ISSN 2229-5518

Data Mining used in current scenario through Associating Rule Mining

Shivani Yadav, Meer Shizan Ali

Abstract— Association rule mining satisfies the predefined minimum support and confidence from a given database. The problem is usually decomposed into two sub-problems. One is to find the frequent or large itemsets in the database. The second problem is to generate association rules from those large itemsets with minimal confidence. Apriori algorithm is an algorithm proposed to mine the data warehouses to find the associations. More improvements and alternatives have been suggested to overcome the inefficiency of Apriori algorithm. DSIM (Data-Set Intersection Method) algorithm avoids generation of vast volume of candidate itemsets. This process is performed by deleting items in infrequent itemset and merging duplicate transaction repeatedly.

Index Terms— Association rule mining, data mining, maximum frequent itemsets, candidate itemsets, Market Basket Analysis.

1 Introduction

DATA MINING is an emerging concept in databases through Knowledge discovery, prediction, clustering, and classifications. The choice of the algorithm to retrieve the relationship between the variables for a given application is a challenging task, where the accuracy, efficiency, latency, throughput, and security matters as resources are limited. Several algorithms leading to optimal conclusion have been developed and practiced on the datasets to extract patterns. Given the existence of itemsets, association rules make it possible to predict the existence of one or more items based on the knowledge that is gathered by classifying the data warehouses.

Association rule learning is mostly explained by one of its common applications – retrieving the association between the items that customers purchase. Apart from the **Market Basket Analysis**, association rules are also used in mining web usage, intrusion detection, and bioinformatics. Statistical bias caused by suggesting the hypothesis by a narrow sample to match the hypothesis intentionally or unintentionally is defined as data-snooping bias, which leads to a wrong decision in scientific calculations including a highly distributed network. Apriori algorithm is the mostly learned algorithm for association rule mining. It is considered as the fundamental algorithm for data mining for associations, though there exists many variants of algorithm

2 Data Mining Concepts

Data mining is frequently described as "The process of extracting valid, authentic, and actionable information from large databases." Data mining derives patterns and trends collected together and defined as a mining model. Mining model can be applied to specific business scenarios, such as:

- Forecasting sales.
- Targeting mailings toward specific customers.
- Determining which products are likely to be sold together.
- Finding sequences in the order that customers add products to a shopping cart.

This process can be defined by using the following six basic steps:

1. Defining the problem
2. Preparing Data
3. Exploring Data
4. Building Models
5. Exploring and Validating Models
6. Deploying and Updating Models

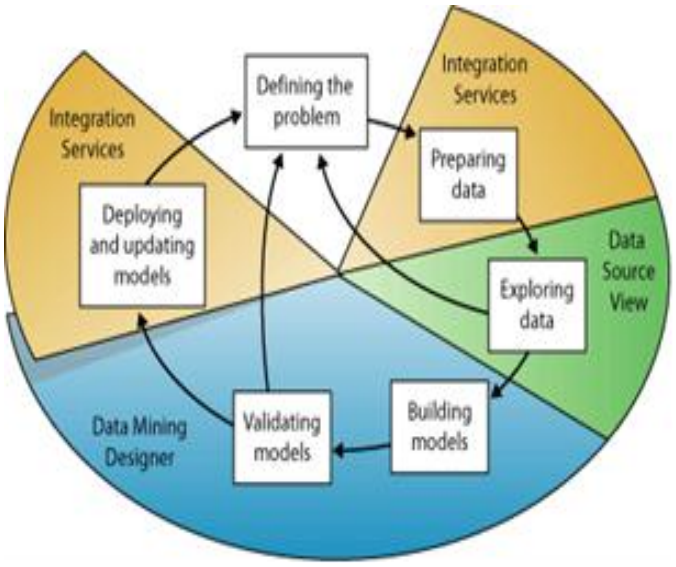

Fig. 2.1 The above diagram describes the relationships between each step in the process

.

- *Shivani Yadav is currently pursuing Masters Degree from Vindhya Institute of Technology and Science, Indore. E-mail: shivani.yadav080@gmail.com*
- *Meer Shizan Ali, Assistant Professor MIT, Indore E-mail:mshizan@gmail.com*

Although the process that is illustrated in the diagram is circular; each step does not necessarily lead directly to the next step. Creating a data mining model is a dynamic and iterative process. User may build several models and realize that they do not answer the problem posed when he/she defined the problem, and therefore look for more data.

