Finding Association Rules from Sales Data with Apriori Algorithm

Shwe Sin Thein
University of Computer Studies, Mandalay
shwesintheinn@gmail.com

Abstract

The association rule mining is one of the primary sub-areas in the field of data mining . This type of mining ,the association rule searches for interesting relation among item in a given data set , has been used in numerous practical applications : market basket analysis , catalog design and loss-leader analysis . In this paper, Apriori algorithm is implied for mining frequent itemsets for Boolean association rules with a large amount of items in a database. Consequently, every itemset is employed with level-wise search to provide achieving more frequent itemsets found. Furthermore , the user is eventually allowed to understand this system with two step process including join and prune actions. As a result, the valuable information is allowed to the user with a variety of threshold value. The association rule resulted from Apriori algorithm is also improved into strong rule according to these threshold values.

1. Introduction

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 [1]. 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 Apriori requires the full scan of the database for every itemset and it may take a very long for the Apriori with a hashing scheme that is used in each large database . Apriori technique can be used to reduce the size of candidate itemsets , when scanning each transaction in the database to generate the frequent itemsets. It is based on the support and confidence framework.

Market basket data analysis has been well- addressed in mining association rules for discovering the set of large items. Large items refer to frequently purchased items among all transactions and 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 [5].

The structure of this paper is organized as follows, section 2 is introduced the related work of rule mining. Section 3 is explained the principles of the single level association rules and Apriori technique. Section 4 is organized the design of the system and Market implementation including a data set is also presented in section 5.Finally, conclusion is described in section 6, respectively.

2. Related work

Direct Hashing and Pruning (DHP) algorithm is in fact a variation of the well-known Apriori algorithm. In the DHP algorithm, a hash table is used in order to reduce the size of the candidate itemsets. Iteration to prune some candidates before the database passes. It prunes candidates by using dynamically generated hash tables, thus reducing the number of database blocks read [4]. The system develops to generate the interesting patterns from the transaction database. It is based on the support-confidence framework. Mining data that has been collected from web server log files is not only useful for studying customer choices, but also helps to better organize web pages. Then interesting rules are mined by using Apriori algorithm. Usage mining is the application of data mining techniques

to discover usage patterns from web data. The importance of discovered patterns depends on statistical measures like support and confidence which are usually computed by the mining application.

In electronic market system, one of the best known data mining techniques is the discovery of association rules. The main goal of these rules 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 set. Apriori, Tree Projection algorithms and the FP tree algorithms are some of well-known algorithms for finding association rules from databases [2]. Identifying all the data required to make a decision gathering it together organized as meaningful information. Specialized type of data analysis developed to enhance the business decision process [3].

3. Apriori algorithm

Apriori is a classic algorithm for learning association rules. Apriori is designed to operate on databases containing transactions. As is common in association rule mining , given a set of *itemsets*, the algorithm attempts to find subsets which are common to at least a minimum number C of the itemsets .

Apriori uses a " bottom up " approach , where frequent subsets are extended one item at a time (a step known as *candidate generation)* , and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found. The Apriori algorithm was implemented in commercial packages. As input, this algorithm uses a table with purchase transactions. Each transaction contains the customer identification and the purchased items, as follow (customer, item). Input parameters are defined as the maximum number of acquired items (max k) and the minimum support (min sup) of a certain basket.

The first step of the Apriori algorithm generates sets of market baskets. I_k is defined as the set of frequent items with k items bought together. Firstly, the algorithm filters the items with a frequency that is higher than the min sup, generating I_1. In the following stages, for each I_k it generates the I_k+1 candidate, such as I_k _I_k+1. For each I_k+1 candidate, the algorithm removes the baskets, which are lower than the min sup. The cycle ends when it reaches I_{max} k.

In the second step, the Apriori algorithm generates sets of market baskets and then generates association rules L $\Rightarrow$ *R*. For each rule, the support measure and the confidence measure are calculated. In order to implement the cross-selling strategy the data analysis choose, firstly, the dimension of the basket, secondly, they choose the rules with the highest support measure. Finally, those with the highest confidence measure are chosen, among those with the highest support measure.

3.1. Generating frequent itemset

A set of items is referred to as an itemset. The occurrence frequency of an itemset is the number of transactions that contains the itemset. This is also known as the frequency , support count ,or count of item set . An itemset satisfies minimum support if the occurrence frequency of the itemset is greater than or equal to the product of min _sup and the total number of transactions in D. The number of transactions required for the itemset to satisfy minimum support, such itemsets are called as frequent itemsets.

There are several ways to reduce the computational complexity of frequent itemset generation

1. ***Reduce the number of candidate itemsets.*** The Apriori principle described in the next section is an effective way to eliminate some of the candidate itemsets without counting their support values .

