A Study and Analysis of Association Rule Mining Algorithms In Data Mining

N. Yuvamathi *1, R. Porkodi 2

*1PG Scholar, Department of Computer Science, Bharathiar University, Coimbatore, Tamilnadu, India

2 Assistant Professor, Department of Computer Science, Bharathiar University, Coimbatore, Tamilnadu, India

ABSTRACT

The data mining is a technology that has been developed rapidly. It is based on complex algorithms that allow for the segmentation of data to identify pattern and trends, detect anomalies, and predict the probability of various situational outcomes. The raw data being mined may come in both analog and digital formats depending on the data sources. There are many trends that are available in data mining some of the new trends are Distributed Data Mining (DDM), Multimedia Data Mining, Spatial and Geographic Data Mining, Time Series and Sequence Data Mining, Time Series and Sequence Data Mining . This paper is based on Association rule mining. In the field of association rule mining, many algorithms exist for exploring the relationships among the items in the database. These algorithms are very much different from one another and take different amount of time to execute on the same sets of data. In this paper, a sample dataset has been taken and various association rule mining algorithms namely Apriori, FP-Growth, Tertius have been compared. The algorithms of association rule mining are discussed and analyzed deeply. The main objective of this paper is to present a review on the basic concepts of ARM technique and its algorithms.

Keywords : Data Mining, Association Rule, Apriori, FP-Growth, Tertius.

I. INTRODUCTION

Data mining concept was raised at the ACM conference in the United States in 1995.It is the computing process of discovering pattern in large data sets involving methods at the intersection of machine learning, statistics and database systems. It is an interdisciplinary subfield of computer science. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Data mining is the analysis step of the “Knowledge Discover of Database” process, or KDD. It is also called as KDD process. Data Mining plays an important role in many areas. Some of the important areas are Future Healthcare, Market Basket Analysis, Education, Manufacturing Engineering, Customer Relationship Management (CRM), Fraud Detection, Intrusion Detection, customer Segmentation, Lie Detection, Financial Banking, Bio informatics, Research Analysis, etc [1].

Recently, there are several data mining techniques that have been developed and used in data mining projects including classification, clustering, association rule, prediction, sequential patterns and decision tree. The Association rule is one of the best known data mining technique. Association rules are if/then statements that help uncover relationships between apparently unrelated data in relational database or other information repository. It has two parts, an antecedent (if) and a consequent (then). An antecedent is an item that found in the data. A consequent is an item that is found in combination with the antecedent [2]. Association rules are created by analyzing data for frequent patterns and using the criteria support and confidence to identify the most important relationships. Association rules use different algorithms to find frequent patterns in the data base.

The section I discuss about the introduction of data mining and the Association rule mining. Section II gives the literature review of the various journals. Section III explains the methods that are used in Association rule mining algorithms. The results and discussions are explained in Section IV and Section V concludes this analysis work.

II. LITERATURE REVIEW

Many journals and articles concerning association rule mining algorithms were studied from year 2000 to 2016. Some compared association rule mining algorithms while some modified the existing algorithms to improve the performance.

XIAO-FENG GU et.al [3]. presented a paper in that the algorithms apriori and FP-Growth are discussed and compared with the experimental result and the new algorithm Growth-FP was proposed and compared with the existing algorithm and the growth fp algorithm is concluded as the best algorithm.

MS.Pooja Jardosh, et.al [4]. presented a paper in that the XML mining and apriori algorithm are discussed briefly. And concluded the paper with the Strategies which are already defined, can be combined together for improvement of apriori algorithm which can overcome most of the problems faced with algorithms.

T.Karthikeyan, et.al [5]. presented a paper that gives an theoretical survey on Association rule mining technique. They concluded with the statement that is in future it is needed to design an efficient algorithm with decreased I/O operation by means of reducing the spells of database scanning.

Subrata Bose, et.al [6]. Adopted a balanced approach for Frequent pattern selection and proposed a method named Weighted support. And the result proved the effectiveness of the proposed Weighted support Algorithm for frequent set generation.

