MINING ASSOCIATION RULES FROM LARGE DATA BASES- A REVIEW

R.Priya[1], Ananthi Sheshasaayee[2]
[1]Research Scholar, Asst. Prof, M.C.A. Dept, VELS University, Chennai, India
[2]Associate Professor and Head, Dept of Computer Science,
Quaid-E-Millath Govt. College for women (Autonomous), Chennai, India

ABSTRACT

In today's world, the amount of data transfer has been increasing in a fast pace in all fields due to the rapid advancements in Information technology. It is always intended to obtain meaningful, valuable information that isn't explored earlier from datasets by applying data mining technique .In data mining techniques, association rules are one of the most preferred techniques. Apriori algorithm is the extensively used one for Association mining. Usually the user is interested in finding relationships between certain attributes instead of considering the whole dataset. The preference in selecting the target attribute in the association algorithms aids in generating association rules. The generated rules and the related results will allow us to take decisions for the future. This paper provides preliminaries of basic concepts about Associative Rule Mining and its techniques. Associative Rule Mining is one of the most important data mining techniques designed to group objects together from large databases to find interesting relation or correlations from huge dataset. In this article we provide a brief review and analysis of frequent pattern mining paving new directions on research.

Keywords: *Data Mining, Association rules, Apriori algorithm, Frequent item set, Pattern mining, Data Reduction, Prediction.*

I. INTRODUCTION

`Data Mining: Data mining is the technique of extracting meaningful information from large and mostly unorganized databanks.[1].It is the process of performing automated extraction and generating predictive information from large data banks ,enabling us to understand the current methods and techniques to be applied to large datasets. Data mining involves various steps like cleaning and integrating data from various databases, pre-processing of selected data, transforming data, mining the required knowledge, evaluation and presentation of knowledge. In Data Mining, association rule mining is one of the utmost techniques in identifying patterns between stored data in huge repositories. The mined information is represented as a model of semantic structure, which can be used on new data for prediction or classification. This technique is applied in various fields including marketing, manufacturing, process control, fraud detection and network management.[2].This paper is organized as follows: Chapter II deals with the research methods which elaborates the conceptual methodologies of Association rule mining. Chapter III explains about the various ways of increasing the efficiency of Association rule algorithms. Chapter IV discusses some of the recent advances in Association rule for better solution. Chapter V gives conclusion of Association rule mining through comprehensive study of various algorithms and methodologies for pattern matching.

II. RESEARCH METHODS

2.1Association Rule Mining:

Association rule learning in data mining is a conventional and well researched method for determining interesting relations between attributes in large databases.[3].Association rule mining is mainly intended to recognize strong rules from databases using different measures like support and confidence.

The essentials for performing data mining on any data are discussed below.

Let I={I1,I2,I3….I m } be a set items. Let D, the task relevant data, be a set of database transactions where each transaction T€ I. Each transaction is an association with an identifier, called transaction identification.(TID).Let A be a set of items. A transaction T is said to contain A if and only if A€T. An Association rule is an implication of the form A=> B, where A c I, B c I and AUB=0. [3].Support(S) and Confidence(C) are two measures of rule interestingness.

Support: I= {I1, I2, I3…I m} is a collection of items .T be a collection of transactions associated with the items. T be a collection of transactions associated with the items .Every transaction has an identifier TID [4].

Association rule A=>B is such that, A€I, B€I.A is called as premise and B is called Conclusion.[5].The support, S, is defined as the proportion of transactions in the dataset which contains the item set.

Support(X=>Y) =Support (X U Y) =P (X U Y)

Confidence: The confidence is defined as a conditional probability.

Confidence(X=>Y) =Support (X U Y) / Support(X) = P(Y/X)

Table 1: Transaction Database Table

TID	Item set
1	I1,i2
2	I1,i3,i5
3	I2,i3,im
…	….
M	I1,i3,im

A set of items is referred as an item set. An item set that contains K items is a k-item set. The occurrence frequency of an item set is the number of transactions that contain the item set. If the relative support of an item set I satisfies a prescribed minimum support threshold, then I is a frequent item set.

The association rule mining can be viewed as a two step process:

1)Find all the frequent item sets: Each of these item sets will occur at least as frequently as a predetermined minimum support count.

2)Generate strong association rules from the frequent item sets: The rules must satisfy minimum support and confidence.[3]

The rules which satisfy minimum support and minimum confidence leads to strong rules.

