Review on Mining Association Rule from Semantic Data

Kalyani A.Kale  
Computer Science and Engg. Department  
Amravati University  
P.R.M.C.E.A.M Badnera,  
Amravati [MH] India

Prof R.P.Sonar  
Computer Science and Engg. Department  
Amravati University  
P.R.M.C.E.A.M Badnera,  
Amravati [MH] India

Abstract: The amount of ontology and semantic annotations for various data of broad applications is constantly growing. This type of complex and heterogeneous semantic data has created new challenges in the area of data mining research. Association Rule Mining is one of the most common data mining techniques which can be defined as extracting the interesting relation among large amount of transactions. Since this technique is more concerned about data representation, it is the most challenging data mining technique to be applied on semantic data. Moreover, the Semantic technologies offer solutions to capture and efficiently use the domain knowledge. This paper presents review on researches done in association rule mining.

Keywords: Data mining, Association rules, semantic data

I. INTRODUCTION:
The main Purpose of data mining is to disclose the hidden information from the database. Due to the growth of data volume in an organization sectors like banking, marketing, telecommunication, manufacturing, and transportation etc., a different technique for deletion of repetitive data and conversion of data to more usable forms has been proposed under data mining. Data mining also known as knowledge discovery is used to discover useful patterns from the database. Many techniques have been developed in data mining amongst which association rule mining is very important. Apriori is one of the best algorithms for the association rule mining. The Apriori algorithm discover the frequent patterns from database whose support and confidence must satisfy the minimum support and confidence.

II. DATA MINING:
Data mining is an inter-disciplinary subfield of computer science, also called as "Knowledge Discovery in Databases" process. It is a process to extracting useful and interesting knowledge from large datasets. Data mining is the computational process involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The knowledge modes data mining discovered have a variety of different types. Including models like association model, classification model, class model, sequence pattern, etc. The overall goal of the data mining process is to extract information from a database and transform it into an understandable structure so that anyone can use it in future. The actual data mining task extract previously unknown interesting patterns such as groups of data records by cluster analysis, unusual records called anomaly detection and dependencies by association rule mining. Mining association rules is one of the most important aspects in data mining.

III. SEMANTIC DATA:
The Semantic data is organized in such a way that it can be interpreted meaningfully without human intervention. Since, 1970 to current semantic data issued in a wide variety of data management systems and applications. Based on relationships between stored symbols and the real world it is a software engineering model. The designed Goals of Semantic Data system is to represent the real world as accurately as possible within some data set. There is linear and hierarchical organization of data to give certain meanings like in below example. Semantic data allow the real world within data sets by representing, machines to interact with worldly information without human interpretation. This semantic data is organized on binary models of objects, mostly in groups of three parts consisting of two objects and their relationship. Consider example, if one wanted to represent a pen is on a letter book, the organization of data might look like: PEN LETTER BOOK. The objects (pen and letter book) are interpreted with regard to their relationship. The data is organized linearly, telling the software that as PEN comes first in the line, it is the object that acts. i.e., the position of the word makes the software to understand that the pen is on the letter book and not that the letter book is sitting on the pen. Databases designed in this concept have greater applicability and are easily integrated into other databases. Since, this semantic data is developing from 1970; its uses are growing on increasing and reach too many important applications. It has very important applications for the enterprise world. Database Management Systems can be integrated with one another and compared. It is helpful model for streamlining the relationship between company and vendors, making database sharing and integration much simpler.

IV. ASSOCIATION RULES:
Association Rule Mining (ARM) is the most important and researched techniques of data mining. ARM was first introduced by Agrawal et al. 1993. It is association tools for
an analyzing customer purchasing habit, such as market-basket analysis. ARM aims to extract interesting frequent patterns, association among set of items or database. Association Rules are if/then statements that help to discover relationships among unrelated data in a data repository. Many algorithms are proposed for finding frequent item sets for large datasets. Association rule uses two criteria support and confidence to identify the relationships and rules are generated by analyzing data for frequent if/then pattern. Association rules are generally needs to satisfy a minimum support and a minimum confidence at the same time.

**V. ASSOCIATION RULE MINING:**

In Data Mining Association rule learning is a method for discovering interesting relations between variables in large database. Association rule discovers interesting association/correlation among a large set of data items. The sales of Super market would indicate that if a customer buys onions and potatoes together, he or she is likely to also buy burger. This information will help business to know the behavior of the customers. Shopping centers also uses the association rule mining to place the items next to each other so that user buy more items. Another application of Association mining is the google auto complete, where we type a word it searches frequently associated words that user type after that particular word.

**VI. CONCEPT OF ASSOCIATION RULE MINING:**

Support(S) of an association rule is defined as the percentage/fraction of records that contain X∪Y to the total number of records in the database. Suppose the support of an item is 0.1%, it means only 0.1 percent of the transaction contain purchasing of this item. (If x and y are two items in database then both comes together).

\[
\text{Support}(X \cup Y) = \frac{\text{Support count of }(X \cup Y)}{\text{Total number of transaction}}
\]

Confidence(C) of an association rule is defined as the percentage/fraction of the number of transactions that contain X∪Y to the total number of records that contain X. Confidence is a measure of strength of the association rules, suppose the confidence of the association rule X⇒Y is 80%, it means that 80% of the transactions that contain X also contain Y together.

\[
\text{Confidence}(X\Rightarrow Y) = \frac{\text{Support}(X \cup Y)}{\text{Support}(X)}
\]

**Item:** It is a field of transactional database.

Consider the following Transactional database Table-I:

<table>
<thead>
<tr>
<th>Transaction ID</th>
<th>Milk</th>
<th>Bread</th>
<th>Butter</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

In Table I, 1 represent the presence of item and 0 represent the absence of items. Now let’s count the support and confidence.

