Modern Research Trends in Association Data Mining Techniques

Sonal Vishwakarma  
Information Technology Dept.  
RKDF IST, Bhopal, India

Shrikant lade  
Head of Department  
RKDF IST, Bhopal, India

Abstract
Extracting needful information from the large pool of information is a primary task before predicting. Especially from a huge amount of incomplete, noisy, redundant and randomly scattered data. Data mining provides a framework that automatically discovers required patterns from data set which will be using these predict future occurrence in analogous scenario. Data mining approach modeled and extracts multiplicity of category and various time granularities to fulfill the need of various users or uses. Association rule mining is the core mechanism of data mining to help us. An association rule mining has grown to be central field in modern data mining research context [1]. In this article we have surveyed various techniques of association rule mining and their significance.

Keywords
Association, Apriori, Data Mining, FP, FP Tree, Positive and Negative FP Tree.

1. Introduction
Extracting needful information from the large pool of information is a primary task before predicting. Especially from a huge amount of incomplete, noisy, redundant and randomly scattered data. Here Data mining plays a significant role to specify the model which extracts the knowledge from the large information which will be the decisive next (data). Broadly Data mining approach can be classify into two-description and prediction. In Descriptive type mining characterization of common distinctiveness of data while Predictive is to predict on the basis of the referring of data [10].

Data mining provides a framework that automatically discovers required patterns from data set which will be using these predict future occurrence in analogous scenario. Data mining approach modeled and extracts multiplicity of category and various time granularities to fulfill the need of various users or uses. The author of [31] divides data mining tasks in following (a) conceptual description; (b) correlation analysis; (c) classification and prediction; (d) clustering; (e) outlier analysis; (f) evolution analysis.

When there is situation if this is what then will be. Such situation are significantly used in market analysis, medical clinical analysis of critical diseases those are depends of if then rule along with support and confidence support metrics. In such situation Association rule mining is the core mechanism of data mining to help us. An association rule mining has grown to be central field in modern data mining research context [1]. Still there is a need of enhancement in association rules due to the large number of wrong association rules generated by conventional support-confidence methods.

Originally Rakesh Agrawal et al [2] was anticipated association rule mining concept in 1993. Conventionally it is designed for mining customer transaction relationship among two sets of items, which are in the form of A ⇒ B, in high frequency, high correlation rules [3-8].

Association rule mining is the most popular technique among modern researchers to solve many problem like market analysis (share market as well), medical application, intrusion and detection and prevention system and modern communication network such as WiMax, MANET etc.

In this article we have surveyed various techniques of association rule mining and their significance.

Rest of the paper is organized as follow: section 2 gives brief definition and working of association. Section 3 gives detail about the various methods of Association, section 4 presents the literature survey on association then section reveals the concept FP tree and presents the basic idea on negative and positive association rule mining. And finally section 6 concludes the article.

2. Association rule mining
Author of [9] ha illustrated the working of Association rule mining in article, According to the author defined as discovers the interesting associations among items in a given data set. Suppose a Database T is a collection of m transactions, \{T₁, T₂, . . . , Tₘ\} and I is the set of all
items, \( \{i_1, i_2, \ldots, i_n\} \), where each of the transactions \( T_k \) \((1 \leq k \leq m)\) in the database \( T \) represents a set of items \( (T_k \subseteq I) \). An item set is defined as a non-empty subset of \( I \).

Now the association rule defined as: \( X \rightarrow Y(c, s) \), where \( X \subseteq I \), \( Y \subset I \) and \( X \cap Y = \emptyset \).

Where \( s = \text{support and} \ c = \text{confidence} \)

The support calculated as is the percentage of the transactions in which both variable \( X \) and \( Y \) is appeared in the similar transaction while confidence is the proportion of the number of transactions which contains both \( X \) and \( Y \) to the number of transactions that contain only \( X \).

It can be formulated as follows:

\[
\text{Support} (X \rightarrow Y) = P (X \cup Y) \\
\text{Confidence} (X \rightarrow Y) = P (Y|X)
\]

Developing precise and proficient classifiers for large volume of databases is an inherent modus operandi of data mining process. Classification model is build for making future calculation of data objects for which the class is anonymous [29][30].