Association rules are usually required to satisfy a user-specified minimum support and a user-specified minimum confidence at the same time. Association rule generation is usually split into two separate steps:

1. Minimum support is applied to find all frequent item sets in a database.
2. These frequent itemsets and minimum confidence constraint are used to form rules.

3 Market Basket Analysis

A typical example of frequent item-set mining is market basket analysis. Market Basket - a basket of goods purchased by the buyer within a particular transaction, which is the very characteristic of the operation performed. One of the most common tasks carried out in the analysis of such databases is to find products or sets of items (item-set), which also occur in many transactions. Marketing division can use these templates in order to more accurately place the goods in the shops or restructure pages product catalogues and pages Web.

A set consisting of 'I' goods, is called i-set-element (i-item-set). The percentage of transactions containing a given set, called the support (provision) set. It is believed that its support must be above a user-defined minimum, such sets are called frequent (frequent).

Table 1 describes several transactions (T100, T200, etc), stored in a relational database. And in the corresponding column, it's mentioned the relevant list of item ids for the particular transaction. As an example "T200" contains "I2" and "I4" item ids.

TABLE 1

Transactional data for an *AllElectronics* branch.

TID	List of item_IDs
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

Frequent item-set mining derive rules based on minimum support and the minimum confidence that reflect the usefulness and certainty of the discovered rules. Association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold.

4 Basic Concepts & Association Rule Algorithms

An association rule is an implication in the form of X➔Y, where X, Y **C** I are sets of items called itemsets, and X ∩ Y =∅. X is called antecedent while Y is called consequent, the rule means X implies Y. The basic measures for association rule are support (s) and confidence (c).

Support(s) of an association rule is defined as the percentage/fraction of records that contain X **U** Y to the total number of records in the database.

Confidence of an association rule is defined as the percentage/fraction of the number of transactions that contain X **U** Y to the total number of records that contain X.

Generally, an association rules mining algorithm contains the following steps:

- The set of candidate k-itemsets is generated by 1-extensions of the large (k -1) - itemsets generated in the previous iteration.
- Supports for the candidate k-itemsets are generated by a pass over the database.
- Itemsets that do not have the minimum support are discarded and the remaining itemsets are called large k-itemsets.

This process is repeated until no more large itemsets are found.

Apriori is more efficient during the candidate generation process. Apriori uses pruning techniques to avoid measuring certain itemsets, while guaranteeing completeness.

These are the itemsets that the algorithm can prove will not turn out to be large. However there are two bottlenecks of the Apriori algorithm. One is the complex candidate generation process that uses most of the time, space and memory. Another bottleneck is the multiple scan of the database. Based on Apriori algorithm, many new algorithms were designed with some modifications or improvements.

5 Increasing The Efficiency Of Association Rule Algorithms

The computational cost of association rules mining can be reduced in four ways:

- By reducing the number of passes over the database
- By sampling the database
- By adding extra constraints on the structure of patterns
- Through parallelization.

In recent years much progress has been made in all these directions.

5.1 Reducing The Number Of Passes Over The Database

FP-Tree, frequent pattern mining, in association rule mining breaks the main bottlenecks of the Apriori. The frequent itemsets are generated with only two passes over the database and without any candidate generation process. FP-Tree scales much better than Apriori because as the support threshold goes down, the number as well as the length of frequent itemsets increase dramatically. The frequent patterns generation process includes two sub processes:
a. Constructing the FT-Tree, and
b. Generating frequent patterns from the FP-Tree

The efficiency of FP- Tree algorithm account for three reasons: First the FP-Tree is a compressed representation of the original database because only frequent items are used to construct the tree. Secondly this algorithm only scans the database twice. Thirdly, FP-Tree uses a divide and conquers method that considerably reduced the size of the subsequent conditional FP-Tree.

5.2 Sampling

This approach can be divided into two phases:-
During phase 1 a sample of the database is obtained and all associations in the sample are found. These results are then validated against the entire database. Lowered minimum support on the sample maximizes the effectiveness of the overall approach. Since the approach is probabilistic (i.e. dependent on the sample containing all the relevant associations) not all the rules may be found in this first pass. Those associations that were deemed not frequent in the sample but were actually frequent in the entire dataset are used to construct the complete set of associations in phase 2.