Consider X = milk and Bread, Y = Butter.

Support {milk, Bread} → {Butter} = Support (X⇒Y) = 1/5 = 0.2 (20%)

Confidence {Milk, Bread}⇒{Butter} = Confidence (X⇒Y) = 0.2/0.4 = 0.5 (50%)

Support says that milk butter and bread all purchased together while confidence says that whenever milk and bread purchased there is also possibility of butter.

Association rule mining usually split into two separate steps:

1. Apply minimum support value to find all frequent item sets in a database. This steps required more attention
2. Form rules by using frequent item sets and the minimum confidence value.

**VII. GENERATION MODEL OF ASSOCIATION RULE:**

Association Rules Generation contains many process, they can easily understand by the following related model. The model of Association Rule Generation is in Fig. 2.
Association Rule Generation contain of processes as selection of database from the large repository, then preprocessing on selected data, after that mine the candidate frequent item sets from the preprocessed data, then prune the frequent item sets according to a given threshold. Such a way rules are generated then association rules are mined according to given support and confidence. ARM is mainly used for mining the frequent and infrequent item sets from the large databases. It is based on two principles - support and confidence. Association rules are the if/then sentences to show the relationship among the item sets. One of the most important ARM algorithms is Apriori algorithm.

Apriori Algorithm:
The Apriori algorithm is one of the most popular algorithm in the mining of association rules in a centralized database. Farah Hanna AL-Zawaidah, Yosef Hasan Jbara in [3] proposed the Apriori Algorithm for finding the frequent itemsets. The Apriori Algorithm proposed to finds frequent items in a given data set. The name of Apriori is based on the fact that the algorithm uses a prior knowledge of frequent item set properties. Apriori employs an iterative approach known as a level wise search, where k item sets are used to explore (k+1) itemsets. This algorithm contains a number of passes over the database. During pass k, the algorithm finds the set of frequent item sets Lk of length k that satisfy the minimum support requirement. Apriori is designed to operate on databases containing transactions. The purpose of the Apriori Algorithm is to find associations between different sets of data. It is sometimes referred to as "Market Basket Analysis". Each set of data has a number of items and is called a transaction. The output of Apriori is sets of rules that tell us how often items are contained in sets of data. To illustrate the concepts, there is a small example from the supermarket domain. The set of items is I = {milk, bread, butter, beer} and a small database containing the items. An example association rule for the supermarket could be {milk, bread} => {butter} meaning that if milk and bread is bought, customers also bought butter. Apriori uses bottom up strategy. It is the most famous and classical algorithm for mining frequent patterns. Apriori algorithm works on categorical attributes. Apriori uses breadth first search.
of web usage mining. They stated that as commercial companies as well as academic researchers developed an array of tools that perform several data mining algorithms on log Files coming from web servers in order to identify user behavior on a particular web site. Performing this kind of investigation on the web site can provide information that can be used to better accommodate the user’s needs [6].

- “Ming-Cheng Tseng · Wen-Yang Lin and Rong Jeng” in their paper titled “Updating generalized association rules with evolving taxonomies” stated the problem of updating the discovered generalized association rules under evolving taxonomies. For this purpose they proposed two algorithms Diff ET and Diff ET2 are used for updating generalized frequent item sets. And evaluation showed that both algorithms are effective and have good linear scale-up characteristics [7].

- Zahir Tari and Wensheng Wu in their paper titled “ARM: A HYBRID ASSOCIATION RULE MINING ALGORITHM” stated that Most of the approaches for association rule mining focus on the performance of the discovery of the frequent item sets. They are based on the algorithms that require the transformation of from one representation to another, and therefore excessively use resources and incur heavy CPU overhead. They Proposes a hybrid algorithm that is resource efficient and provides better performance. In addition, they propose a comparison algorithm (CmpApr) that compares candidate item sets with a transaction, a filtering algorithm (FilterApr) that reduces the number of comparison operations required to find frequent item sets. ARM has better performance and scales linearly [8].

IX. CONCLUSION:
Association rule mining is an interesting topic of research in the field of data mining. The paper gives a basic idea about the terms related to association rule mining. Association rules are basic data mining tools for initial data exploration usually applied to large data sets, seeking to identify the most common groups of items occurring together. There are various association rule mining algorithms. This paper studied the most frequently used association rule mining algorithms i.e. Apriori algorithm which is used for discovering all significant association rules between items in a large database of transactions. This paper also studied and represents the reviews on the recent researches done in the field of association rule mining. However, association rule mining is still in a stage of exploration and development.

REFERENCES:
[2] Rakesh Agrawal, Tomasz Imieliński and Arun Swami, “Mining Association Rules between Sets of Items in Large Databases” in IBM Almaden Research Center 650 Harry Road, San Jose, CA 95120 2012