2.2Apriori algorithm:

The Apriori Algorithm[1] is an influential algorithm for mining frequent item sets for Boolean association rules.

Frequent Item sets: The set of item which has minimum support(i th Item set).

Apriori Property: Any subset of frequent item set must be frequent.

Join operation: To find Lk, a set of candidate k- item sets is generated by joining Lk-1 with itself.

The objective is to find the frequent item sets : the sets of items that have minimum support. A subset of a frequent item set must also be a frequent item set and Iteratively find frequent item sets with cardinality from 1 to k(k-item set) and use the frequent item sets to generate association rules.

The following lines state the steps in generating frequent item set in Apriori algorithm[6].

Let Ck be a candidate item set of size k and Lk as a frequent item set of size k. The main steps of iteration are:

1. Find frequent setLk-1.
2. Join step: is generated by joining Lk-1 with itself.(Cartesian product Lk-1 x Lk-1)
3. Prune step (Apriori property):Any(k-1) size item set that is not frequent can't be a subset of a frequent k size item set ,hence should be pruned/removed.

4. Frequent set Lk has been achieved.[3].

2.3Pattern Mining and Data Reduction

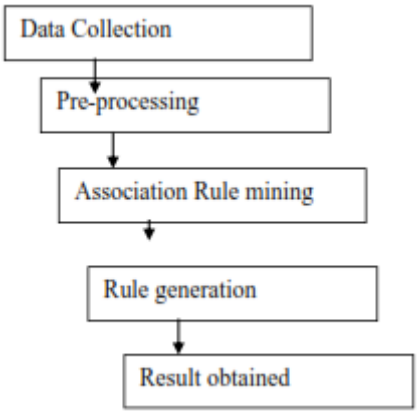

Fig 1. Working Methodology

Usually sample data are collected, pre-processed .Using software tools or with the researcher's mathematical model(algorithm) association rules are generated for different values of confidence and support .On the basis of those rules results are obtained. At the outset the dataset goes through the following phases.

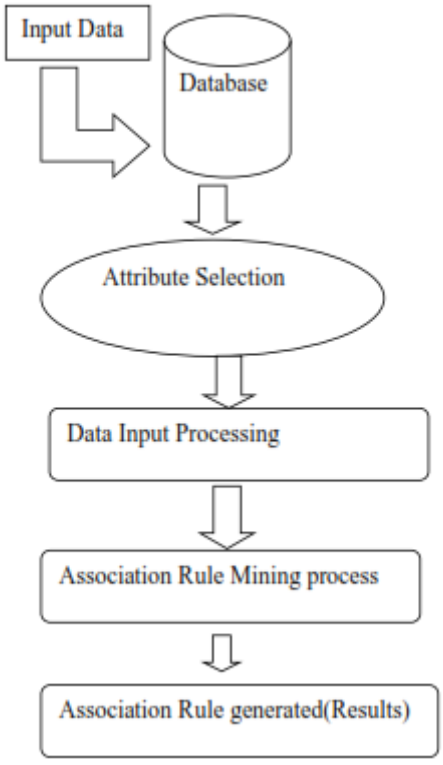

Fig 2.Phases for generating Association rule

III. INCREASING THE EFFICIENCY OF ASSOCIATION RULES ALGORITHMS

The computational cost of association rule mining can be reduced through four ways:[7].

1. By reducing the number of passes over the database.
2. By Sampling the database.
3. By adding extra constraints on the structure of patterns.

4. Through parallelization.

FP-Tree [8],Frequent pattern mining is another milestone in the development of
Association rule mining which breaks bottle necks of Apriori. It doesn't have candidate key generation and frequent item sets are generated in just two passes ,.but can't be used for interactive mining.

Tree Projection is another efficient algorithm recently proposed in [9].This constructs lexicographical tree. Transaction projection limits to small space. The number of nodes in tree is exactly same as that of frequent itemsets.Wang and Tjortis [10] presented PRICES, an algorithm for mining association rules. It is an efficient algorithm which reduces time, scans database only once and uses logical operations. Another algorithm for efficient generating large frequent candidate sets is proposed by[10],called Matrix algorithm. It generates a matrix with 1 or 0 entries. Finally association rules are mined which are effective than apriori. Toivonen[11],presented an association rule mining using sampling which is divided into two phases. Partasarathy[12] presented an efficient method to progressively sample for association rules. Chang et al[13] explored another sampling algorithm called SEE(Sampling Error Estimation) which aims to identify appropriate sample size. Some studies considered the usage of sampling techniques for reducing the processing overhead.[24,25,29].