Anubha Sharma, et.al [7]. presented a paper that provides a major advancement in the approaches for association rule mining using genetic algorithm. They proposed to use multi-point crossover operator. The new algorithm significantly reduces the number of rules generated in the data sets.

Bindiya Sagpariya, et.al[8]. presented a paper in that the association rule hiding approaches and related works are discussed briefly. And proposed a technique called hybrid technique that can be found to reduce the side effects and increase the efficiency by reducing the modifications on database, while hiding the association rules.

J.Manimaran, et.al [9]. presented a paper in that Association rule mining and text mining concepts are discussed briefly. Role of association rule mining and apriori algorithm in text application are also discussed. This survey concludes that among the different algorithms used to analyze text data by using ARM technique, Apriori algorithm is suitable and mostly utilized in their chosen domains.

Vidisha H.Zodape, et.al [10]. Presented a paper in that number of papers are taken based on the privacy preserving in mining association rule and all are discussed. This survey shows that the research in this area is increasing.

R.Nedunchezhian, et.al [11]. presented a paper that effectively discuss about Big data, Association rule mining concepts and basic methods of frequent pattern mining and frequent pattern from enormous dataset. And the merits and demerits in all the algorithms are discussed.

Haibat Jadhav, et.al [12]. presented a paper that explored the DES algorithm in client side for generating the secret key to encrypt the items of the support table. This is to achieve more security in the client side. It applies DES algorithm to achieve more security and prevent intruders attack. It also use the Rob frugal method to add the number of fake pattern.

III. METHODOLOGY

The proposed research methodology consists of three phases shown in fig.1. The first phase is preprocessing. The second phase is Association rule mining in that the three algorithms namely Apriori, FP-Growth, Tertius are implemented. And the final phase concludes the performance Evaluation.

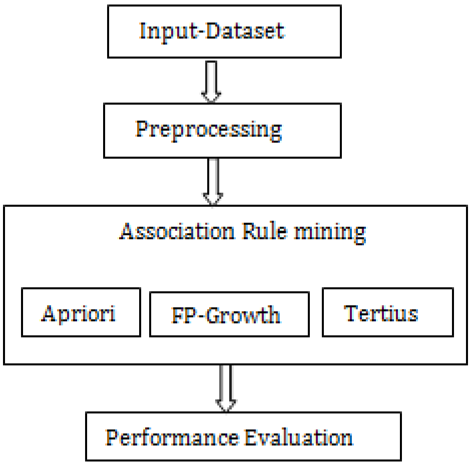

Figure 1. Methodology

1). Association Rule Mining

Association rule mining is a method and it is used to find frequent patterns, correlations, associations, or causal structures from data sets which are found in various kinds of databases like relational databases, transactional databases, and data repositories. Association rules are being used widely in various areas such as telecommunication networks, risk and market management, inventory control, medical diagnosis/drug testing etc.

Let $I = \{i_1, i_2, ...i_n\}$ be the set of items in a dataset and $T=\{t_1, t_2, \ldots t_m\}$ is the set of transactions in the dataset ,it contains m transactions. Association rules are expressed in the form of X=>Y, where X, YUI are item sets, and $X \cap Y = \phi$. Where, X is antecedent and Y is consequent.

Support Count (σ): Frequency of occurrence of item set in transactions σ ({X,Y}).

Support(s): Fraction of transaction that contains an itemset, it also determines how often a rule is applicable to a given data set.

Support, $s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{m}$

Confidence(c): It is a measure that measures how often items in Y appear in transaction that contain X.

Confidence,

$$c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

There are a few commonly used terms that must be defined [3]

Itemset: An itemset is a set of items. A k-itemset is an itemset that contains k number of items.

Frequent itemset: This is an itemset that has minimum support.

Candidate itemset: Set of itemsets that require testing to see if they fit in a certain requirement.

Some other interestingness measures are:

- Expected Predictability: The frequency of occurrence of the item Y is said to be its expected predictability
- Lift: It is the ratio of confidence or predictability to expected confidence or expected predictability i.e. the number of transactions that include the consequent or the right hand side of the rule divided by the total number of transactions [3].