If data comes as a stream flowing at a faster rate than can be processed, sampling seems to be the only choice. How to sample the data and how big the sample size should be for a given error bound and confidence levels are key issues for particular data mining tasks.

5.3 Parallelization

Association rule discovery techniques have gradually been adapted to parallel systems in order to take advantage of the higher speed and greater storage capacity that they offer. The transition to a distributed memory system requires the partitioning of the database among the processors, a procedure that is generally carried out indiscriminately.

FDM algorithm is a parallelization of Apriori. At every level and on each machine, the database scan is performed independently on the local partition. Then a distributed pruning technique is employed.

Another efficient parallel algorithm **FPM (Fast Parallel Mining)** for mining association rules on a shared-nothing parallel system adopts the count distribution approach and has incorporated two powerful candidate pruning techniques, i.e., distributed pruning and global pruning. It has a simple communication scheme which performs only one round of message exchange in each iteration. A new algorithm, **Data Allocation Algorithm (DAA)**, is presented that uses Principal Component Analysis to improve the data distribution prior to FPM.

5.4 Constraints Based Association Rule Mining

The goal of data mining techniques is to discover all the patterns whose frequency in the dataset exceeds a user-specified threshold. Data mining systems should be able to exploit such constraints to speedup the mining process. Techniques applicable to constraint-driven pattern discovery can be classified into the following groups:

- Post-processing (filtering out patterns that do not satisfy user-specified pattern constraints after the actual discovery process);
- Pattern filtering (integration of pattern constraints into the actual mining process in order to generate only patterns satisfying the constraints);
- Dataset filtering (restricting the source dataset to objects that can possibly contain patterns that satisfy pattern constraints).

Rapid Association Rule Mining (RARM) is an association rule mining method that uses the tree structure to represent the original database and avoids candidate generation process. In order to improve the efficiency of existing mining algorithms, constraints were applied during the mining process to generate only those association rules that are interesting to users instead of all the association rules.

6 Categories Of Databases In Which Association Rules Are Applied

Transactional database refers to the collection of transaction records. Data mining on transactional database focuses on the mining of association rules, finding the correlation between items in the transaction records.

Association rule mining with taxonomy is potential to discover more useful knowledge than ordinary flat association rule mining by taking application specific information into account. The more general the items chosen the higher one can expect the support to be.

Spatial databases usually contain not only traditional data but also the location or geographic information about the corresponding data. These describe the relationship between one set of features and another set of features in a spatial database. The form of spatial association rules is also X➔Y, where X, Y are sets of predicates and of which some are spatial predicates, and at least one must be a spatial predicate.

7 Conclusion

Association rule mining has a wide range of applicability such as market basket analysis, medical diagnosis/ research, website navigation analysis, homeland security and so on. The conventional algorithm of association rules discovery proceeds in two steps. All frequent itemsets are found in the first step. The frequent itemset is the itemset that is included in at least *minsup* transactions. The association rules with the confidence at least *minconf* are generated in the second step.

End users of association rule mining tools encounter several well known problems in practice. First, the algorithms do not always return the results in a reasonable time. It is widely recognized that the set of association rules can rapidly grow to be unwieldy, especially as we lower the frequency requirements.

REFERENCES

[1] Data Mining, Concepts and Techniques, 2nd Edition. Jiawei Han and Micheline Kamber.

[2] Database System Concepts, 5th Edition. Abraham Silberschatz, Henry F.Korth, and S.Sudarshan.

[3] Association Rules Mining Algorithm Zhihua Xiao Department of Information System and Computer Science National University of Singapore Lower Kent Ridge Road Singapore.

[4] Association Rule and Quantitative Association Rule Mining among Infrequent ItemsLing Zhou Stephen Yau

[5] An Algorithm for Frequent Pattern Mining Based On Apriori Goswami D.N. et. al. / (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 04, 2010, 942-947

[6] Fast Algorithms for Mining Association Rules Rakesh Agrawal Ramakrishnan Srikant_ IBM Almaden Research Center 650 Harry Road, San Jose, CA 95120