Volume 5, Issue 9, September 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Survey on Association Rule Mining

S. Venkata Krishna Kumar
Associative Professor,
Department of Computer Science,
PSG College of Arts and Science,
Coimbatore, Tamilnadu, India

P. Kiruthika
Researcher Scholar,
Department of Computer Science,
Dr.N.G.P Arts and Science College,
Coimbatore, Tamilnadu, India

Abstract*— Association Rule Mining has been the area of interest for many researchers for a long time and continues to be the same. It is one of the important tasks of Data mining. Association Rule Mining that used to find out correlations, association between a set of transactions is the databases and data warehouse. The knowledge obtained from these database are used for different applications like super market sales prediction, fraud detection etc. This paper presents a review on the basic concepts of ARM technique along with the recent related work that has been done in this field. Therefore this survey guides the researchers to know the progress of pattern mining using association rule for the intended purposes.

Keywords: ARM, Temporal mining, Utility mining, Statistical mining, AIS.

I. INTRODUCTION

Association rule mining is an important technique used in data mining proposed by Agrawal et.al.in 1993. Association rule mining is used for discovering interesting patterns and associations between a set of transactions in the databases. Association rules are basically used in areas like market analysis and inventory control.

II. DATA MINING

Data mining also known as knowledge discovery in databases has established its position as a prominent and important research area. The goal of data mining is to extract higher-level hidden information from an abundance of raw data. Data mining has been used in various data domains. Data mining can be regarded as an algorithmic process that takes data as input and yields patterns, such as classification rules, item sets, association rules, or summaries, as output. Data mining tasks can be classified into two categories, Descriptive Mining and Predictive Mining. The Descriptive Mining techniques such as Clustering, Association Rule Discovery, Sequential Pattern Discovery, is used to find human-interpretable patterns that describe the data. The Predictive Mining techniques like Classification, Regression, Deviation Detection, use some variables to predict unknown or future values of other variables.

III. ASSOCIATION RULE MINING

Mining association rules is one of the research problems in data mining. Given a set of transactions where each transaction is a set of items, n association rule is an expression of the form XY, where X and Y are sets of items, the problem is to generate all association rules that have support and confidence greater than the user-specified minimum support and minimum confidence

$$\begin{tikzcd} & Support = \dfrac{frq(X,Y)}{N} \\ Rule:\ X \Rightarrow Y \arrow[ru] \arrow[rd] & \\ & Confidence = \dfrac{frq(X,Y)}{frq(X)} \end{tikzcd}$$

Supports(S)

Supports(S) of an association rule is defined as the percentage/fraction of records that contain XUY 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.

Support (XY) = Support count of (XY)/ Total number of transaction in D

Confidence(C)

Confidence(C) of an association rule is defined as the percentage/fraction of the number of transactions that contain XUY 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 also contain together.

Confidence (X|Y) = Support (XY)/ Support (X)

Association rule mining is done to find out association rules that satisfy the predefined minimum support and confidence from a given database. The problem of finding association rule is usually decomposed into two sub problems.

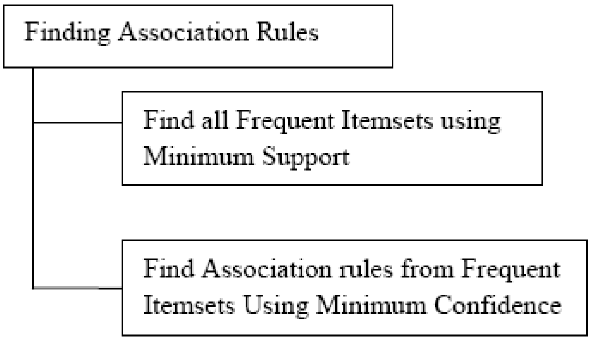
Finding Association Rules
Find all Frequent Itemsets using Minimum Support
Find Association rules from Frequent Itemsets Using Minimum Confidence

Fig 1: Generating Association Rules

In the example database, the item set {bread, egg, butter} has a support of 1/5=0.2 since it occurs in 20% of all transactions (1 out of 5 transactions). The rule{bread, egg}=>{butter}has a confidence of 0.2/0.4=o.5, which means that for 50% of the transactions contain bread and egg (50% of the times a customer buys bread and egg, butter is bought as well).