(A). Association Rule Mining Process

Association rule mining is a two step process:

1. Find all frequent item sets: All the item set that occurs at least as frequently as the user specified minimum support count.
2. Generate strong Association rules: These rules must satisfy user defined minimum support and minimum confidence.

(B). Algorithms Used

There are many algorithms for association rule mining. Some of the algorithms are Apriori algorithm, FP-Growth algorithm, Continuous Association Rule Mining Algorithm, Rapid Association Rule Mining, Genetic Algorithm.

In this paper I have implemented and given comparison between three association rule algorithms using the data mining tool Weka.

1. Apriori Algorithm
2. FP-Growth Algorithm
3. Tertius

(C). Apriori Algorithm

Apriori Algorithms is an influential algorithm for mining frequent itemsets for Boolean association rules. It uses 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). Firstly the candidate itemsets are generated. Then the database is scanned for checking the support of these itemsets to generate frequent 1-itemsets. During the first scan, 1-itemsets are generated by rejecting those itemsets whose support is below the threshold. In the consequent passes, the candidate k-itemsets are generated after (k-1)Th pass over the database by joining (k-1) itemsets. The pruning of the non-interesting itemsets is done according to the Apriori property which states that the subset of a frequent itemset must also be frequent [3][13].

Advantages

- Apriori algorithm generates the candidate item sets by using Apriori, which greatly compress the candidate item sets and the size of the frequent item sets, and obtain good performance.
- It is easy to implement.

Disadvantages

- The need for multiple scan database system I/O load is quite large. The time of each scanning will be very long, resulting in a relatively low efficiency of Apriori algorithm.
- It may produce huge candidate item sets. In the worst case, it produces the considerable proved to be non-frequent candidate item sets and the cost of counting is quite high, especially when the candidate set is relatively long, time and space is a challenge.

(D). FP-Growth Algorithm

FP-Growth stands for frequent pattern growth. It is a scalable technique for mining frequent pattern in a database. The algorithm only scans the database for two times. It is a depth first search algorithm combined with direct counting using recursive strategy of pattern growth, it need not generate candidate sets, and instead, the transaction database is compressed into a tree structure that stores only the frequent items [3][13].

The FP-Growth algorithm works in two steps:

1) Construct FP-Tree

- The database is scanned to discover 1-itemsets.
- The items are arranged in an order of decreasing support.
- The database is again scanned to construct FP-Tree.

2) Discovering frequent itemsets using FP-Tree

- Frequent itemsets are found recursively with common suffix, ending with items having lower support first.

Advantages

- The tree - FP algorithm uses a compressed storage tree structure to access transaction records, only for the two scan, the scan time consumption is less than Apriori algorithm.
- FP - tree algorithm does not generate candidate sets completely, and does not count the candidate item sets, so the time performance is better than the Apriori algorithm.

Disadvantages

- FP-tree algorithm in mining frequent patterns inevitably needs to create additional data structures, which will consume a lot of time and space.
- For the FP - tree algorithm, the performance of the algorithm will be affected if the condition tree is very rich (in the worst case).
- The FP - tree algorithm can only be used to excavate the Boolean association rules of a single dimension.

(E). Tertius

This algorithm finds the rule according to the confirmation measures. It uses first order logic representation. It includes various option like class index, classification, confirmation threshold, confirmation values, frequency threshold, missing values, negation, noise Threshold, number literals, report literals, values output etc [13].

Disadvantages

- It is relatively long runtime, which is largely dependent on the number of literals in the rules.
- Running Tertius can take up to several hours for some larger tests.

IV. RESULTS AND DISCUSSION

The experimental dataset supermarket dataset [14] has 217 attributes and 4627 instances. It contains details of products available in the supermarket of different departments.

The preprocessing phase is mainly used in filtering the unique or distinct items from the transaction data base. It was done by using the numeric to nominal conversion in order to convert all the numeric values to nominal. Each item is selected and the results as label, count, missing percent, distinct values and unique percent is used to visualize the preprocessed dataset in bar graph representation.

