support and confidence, this system can also support the decision making process for a market expert. Moreover, the processing time of ECLAT and Apriori algorithms is also measured and compared in this system. This system is implemented by using C# and Microsoft Access Database.

Keywords: ECLAT, Apriori and Association Rule.

I. INTRODUCTION

A great deal of business transaction data implicit much of useful knowledge for business decision, but association rule mining method finds the interesting association or correlation relationships among a large set of data items. With massive amounts of data continuously being collected and stored, a new research subject arise how interesting association relations can be found out of a large quantity of business transaction records to help make commercial decisions such as catalogue design, cross-marketing and loss-leader. Association rule is one of the most researched areas of data mining and has recently received much attention from the database community. The process of finding association rules has two separate phases. In the first phase, find all combinations of items that have transaction support above the minimum support count. In the second phase, use the frequent item sets to generate the desired rules. Most of the previous algorithms are based on the traditional horizontal database format for mining. In vertical database each item is associated with its corresponding transaction id (TIDset). Mining algorithms using the vertical format have shown to be very effective and usually outperform horizontal approaches because frequent itemsets can be countered via TIDset intersections in the vertical approach.

This system is mined the frequent itemsets on the transaction data of Electronic Shop by using ECLAT and Apriori algorithms and then the important decisions are made by applying strong association rule. Moreover, this system intends to compare Apriori (horizontal data format) and ECLAT (vertical data format) for sale analysis system. Electronic Shop is promoted sales and developed by using this system. The purposes of the Market analysis system are as follows:

• To mine association rules from frequent item sets of Electronic shop.
• To guide the mining procedure to discover the interesting associations.
• To help retailers, buyers, planners, merchandisers, and store managers to plan more profitable advertising and promotions, attract more customers and increase the value of the market basket.

The paper is organized as follows. In Section II, we define the related work. In Section III, we introduce background theory which includes data mining, mining association rule and algorithms of ECLAT and Apriori. In Section IV, we discuss proposed system with diagram and explanation of the system with examples. We conclude this proposed system in Section V.

II. RELATED WORK

R. Srikant and R. Agrawal [6] proposed the algorithm for mining frequent itemsets for boolean association rules. Apriori employs an iterative approach known as level wise search, where k-itemsets are used to explore (k+1)-itemsets. The set of frequent 1-itemsets is found by scanning the database to accumulate the count for each item and collecting these items that satisfy minimum support. M. J. Zaki [4] presented how frequent itemsets can also be mined efficiently using vertical data format, which is the essence of the equivalence class transformation algorithm. 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. G. Grahe and J. Zhu [6] presented a novel array-based technique that greatly reduces the time to spend traversing FP-tree. Furthermore, they also presented new algorithms for mining maximal and closed frequent item sets.

### III. BACKGROUND THEORY

This system is implemented to analyze the transaction data from Electronic Shop by using ECLAT and Apriori algorithms within association rule mining. And then, this system compared the performance of these two algorithms.

#### A. Market Basket Analysis

Market basket analysis may be performed on the retail data of customer transactions at your store. This process analyzes customer buying habits by finding associations between the different items that customers place in their “shopping baskets”. The discovery of such associations can help retailers develop marketing strategies by gaining insight into which items are frequently purchased together by customers. In a supermarket with a large collection of items, typical business decisions that the management of the supermarket has to make include what to put on sale, how to design coupons, how to place merchandise on shelves in order to maximize the profit, etc. Analysis of past transaction data is a commonly used approach in order to improve the quality of such decisions [3].

#### B. Benefits of Association Rule

The most famous application of association rules is its use for market basket analysis. A supermarket setting is considered where the database records items purchased by a customer at a single time as a transaction. The planning department may be interested in finding “associations” between sets of items with some minimum specified confidence. Such associations might be helpful in designing promotions and discounts or shelf organization and store layout. However, association rules have many other fields in which it has been helpful. Association rules mining is used in the telecommunications and medical fields for performing partial classification. This type of mining has been also used on other typed of data sets. It has been used to mine web servers log files to discover the patterns that access different resources consistently and occur together or the access of a particular place occurring at regular times [9].

#### C. Equivalence Class Transformation (ECLAT)

In the ECLAT (Equivalence CLASS Transformation), mining frequent patterns from a set of transactions in item-TID-set format (that is, {items: 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. First, transform the horizontally formatted data to the vertical format by scanning the data set once. Mining can be performed on this data set by intersecting the TID-sets of every pair of frequent single item. The support count of an
Performance Comparison of Data Mining Algorithms

The previous pass for each transaction t ∈ D do
1. C = cover (\{i\}) \cup \text{cover (\{j\})}
2. if |C| ≥ s then
3. D' = D \cup \{j, C\}
4. end if
5. end for
6. //Depth-first recursion
7. Compute F[I \cup \{i\}] \cup D'
8. end for

D. Apriori

Apriori is a classic algorithm for frequent item set mining and association rule learning over transactional databases [10]. Apriori algorithm is based on the fact that the algorithm uses prior knowledge of frequent itemsets properties. This technique uses the property that any subset of a large itemset must be a large itemset. Apriori generates the candidate itemsets by joining the large itemsets of the previous pass and deleting those subsets which are small in the previous pass without considering the transactions in the database. An association rule is valid if its confidence and support are greater than or equal to corresponding threshold values [2]. Apriori employs an iterative approach known as a level-wise search, where k-itemsets are used to explore (k+1)-itemsets. First, the set of frequent 1-itemsets is found. This set is denoted L_1. L_1 is used to find L_2, the frequent 2-itemsets, which is used to find L_3, and so on, until no more frequent k-itemsets can be found. The finding of each L_k requires one full scan of the database [3].