Association rule mining is to help find the relationship between the itemsets in a large number of databases. By describing the potential rules between the items in databases, dependencies between multiple domains which meet the given support and the confidence threshold are found. The relationship is hidden and unknown in advance, is not got by the logic operation of a database or statistical methods. They are not the inherent properties of the data itself, but on the characteristics of data items simultaneously. The most typical example of association rules cases is: "80 percent of customers buy beers also buy diapers at the same time", its intuitive meaning is, how larger the tendency of customers buy certain products while they will buy other goods. Discovering the rules like this is valuable for setting marketing strategy. Association rules can be applied to analysis of customer shopping, product shelf design, mailing of commercial advertising, catalog design, additional sales, storage planning, network fault analysis, and classification of the users based on buying patterns.

**Fundamental model of Association**

As [10] describe in the article, the task of association rules mining is to find out all the strong association rules in transaction database \( D \) with user-given minimum support and minimum confidence. Corresponding itemsets of strong association rules \( A \rightarrow B \) must be frequent itemsets, and the confidence of association rules \( A \rightarrow B \) derived from frequent itemsets \( A \) \( B \) is calculated by the frequent itemsets \( A \) and \( A \cup B \)'s support. Therefore, the association rules mining can be decomposed into two steps: The first step is to find out all the frequent itemsets in \( D \) quickly and efficiently, which is the central issue of association rules mining.

The second step is to produce a strong frequent itemsets. We use frequent itemsets to generate the required association rules, based on the user-given minimum confidence to select. Finally, strong association rules are generated.

Agrawal et al. has attest the association rules have the following characteristics-

1. The subset of the frequent items is also frequent items.
2. The superset of the non-frequent items is also non-frequent.

**3. METHODS OF ASSOCIATION**

Author of [10], mentioned the techniques most commonly used in association rule mining.

Among the algorithms, Apriori [11] algorithm is the most classical association rule mining algorithms. The algorithm was first proposed by Agrawal et al in 1993.

This algorithm is decomposed into two steps: discover frequent itemsets from candidate frequent itemsets, generate rules from frequent itemsets and pioneering the use of support-based pruning technology, and control the exponential growth of the candidate set.

Apriori algorithm strategy is to separate association rule mining tasks into two steps:

1) The generation of frequent itemsets: by iteration, finding all the itemsets that satisfy minimum support threshold, that is, find all frequent itemsets.

2) Generating association rules: extract high confidence rules from the frequent itemsets found by the former step, which are the strong association rules. The first step for mining frequent itemsets, the algorithm will produce a large number of candidate items, and in order to generate frequent itemsets, scanning databases need to loop through pattern matching to inspect candidates, the cost of computation time of this step will be much...
larger than the second step, and first step is the core of the algorithm.

Here, the basic idea of the generating frequent itemsets is: The algorithm requires multi-step processing of data sets. The first step, statistics the frequency of the set with an element, and identify those itemsets that is not less than the minimum support, that is, the maximum one-dimensional itemsets. Then start the cycle processing from the second step until no more maximum itemsets generated. The cycle is: in the first step k, k-dimensional candidate is generated form (k-1) dimensional maximum itemsets, and then scans the database to get the candidate itemset support, and compare with the minimum support, k-dimensional maximum set is found. Apriori algorithm has two fatal performance bottlenecks: huge numbers of candidates are produced; with multiple scans of transaction database, tedious workload of support counting for candidates will be spent.

3.1 More commonly used Algorithms For Association-

Author of [10] listed various methods of association mining. Some other techniques are as follow:

a. **Data Set Partitioning Algorithm** - Data set partitioning algorithms include Partition algorithm proposed by Savasere [12], DIC algorithm [13] proposed by Brin. Partition algorithm logically divides entire database into a few of separate memory blocks that can be stored in the data processing, saving time of accessing external memory I / O. It is considered separately for each logic block to generate the appropriate frequent sets, and then use "frequent itemset in at least one partition is frequently" the nature to unite the frequent sets in all the logic blocks to generate all possible global candidate sets. Finally, scans the database again to calculate global count of support of sets.

b. **Depth-First Algorithm** - Common depth-first algorithm includes FP-growth algorithm algorithm [14], OIP algorithm [15], TreeProjection algorithm [16]. FP-growth algorithm is one of the latest and most efficient algorithms in depth-first algorithm. The algorithm does not subscribe to generate - and test paradigm of Apriori. Instead, it encodes the data set using a compact data structure called and FP-tree and extracts frequent itemsets directly from this structure.