2. ***Reduce the number of comparisons.*** Remove of matching each candidate itemset against every transaction , the number of comparisons can be reduced by using more advanced data structures, either to store the candidate itemsets or compress the data set .

3.2. Support and confidence

The quality of association rules is commonly evaluated by looking at their support and confidence .The support s, of a rule measures the occurrence frequency of the pattern in the rule. The support gets dividing the number of transactions containing both A and B by number of transactions. The variable s is measured by the fraction of transaction contains both A and B:

tuples_containing_

$$\text{Support (“A}\rightarrow\text{B”)} = \frac{\text{Both A and B}}{\text{total\_\#\_ of tuples}}$$

The confidence is defined as the measure of certainly or trustworthiness associated with each discovered pattern. The confidence c , is the measure of the strength of implication . The variable c is defined to maintain the percentage of transaction in D containing A and B :

$$\text{Confidence (“A}\rightarrow\text{B”)} = \frac{\text{\# tuples\_containing\_ Both A and B}}{\text{\# tuples \_containing A}}$$

Where support-count (A→B) is the number of transactions containing the itemsets A∪B, and support count (A) is the number of transactions containing the itemset A.

3.2.1. Calculating minimum support count. The minimum transaction support count required is obtained by dividing the user specified minimum support percentage by 100 and multiplying the result with number of transactions. For example, the minimum transaction support count 3 is gained by dividing the 40% (user specified support percentage) with 100 and multiplying result (0.4) with number of transaction (6).

3.3. Association rules

Let J= $\{i_1, i_2, i_3, \ldots, i_m\}$ 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⊆J. 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 an only if A⊆T . An association rule is an implication of the form A ⇒B , where A⊂ J, B ⊂ J and A ⇒ B = f .The rule A ⇒ B holds in the transactions set D with supports s ,where s is the percentage of transactions in D that contain A ∪ B (i.e., both A and B). This is taken to be the probability, P (A ∪ B). The rule A ⇒B has confidence c in the transaction set D if c is the percentage of transaction in D containing A that also contain B . This is taken to be the conditional probability , P (B \ A). That is,

$$\textit{Support } (A \Rightarrow B) = P (A \cup B)$$

$$\textit{Confidence } (A \Rightarrow B) = P (B \cup A)$$

3.3.1. Single-dimensional-boolean association rule. It concerns single dimensional, single level and Boolean association rule. The items or attributes in an association rule reference only one dimensional then it is a single dimensional association rule. The items or attributes in an association rule reference only one level then it is a single level association rule. It concerns associations between the presence and absence of items.

3.3.2. 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 :

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

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

For each frequent item set I , generate all non-empty subsets of I . For every non-empty subset of I , output the rule " s ⇒ (I-s) " if $\textit{Support\_count(I)/Support\_count(s)} \geq \textit{min\_conf}$, 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 hash tables along with their counts so that they can be accessed quickly.

4. System design

This system is used to enter the updated sale Items list into the system. Thus , the system analyses this transaction database and then generates the Apriori Algorithm(see Figure 1).

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. To detect relations between data items, the concept of association rules can be used. Association rules aim at detecting uncovered relations between data items. All possible rules are generated since the frequent itemsets are found. The rules are then examined whether they are strong or not based on the minimum confidence. The rules, having equal to or greater confidence than user specified one, are considered to be strong. Finally, reasonably strong rule implications are given and others are discarded. These processes are truly operated through out the former sequence of implementation in Apriori algorithm .

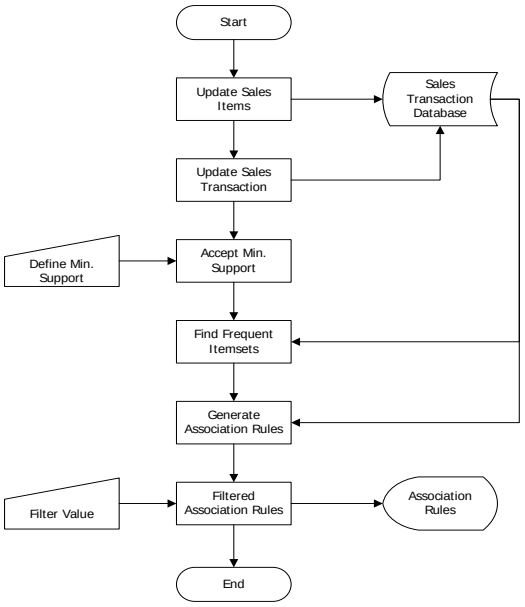

Figure 1: System flow diagram

5. Market implementation

The transactional database may have additional tables associated with it, which contain other information regarding the sale, such as the customer ID number of the sales person and of the branch at which the sale occurred, and so on.