Constraint driven pattern discovery is classified into three groups:

1. Post processing
2. Pattern filtering
3. Dataset filtering.

Wojciechowski and Zakrrewicz[14] focus on improving the efficiency by using dataset filtering techniques. Tien Dung Do [15] proposed the category based apriori algorithm which reduces the computational complexity. Rapid Association Rule Mining(RARM)[16] is an association rule mining method that uses tree structure for original database and avoids candidate generation process.

Association rule discovery techniques have gradually been adapted to parallel systems in order to take advantage of higher speed and greater storage.[17].Cheung et al[18] presented an algorithm called FDM. Another efficient parallel algorithm FPM(Fast Parallel Mining) has been proposed by[19].,which adopts powerful candidate pruning. Parthasarathy et al[23] have presented an excellent survey on parallel association rule mining with shared memory architecture covering most of the challenges and approaches adopted for parallel data mining. Association rule clustering is useful when the user desires to segment the data .Lent et al[26] proposed a clustering association rule in which they measure the quality of segmentation by ARCS(Association Rule Clustering System).Pi et al[27] proposed a new fuzzy clustering algorithm on Association rules for Knowledge management. Gupta et al[28] recently proposed a cluster based algorithm that uses a novel approach to the insignificant transactions dynamically.

IV. RECENT ADVANCES IN ASSOCIATION RULE DISCOVERY

To address redundant association rule Jaroszewicz and Simovici[20] presented a solution using maximum entropy approach. Techapichchetvanich and Datta[21] presented three step visualization method for mining. Other measures such as all confidence and all –bond other than support can be viewed as a significant measure.[22].Some of the advanced Association Rule Techniques[30] includes the following

1. Generalised Association Rules.
2. Multiple –level Association Rules.
3. Quantitative Association Rule.
4. Using Multiple Minimum Support.
5. Correlation Rules.
6. Temporal Association Rule.

V. CONCLUSION AND FUTURE WORK

In this paper we have seen an overview of Association rule mining and future directions of pattern mining. We have also done a comprehensive study of some algorithms and methodologies for pattern mining .As an enhancement process it is vital to analyze different properties and interestingness

measures for pattern mining algorithms .We conclude that frequent pattern mining has a wide range of applicability, proving its niche in solving a number of problems leading to knowledge discovery. It also paves way for the research scholars to explore new applications for frequent pattern mining.
Association rule mining has a wide range of applicability such as market basket analysis, medical diagnosis/research, website navigation analysis, homeland security, in Educational institutes and so on. To make an improvement, it is vital to analyze different attributes and solutions of the works interested by pattern mining algorithms and design our mathematical model which can outperform the existing mining algorithms through performance .On the outset we require to conduct deep research based on various critical problems so that this domain will have an impact in data mining applications.