Table I: Sample database for finding association rule

T	Bread	Egg	Butter	Cheese
T1	1	1	0	0
T2	1	1	1	0
T3	1	0	1	1
T4	0	1	1	0
T5	1	1	0	0

Association rule mining (ARM) is a popular technique for finding co-occurrences, correlations, and frequent patterns, associations among items in a set of transactions or a database. Rules with confidence and support above user-defined thresholds (minconf and minsup) were found. As data continues to grow and its complexity increases, never data structures and algorithms are being developed to match this development. Association Rule mining process can be divided into two steps. The first step involves finding all frequent item sets (or say large item sets) in databases. Once the frequent item sets are found association rules are generated ARM is widely used in market-basket analysis. For example, frequent item set can be found out by analysing market basket data and then association rules can be generate d by predicting the purchase of other items by conditional probability.

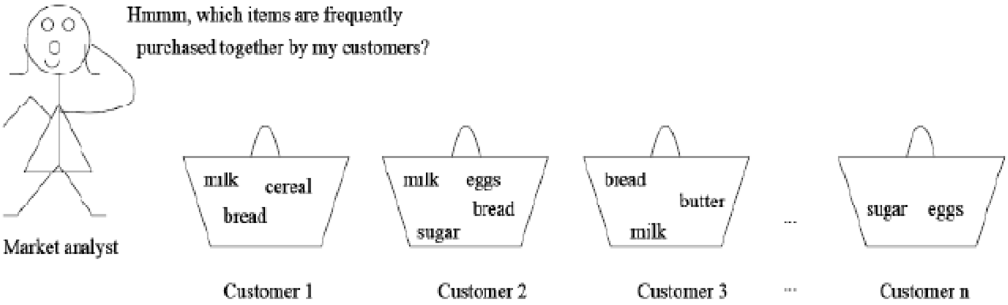

Fig 2: Market Basket Analysis

3.1 Basic Algorithms

Many algorithms for generating association rules have been presented over time. Some of the well known algorithms are Apriori, FP-growth, AIS, Apriori-TID, Apriori Hybrid, Partitioning algorithms, FP-growth Algorithm and many more. The AIS algorithm was the first algorithm to generate all large item sets in a transaction database. The algorithm was the first algorithm to rules. The technique is limited t only item in the consequent. The main problem of the AIS algorithm is that it generates too many candidates that later turn out to be small. Another drawback of this algorithm is that the data structures required for maintaining large candidate item sets are not specified. The Apriori algorithm developed by is the most well known association rule algorithm. Apriori means "from what comes before" and uses breadth first search technique. Its implementation is easier than other algorithms and consumes less memory. However it has certain disadvantages also. It only explains the presence and absence of an item in transactional databases and requires a large number of database scan. Moreover the minimum support threshold used is uniform and the number of candidate item sets produced is large. To overcome some of the bottlenecks of the Apriori algorithm FP-growth algorithm was designed which is based on tree structure.

Table II: The advantages and disadvantages of some of the association rule mining algorithms

Association Rule Mining Algorithm	**Advantages**	**Disadvantages**
AIS	1. An estimation is used in the algorithm to prune those candidate item sets that have no hope to be large.	1. It is limited to only one item in the consequent. 2. Requires Multiple passes over the database.
Apriori	1. This algorithm has least memory consumption. 2. Easy implementation. 3. It uses Apriori property for pruning therefore, item sets left for further support checking remain less.	1. It requires many scans of database. 2. It requires only a single minimum support threshold. 3. It is favourable only for small database. 4. It explains only the presence or absence of an item in the database.
FP-growth	1. It is faster than other association rule mining algorithm. 2. Repeated database scan is eliminated.	1.The memory consumption is more. 2. It cannot be used for interactive mining and incremental mining.