Transaction-ID	Items
1.	Cake, Milk, Bread
2.	Cola, Bread, Eggs
3.	Cake, Cola, Bread, Eggs
4.	Cola, Eggs

Let's look at a concrete example of Apriori , based on a sample transaction database D. There are four transaction in this database , that is ,
D = 4. Suppose that the minimum transaction support count required is 2 .

1-itemsets

Itemset	Support count
{ Cake }	2
{ Cola }	3
{ Bread }	3
{ Eggs }	3

Initially , every item is considered as a candidate 1-itemset in above transaction. After counting their support, the candidate itemsets {Milk} are discarded because they appear in fewer than two minimum transaction support.

2-itemsets

Itemset	Support count
{ Cake, Bread }	2
{ Cola ,Bread }	2
{ Cola, Eggs }	3
{ Bread , Eggs }	2

In the next iteration, candidate 2-itemsets are generated. After counting their supports , the candidate itemsets { Cake , Cola } , {Cake ,Eggs} are discarded because the Apriori principle ensures that all supersets of the frequent 1-itemsets must be frequent.

3-itemsets

Itemset	Support count
{ Cola ,Bread, Eggs }	2

The correctness of the Apriori Pruning strategy can be shown by counting the number of candidate 3-itemsets generated . There is no frequent itemsets at this process , Apriori iterated procedures are terminated . Finally ,the following strong rules are produced by 2^k-2 confidence rules (k is itemset) in this system . Thus, the frequent 3-itemset {Cola , Bread ,Cake } have been received with six rules . The first one

and third value (see Table 1) are strong rules in those data.

Let min-conf = 80 %

(1) Cola, Bread → Eggs
Confidence= (2/2) x100=100%,

(2) Cola, Eggs → Bread
Confidence= (2/3) x100=66.67%,

(3) Bread, Eggs → Cola
Confidence= (2/2) x100=100%,

(4) Eggs → Cola, Bread
Confidence= (2/3) x100=66.67%

(5) Bread → Cola, Eggs
Confidence= (2/3) x100=66.67%

(6) Cola → Bread, Eggs
Confidence= (2/3) x100=66.67%

Table 1. Result for strong rules.

Rule	Sup:	Conf:
Cola , Bread → Eggs	50 %	100%
Bread , Eggs → Cola	50 %	100%

As a result , each user is allowed to enter a filter threshold value to produce strong rule. The implementation results are shown in Figure 2. 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.

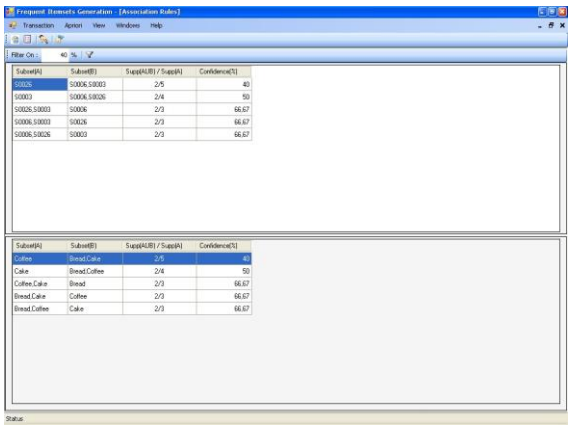

Figure 2: Strong rules production

This system consists of many transactions with itemsets , longest transaction with 16 items. This system uses 200 items as selling data. The memory requirement of the algorithm depends only on the number of distinct items in the database and the minimum support , thus it is independent from the size of the database. The memory requirement of the algorithm decreases as the minimum support increase since with high minimum support , it will only generate fewer results. When the minimum support increases , the number of frequent itemsets decreases ,and as a result of association rules increases .

6. Conclusion

This system is used to implement the association analysis and can be applied to find items that are frequently bought together by customers . The discovered patterns are typically represented in the form of implication rules or features subsets . The rule suggests that a strong relationship exists between the sales of products. Retailers can use this type of rules to help them identify new opportunities for cross-selling their products to the customers . As a result , single level single dimension association rules are generated to provide the market basket area with implemented algorithm .

7. References

[1] A. Savasere, E. Omiecinski, S.Navathe: An Efficient Algorithm for Mining Association Rules in Large Databases, Proc. 21th Int'l Conf. Very Large DataBases, Zurich, Switzerland (1995), pp. 432-444

[2] L. Guichong and Howard J.Hamilton, Department of Computer Science University of Regina , " Basic Association Rules " , S4S OA2, Canada , SK , Regina.

[3] L. Barry: Using the Data Warehouse for Decision Support.http://www.act.ucsd.edu/dw/Forum9806/ index.htm. 1998.

[4] S. Brin , R . Motwani , J . Ullman , S . Tsur : Dynamic Itemset Counting and Implication Rules for Market Basket Data , *Proc of the 1997 ACM SIGMOD International Conference on Management of Data , 1997.*

[5]http://en.wikipedia.org/wiki.Association_rule_mining.