REFERENCES

[1]. Han.j. and Kamber.m. ,(2006),"Data mining :Concepts and Techniques",2nd edition. The Morgan Kaufmann series in Data Management Systems" ,Jim Gray, series editor,2006.
[2]. Klaus Julisch, "Data mining for Intrusion Detection-A critical Review" in proc of IBM Research on Application of Data mining in computer security, chapter 1,2002.
[3]. Anwar, M.A., and Nasser Ahmed, "Knowledge Mining in Supervised and Unsupervised Assessment Data of student's performance."2011 2nd International Conference on Networking and Technology IPCSIT vol.17.2011.
[4]. Lei Guoping ,Dai Minlu, Tan Zefu and Wang Yan,"The Research of CMMB wireless Network Analysis Based on Data mining Association Rules" ,IEEE Conference Paper-project supported by the science and technology Research project of Chongqing municipal education commission under contract noKJ101114 and KJ111103,2011.
[5]. S. Suriya, Dr. S.P. Shantharajah, R.Deepalakshmi,"A Complete survey on Association Rule Mining with Relevance to Different Domain, IJASTR, Issue 2,vol1,2012,ISSN :2249-9954.
[6]. Baha sen , Emine Ucar,.Evaluating the achievements of computer engineering department of distance education students with data mining methods. Procedia technology 1,262-267,2012.
[7]. Sotiris Kotsiantis,Dimitris Kanellopoulos,"Association Rules Mining:A Recent Overview" ,GESTS,International Transactions on Computer Science and Engineering, vol 32(1),2006,p.p 71-82.
[8]. Han.J,Pei.J,2000.Mining Frequent patterns by pattern growth :methodology and implications.ACM SIGKDD Explorations,2,2,14-20.
[9]. Agarwal.R ,Aggarwal.c and Prasad.V."A Tree Projection algorithm for generation of frequent item sets In Parallel and Distributed computing",2000.
[10]. Yuan.Y. Huang. T,a matrix algorithm for mining association rules ,volume 3644,sep2005,pp 370-379.
[11]. Toivonen.H.(1996),sampling large databases for association rules, VLDB journal ,pp 134-145.
[12]. Parthasarathy.S., Efficient Progressive Sampling for association rules, ICDM 2002-354-361.
[13]. Chaung.K, Chen. M. Yang.W., 'Progressive sampling for association rules based on sampling error estimation, vol 3518,jun 2005,pages 505-515.
[14]. Wojciechowski and Zakrrewicz, "Dataset filtering techniques in constraint based frequent pattern mining ,Lecture notes in computer science, vol 2447,2002,pp 77-83.
[15]. Tien Dung Do [15],Siu Cheung , Alvis Fong, "Mining Frequent item sets with category based constraints, lecture notes in computer science, vol 2843,2003,pp76-86.
[16]. Das.A,Woon, Y.K,2001, Rapid association mining ,International conference on information and knowledge management, ACM press,474-481.
[17]. Zaki,M.J, Parallel and distributed association mining: a Survey IEEE concurrency ,special issue on parallel mechanisms for data mining,7(4),14-25,December 1999.
[18]. Cheung.D, Han,Fu.A and Fu.Y,1996,A fast distributed algorithm for mining association rules ,in International conference on parallel and distributed information system ,Miami ,florida,pp31-44.
[19]. Cheung.D and Xiao ,Effect of data skewness in parallel mining of association rules, lecture notes in computer science, vol 1394,1998,pp 48-60.
[20]. Jaroszewicz and Simovici, Pruning redundant association rules using mining association rules, lecture notes in computer science , vol 2336,jan 2002,pp 135-142.
[21]. Techapichchetvanich and Data. Visual mining of market basket association rules, vol 3046,jan 2004,pp 479-488.
[22]. Omiecinski,E.(2003),Alternative Interest measures for mining associations in databases, IEEE transactions on knowledge and Data Engineering, vol 15,no 1,pp 57-69.
[23]. Parthasarathy, S.,Zaki, M.J.,Ogihara, M. ,Parallel data mining for association rules on shared memory systems ,Knowledge and information Systems :An International journal,3(1):1-29,Febraury 2001.

[24]. V.Umarani and M.Punithavalli, "Developing a Novel and Effective Approach for Association rule mining using progressive sampling,ICCEE 2009,Vol.1,pp 610-614.
[25]. V.Umarani and M.Punithavalli,"On developing an Effectual Progressive Sampling Based Approach for Association rule discovery", In the proc. of 2nd IEEE Intl conference on Information and data engineering" (2nd IEEE ICIME 2010),Chengdu,China April 2010.
[26]. B.Lent, A.Swami,J. Wisdom, "Clustering association rules ",in the proc of 13th Int'l Conference on Data Engineering,pp.220.
[27]. Pi Dechang and Qin Xiaolin,"A new fuzzy clustering algorithm on Association rules for Knowledge management". Information Technology Journal, pp119-124, 2008. Asian Network for Scientific Information.
[28]. Rajendra K. Gupta and Dev Prakash Agarwal, 'Improving the performance of Association Rule mining Algorithms by filtering In significant Transactions dynamically", Asian Journal of Information management,pp7-17,2009,Academic Journals Inc.
[29]. Basel A.Mahafazh, Amer f.Al-Badarneh and Mohammed Z.Zakaria,"A New sampling technique for association rule mining",in journal of Information science,vol 35,pp 358-376,2009.
[30]. Margaret H.Dunham,"Data mining Introductory and Advanced Topics", Pearson Education 2008.

AUTHORS

R.Priya, is a research scholar, Asst. Professor in the department of MCA, Vels University. She has 14 years of teaching experience in higher education. Her Area of interest is Data Mining and Object Oriented Programming.

Ananthi Sheshasaayee received her Ph.D in Computer Science from Madras University, India. At present she is working as Associate professor and Head, Department of computer science, Quaid-e-Millath Government College for Women, Chennai. She has published 16 National and International journals. Her area of interest involve the fields of Computer Applications and Educational technology