3.2 Temporal Data Mining

Temporal data mining addresses tasks such as segmentation, classification, clustering, forecasting and indexing of time series, event sequences or sections of time series. Temporal data mining is a single step in factor to the process of knowledge discovery in temporal data, and algorithm that enumerates temporal patterns from or fits models to temporal data is a TDM algorithm. Generally adding time factor to the association rules are called as temporal association rule mining. For example 70% of sales of snacks items jump between 5 pm to 7 pm. This sort of inference can be got by adding temporal rule mining.

3.3 Utility Mining

The traditional association rule mining approaches consider the utility of the items by its presence in the transaction set. The frequency of item sets is not sufficient to reflect the actual utility. Utility miner finds all item sets in a transaction database with utility values higher than the minimum utility threshold. To achieve a user’s goal two types of utilities are stated (i) transaction utility and (ii) external utility. Transaction utility of an item is directly obtained from the information stored in the transaction data set. The external utility reflects user preference and can be represented by a utility table. By considering both transaction database and utility table together, data mining can be guided by the utilities of item sets. Hence, the discovered knowledge is useful for maximizing a user’s goal.

3.4 Statistical Association Rule Mining

Normally, the association rule mining algorithms generates more rules. To present relevant and precise interesting rules to the users is of more main concern. Generally interestingness measures play a good role in decision making process. Decision making can be effectively done with less number of rules. Therefore, the concept of post processing or filtering out the less relevant rules can be done using statistical measures. Finding out the statistical significance among the data items can be done using chi-square, correlation coefficient etc.

IV. CONCLUSION

Association rules are widely used in various areas such as telecommunication networks, risk and market management, medical diagnosis, inventory control etc. This paper presents a review on association rule mining. Firstly a brief introduction about association rule mining is given which is the process of finding co-relations, frequent patterns, associations or casual structures among sets of items in the transaction database or other data repositories. The paper surveys the research work done by various authors in this field. Some of the issues related to this field have also been presented which can help upcoming researchers to carry on their work. The advantages and disadvantages of some of the mining algorithms have also been presented in a tabular form.

REFERENCES

[1] R.Agrawal and R.Srikant, “*Fast Algorithms for Mining Association Rules*,” In Proc. of VLDB ’94, pp. 487-499, Santiago, Chile, Sept. 1994.

[2] G. Piatetsky-Shapiro, “Discovery, Analysis, and Presentation of Strong Rules”, In Proceedings of Knowledge Discovery in Databases.ACM,1991,pp.229-248.

[3] R. Agrawal, T. Imielinski, and A. Swami, “*Mining association rules between sets of items in large databases*”. In Proc. of the ACM SIGMOD international Conference on Management of Data - *SIGMOD '93*. p. 207 Washington, D.C., May 1993.

[4] J. Han, M. Kamber, “*Data Mining Concepts and Techniques*”, Morgan Kaufmann Publishers, San Francisco, USA, 2001, ISBN 1558604898.

[5] R. Agrawal, T. Imielinski, and A. Swami, “ Database mining: A performance perspective,” IEEE Transactions on Knowledge and Data Engineering, 5(6):914{925, December 1993. Special Issue on Learning and Discovery in Knowledge Based Databases.

[6] Flach, P. A., &Lachiche, N. (2001), “*Confirmation-Guided Discovery of First-Order Rules with Tertius*,”*Mach. Learn.*, *42*(1-2), 61-95.

[7] J. Han, Y. Cai, and N. Cercone, “ Knowledge discovery in databases: An attribute oriented approach," In Proc. of the VLDB Conference,pages 547{559, Vancouver, British Columbia, Canada, 1992.

[8] Margaret H. Dunham, “Data Mining Introductory and Advanced Topics,” Publisher, Person education india, 2006, Pages 328