Compared with the Apriori algorithm, FP-growth algorithm has the following advantages: FP-growth algorithm only scans the database twice, to avoid multiple scans database; do not have a large number of candidates, in the mining process significantly reduces the search space, time and space efficiency are increased dramatically. But its difficulty lies in dealing with large and sparse database, in the mining processing and recursive computations require considerable space.

c. **Breadth-First Algorithm** - Breadth-first algorithm, also known as hierarchical algorithms, including Apriori algorithm proposed by Agrawal and others, AprioriTid algorithm[17] and AprioriHybrid [18] algorithm, DHP algorithm [19] proposed by Park et al and so on. Apriori algorithm is breadth-first algorithm, AprioriTid algorithm evolved based on the Apriori algorithm. This algorithm uses Apriori algorithm when scans the database for the first time, when scans for the second time, it no longer re-scans the entire database, but only scans the last generation of candidate items, and calculates the degree of support of frequent itemsets at the same time ,it reduces the time of scanning database and improves the algorithm efficiency. AprioriHybrid algorithm is proposed based on AprioriTid algorithm and Apriori algorithm.

d. **Incremental Update Algorithm**
e. **Sampling Algorithm**

4. Literature Survey

According to author of [19] [20] association rule mining main goal is to identify the relationship sets which are narrative, functional, interesting inside the item set. The strong point with the association is that ought to hit upon association commencing the latest and hiding item. Conversely, most traditional algorithms [11] [21] only concentrate on data.

Author [19] has surveyed about the validity of association rule mining, Agrawal et al. in 1993[11] initially proposes the idea of association rule to find the new rules in customer transactions items database for better analysis and mining. The idea surfaced on the recursion on the frequent item concept.
Main problem with traditional method is generation of large number of candidate itemsets which require a lot of database scan in the memory.

4.1 Limitation of association

Author of [9] addresses the some limitation of association in the article; author surveyed on the basis of various quantitative elements (metrics), association is divided into two category Boolean and quantitative. Ideally, most of the data is quantitative, so t quantitative association rules mining research is very important. The general method to solve quantitative association rules is that the value of the property is divided into several regions by a certain criteria and then is converted to a sequence-<attribute,interval>. Thus quantitative association rule will be transformed into Boolean association rules. However, there are some problems.

On the one hand, if the interval division is too large, confidence of the rules included in the interval will be very low. So that it will cause a small number of rules, and will be a corresponding reduction in the amount of information. If the interval division is too small, support of the rules included in the interval will be very low. So that it will cause a small number of rules. On other hand, if the domain of property is divided into the non-overlapping interval, the discrete data in the database is mapped to the interval. As potential elements near the interval are excluded by clear division, it will lead to some significant interval is ignored. If the domain of property is divided into overlapping intervals, the elements in the border may be in two intervals at the same time. These elements will contribute to the two intervals, resulting in some intervals are overemphasized.

In order to solve the problem of sharp boundary, fuzzy theory is proposed. The membership function is used to define data set in fuzzy sets of the attribute domain, in order to achieve the purpose of softening the border.

Author of [10], illustrate the role of data mining. Data mining is from a large number of incomplete, noisy, ambiguous and random data, extracting implicit in them, people do not know in advance but is potentially useful information and knowledge. Data mining is used to specify the model type that the data mining task to find. Data mining tasks can generally be divided into two categories: description and prediction. Descriptive mining tasks characterize the general characteristics of data in the database. Predictive data mining tasks predict on the base of the referring of current data.

4.2 Pros and cons of Existing Association with evaluating parameters

Author [22] addresses the pros and cons of association techniques.

The primary function of Association rule mining is to find out interesting correlation or relationships among sets of data items. In Association it checks the condition on attributes that arise repeatedly together in particular dataset. Resulting performance degradation due to huge amount of delivered rules, due to this post processing step is hard in order to select interestingness rules out of large volumes of discovered rules.

For evaluating the Associations rules author [22] proposed some parameters like –

1. Scalability – the system should be scalable as the amount of information scaled.
2. quality of filtered rules and
3. User interesting criteria.

Considering above three metrics author compares many association techniques and concludes the performance. Table shows the authors evaluation of various association rule mining techniques -

Table 1. Comparison of various Association Rule Mining Techniques by [22]

<table>
<thead>
<tr>
<th>Parameters</th>
<th>CLOSET</th>
<th>Ontology-Driven Association Rule Extraction</th>
<th>Selecting the Right Objective Measure for Association Analysis</th>
<th>An Approach to Facilitate the Analysis of the Association Rules</th>
<th>Authors Proposed Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scalability</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>User Interesting criteria</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Quality</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Authors [22] has discusses the problem of choosing interesting (right one) association rules out of large volumes of discovered rules.