The Apriori algorithm displays the result such as the schema, relation, instances and the attribute along with the details of generated sets of large itemsets, minimum support 0.1 is taken as a constant value and the confidence values are changed into 0.1, 0.5 and 0.9 and the results are compared. It produces the same count of 10 rules for all confidence values.

The FP-Growth algorithm displays the result such as the schema, relation and instances. The minimum support 0.1 is taken as a constant value and the confidence vales are changed into 0.1, 0.5 and 0.9 and the results are compared. It produces the rules count as 14 for two confidence values such as 0.1, 0.5 and it produce 16 rules for the confidence value 0.9. And the top 10 rules are displayed.

The final result is determined with the details of hypotheses considered and explored. The Threshold values are changed into 0.1, 0.5 and 0.9. For each value it produces the different hypothesis values. The number of hypotheses considered here is 124416 and the number of hypotheses explored is 5276. It generates the set of 10 rules with the total of high values.

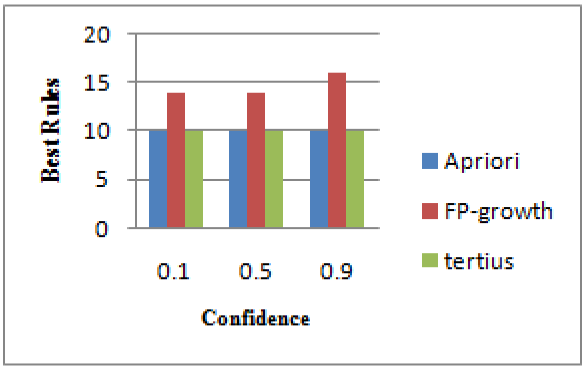

Figure 2. Confidence Vs Rules

If the minimum support degree decreases, the FP-Growth algorithm runs much faster than the Apriori algorithm. If the minimum support degree increases, Apriori algorithm runs faster than the FP-Growth algorithm. Both the algorithms produce the best results based on the minimum support. The FP-Growth Algorithm produces best results than Apriori and Tertius algorithms as shown in fig.2.

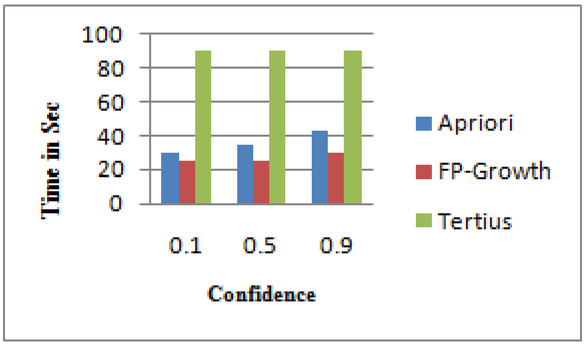

Figure 3. Confidence Vs Time

The time taken for executing these algorithms is depicted in fig.3. The FP-Growth algorithm takes 30 seconds to run the experimental dataset whereas Apriori and Tertius algorithms consume 43 and 90 seconds respectively.

V. CONCLUSION

This paper presents a brief introduction about the association rule mining for finding frequent patterns, co-relation among the items in the database. The paper surveys the research done by the various author in the field. The extensive survey has been conducted in association rule mining area and analyzed various association rule mining algorithms used so far. In this paper I have implemented the three association rule mining algorithms using supermarket dataset and compared the performance of the above algorithms using support and confidence metrics.

It is observed that FP-Growth produces best rules and takes less time to execute the dataset. The second best algorithm is Apriori algorithm it produce best rules but takes more seconds than FP-Growth algorithm. The third best algorithm is Tertius followed by Apriori and it takes too more seconds. Here, the time complexity is an issue to run the algorithms.