Apriori Algorithm: This algorithm is as follows:
Input: Database, D, of transactions; minimum support threshold, min_sup.
Output: L, frequent itemsets in D.
Method:
1. L_1 = find_frequent_1_itemsets (D);
2. for (k=2; L_{k-1} ≠ \emptyset; k++)
3. {
4. C_k = apriori_gen (L_{k-1}, min_sup);
5. for each transaction t ∈ D

IV. SYSTEM DESIGN

The proposed system design, the implementation of the system and experimental results of this system are described in this section.

A. Proposed System Design

Figure 2. Proposed System Design.
The overall proposed system design is shown in Figure 2. The proposed system is implemented to find out which items are commonly purchased together within the Electronic Shop in order to make some selected frequent customers special bundle-offers which are likely to be in their interest. This system searches the interesting relationships among items by using ECLAT and Apriori algorithms. These are step by step processing to generate association rule. Firstly, this system analyzes the transaction database. Second, 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 go on processing. And then, this system can again compare each of them with minimum support count and remove pairs which are less than minimum support count. After finishing these processing, this system produces association rule which is generated by using ECLAT and Apriori algorithm. The rules having equal to or greater confidence than user specified one are considered to be strong association rule. And then, this system compares the processing time as the performance of ECLAT and Apriori algorithms. Finally, this system displays the comparison result of these two algorithms.

B. Implementation of the Proposed System

This system is implemented by using Microsoft Visual Studio 2010, C# programming language and Microsoft Access Database.

1. Transaction Processing: At first, this system imports the transaction data into the system. In this system, the user can choose any desired Microsoft Access Database file as the transaction data. Transaction processing is shown in Figure 3.

2. Generate Association Rule by using ECLAT Algorithm:

In the ECLAT algorithm, this system initially converts from the horizontally formatted data (\{TID: item_set\}) to the vertical format (\{item: TID_set\}) by scanning the data set once. And then, this system searches the support count for each item. After counting their support, the itemsets which is less than minimum support count are discarded. And then, this system generates each frequent itemsets which is equal to and greater than minimum support count from the transaction. After finishing the iterative process, this system generates the association rule. Association rule by using ECLAT algorithm is shown in Figure 4.

3. Generate Association Rule by using Apriori Algorithm:

In the Apriori algorithm, each item is a member of the set of candidate 1-itemsets, \(C_1\) in the first iteration. This system scans all of the transactions in order to count the number of occurrences of each item. This system compares candidate support count with user-defined minimum support count. And then, this system determines the set of frequent 1-itemsets. In the next iteration, this system scans the transactions in database and accumulates the support count of each candidate itemset in \(C_2\). This system continues iterative processing. If there is no more frequent itemsets, this system produces the association rule. Association rule by using Apriori algorithm is shown in Figure 5.

4. Performance Comparison:

This system compares the performance results of ECLAT and Apriori algorithms. From their comparisons, this system proves that the ECLAT performs better than the Apriori algorithm. Performance comparison result is shown in Figure 6.
Performance Comparison of Data Mining Algorithms

Figure 6. Performance Comparison Result.

V. EXPERIMENTAL RESULTS

This system is tested by using 1000 transactions from the Electronic Shop. This system is proposed 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 by using ECLAT and Apriori algorithms. These generated association rules are used to produce the results of analysis report. Mining frequent itemsets using ECLAT algorithm is better than Apriori algorithm in processing time because ECLAT algorithm does not need to scan the database to find the support. Figure 7 shows processing time of ECLAT and Apriori algorithms by changing the various minimum support count.

Figure 7. Comparison of Processing Time by using ECLAT and Apriori Algorithms.

VI. CONCLUSION

In this system, association rule mining is implemented on the basis of ECLAT and Apriori algorithms. Moreover, the processing times of ECLAT and Apriori are also measured and compared for Electronic sale analysis system to ascertain which algorithm is more effective. According to the experimental results, the processing time of ECLAT is always faster than the processing time of Apriori because ECLAT algorithm does not need to scan the database to find the support. Therefore, this system provides the decision maker to give useful information about interesting items. This system is also a provider of several devices and business organizations. The system is implemented by collecting real data from Electronic Shop. Therefore, this system can also support this electronic shop manager who can place the related devices together and advice the customer for the best price and the latest updates.

VII. ACKNOWLEDGMENT

The author would like to express sincere appreciation to the Rector of Mandalay Technological University for kind Permission to prepare for this paper. The author would also like to give special thanks to Dr. Aung Myint Aye, the Head of Department of Information Technology, Mandalay Technological University (MTU). The author is deeply grateful to Dr. May Zin Oo and all teachers in our Department and all who willingly helped the author throughout the preparation of the paper. This paper is dedicated to the author’s parents for continual and full support on all requirements and moral encouragement.

VIII. REFERENCES