Author of [27] deep experiment and proposed a fast association technique, according to author association methods being used to discover associations amongst set of items in knowledge base. The resultant relationships are not the intrinsic properties of the database itself, although it is the co-occurrence of the data items which have derived.

5. FP Tree
To improve the imperfection of Apriori J.Han et al. offers a new technique called FP tree [21] which is not generated any frequent itemsets. FP’s core principal is divide and rule. In this main idea is that positioned itemsets having no candidate sets based on the FP-tree frequent pattern. Then it starts compressing each transaction from database to FP-tree resulting minimum mining time while tree is growing (FP-tree) that’s advances the mining efficiency. It takes more time only if the branch of the tree is large during.

A Frequent pattern tree (FP-tree) is a type of prefix tree [23] that allows the detection of recurrent (frequent) item set exclusive of the candidate item set generation [24]. It is anticipated to recuperate the flaw of existing mining methods. FP –Trees pursue the divide and conquers tactic. The root of the FP-tree is tag as “NULL” value. Childs of the roots are the set of item of data. Conventionally a FP tree contains three fields-item name, node link and count.

To avoid numerous conditional FP-trees during mining of data author of [23] has proposed a new association rule mining technique using improved frequent pattern tree (FP-tree) using table concept conjunction with a mining frequent item set (MFI) method to eliminate the redundant conditional FP tree.

**Definition FP-tree:**
A frequent-pattern tree (or FP-tree) is a tree structure defined below.

1. It consists of one root labeled as “null”, a set of item-prefix subtrees as the children of the root, and a frequent-item-header table.

2. Each node in the item-prefix subtree consists of three fields: item-name, count, and node-link, where item-name registers which item this node represents, count registers the number of transactions represented by the portion of the path reaching this node, and node-link links to the next node in the FP-tree carrying the same item-name, or null if there is none.

3. Each entry in the frequent-item-header table consists of two fields, (1) item-name and (2) head of node-link (a pointer pointing to the first node in the FP-tree carrying the item-name).

5.1 Positive and Negative FP Rule Mining
Author of [28] deep inspect the weakness of association rule mining in the article, according to author [28], mainly positive and negative association rules are identify valuable information enwrap in gigantic data sets, mostly negative rules may returns mutually exclusive relationship between data items. even though a lot of research and experiments on this technique of data mining many challenges still remain in the field positive and negative association rules data mining. Author has proposed a new concept to answered problems like How to determine frequent items? And other one is how to remove ambiguous positive and negative rules?

Author [28] proposed a new method to improve the performance of negative association rule basis on following assumption – “Authors develop a new framework based on new metrics ideology using set of high confidence positive and negative rules to find out recurrent items.”

Author of [25] cleverly explain the concept of positive and negative association rules. According to the [25] two indicators are used to decide the positive and negative of the measure:

1. Firstly find out the correlation according to the value of

   \[ \text{corr}_{P,Q} = s(P \cup Q)/s(P)s(Q) \]

   which is used to delete the contradictory association rules emerged in mining process.

   There are three measurements possible of \( \text{corr}_{P, Q} \) [26]:

   • If \( \text{corr}_{P, Q} > 1 \), Then \( P \) and \( Q \) are related;

   • If \( \text{corr}_{P, Q} = 1 \), Then \( P \) and \( Q \) are independent of each other;
• If corr\(P, Q\) < 1, Then \(P\) and \(Q\) negative correlation;

2. Support and confidence is the positive and negative association rules in two important indicators of the measure.

The support given by the user to meet the minimum support (minsupport) a collection of items called frequent items, association rules mining to find frequent itemsets is concentrating on the needs of the user to set the minimum confidence level (minconf) association rules.

Negative association rules contains itemset does not exist (non-existing-items, for example \(\neg P, \neg Q\)), Direct calculation of their support and confidence level more difficult.

6. Conclusion

Association occupied a significant place in information extraction. The core advantage of this technique is its simplicity due to if-then rule. This article presents a wide survey on association rule mining discussing its pros and cons too. Traditional association rules are not able to produce accurate rule due to the size of database (large) and redundancy in frequent item set produce during mining. In the last section of the article the role of negative and positive is present which guarantee more accurate result then conventional methods.

References:


[17]. Agrawal R, Srikant R. Fast Algorithm for Mining Association Rules in Large Databases[R]. San Jose, IBM Alma den Research Center, 1994


[20]. Jiawei Han, Micheline, Kamber. Data Mining-concepts and techniques, High Education Press. Morgan Kaufman Publishers, 2001.