VI. REFERENCES

[1]. big-datamadesimple.com

[2]. https://searchbusinessanalytics.techtarget.com/definition/association-rules-in-data-mining

[3]. Comparison and improvement of Association Rule Mining Algorithm, XIAO-FENG GU,XIAO-JUAN HOU,CHEN-XI MA,AO-GUANG WANG, IEEE-2015

[4]. A Comprehensive Survey: Association Rule Mining From XML Ms.Pooja Jardosh1 and Dr.Amit Ganatra1 1Department of Computer Science and Applications, Charotar University of Science and Technology.

[5]. A Survey on Association Rule Mining T. Karthikeyan1 and N. Ravikumar2 Associate Professor, Department of Computer Science, PSG College of Arts and Science, Coimbatore, India1 Research Scholar, Department of Computer Science, Karpagam University, Coimbatore, India 2 .

[6]. Frequent Pattern Generation in Association Rule Mining using Weighted Support Subrata Bose Department of Computer Science & Engineering NITMAS Kolkata, West Bengal,

India subratabose@yahoo.co.in Subrata Datta Department of Information Technology NITMAS Kolkata, West Bengal, India subrataju2008@gmail.com

[7]. A Survey of Association Rule Mining Using Genetic Algorithm .Anubha Sharma Department of CSE Shriram College of engineering & Management, Gwalior (MP), India Nirupma Tivari Assistant Professor, DCSE Shriram College of engineering & Management, Gwalior (MP), India

[8]. A Survey on Association Rule Hiding Approaches. Bindiya Sagpariya1 Kruti Khalpada2 1Computer Engineering, AITS Rajkot, Gujarat India 2 Computer Engineering, AITS Rajkot, Gujarat India Address

[9]. A Survey of Association Rule Mining in Text applications. J.Manimaran1, T. Velmurugan2 1Research Scholar, Research & Development Centre, Bharathiar University, Coimbatore, India 2Associate Professor, Research Dept. of Computer science, D. G. Vaishnav College, Chennai, India thavasimaniraj@gmail.com1, velmurugan_dgvc@yahoo.co.in2

[10]. Literature Survey On Formation Of Association Rule Using Secure Mining .Vidisha H. Zodape, Leena H. Patil

[11]. Association Rule Mining on Big Data - A Survey .Dr. R Nedunchezhian Director of Research KIT - Kalaignarkarunanidhi Institute of Technology Coimbatore K Geethanandhini PG Scholar Department of CSE KIT - Kalaignarkarunanidhi Institute of Technology Coimbatore

[12]. Association Rule Mining Methods for Applying Encryption Techniques in Transaction Dataset. Haibat Jadhav Department of Computer Engineering Flora Institute of Technology, Pune Maharashtra, India haibatj4@gmail.com Prof. Pankaj Chandre Department of Computer Engineering Flora Institute of Technology, Pune Maharashtra, India pankajchandre30@gmail.com

[13]. Comparative Analysis of Association Rule Mining Algorithms Neesha Sharma1 Dr. Chander Kant Verma2 1 M. Tech Student 2Assistant Professor 2 DCSA, Kurukshetra University, Kurukshetra, India

[14]. http://storm.cis.fordham.edu/~gweiss/data-mining/wekw-data/supermarket.arff

[15]. Evaluating the performance of apriori and predictive apriori algorithm to find new association rules based on the statistical measures of datasets. Mukesh Sharma Associate.Professor, Jyoti Choudhary Assistant.Professor, Gunjan Sharma 3Mtech Scholar, Department of Computer Science and Engineering, The Technological Institute of Textile and Science,Bhiwani-127021, Haryana - India.

[16]. R. Brice and W. Alexander, "Finding Interesting Things in Lots of Data." 23rd Hawaii Int. Conf. Syst. Sci., Kona. Hawaii, Jan. 1990.

[17]. G. Piatetsky-Shapiro, "Discovery, Analysis, and Presentation of Strong Rules," in Knowledge Discovery in Databases. Cambridge, MA: AAAI/MIT, 1991, pp. 229-248.

[18]. Gregory Piateski , William Frawley, Knowledge Discovery in Databases, MIT Press, Cambridge, MA, 1991

[19]. https://googleweblight.com

[20]. http://en.m.wikipedia.org