International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622
NATIONAL CONFERENCE on Developments, Advances & Trends in Engineering Sciences
(NCDATES- 09th & 10th January 2015)

RESEARCH ARTICLE OPEN ACCESS

Protected Mining of Association Rules In Parallel Circulated Database

[1]Kodavati Venkateswar rao, [2] YVVS Nagalakshmi
[1] M.Tech Student, (Department of Computer Science, CMREC, Hyderabad)
Email: kvenkat58@gmail.com
[2] Associate professor (Department of Computer Science, CMREC, Hyderabad
Email: lakshmiyadamreddy.cse@gmail.com

Abstract:
In recent years, Association Rule Discovery has become a core topic in Data Mining. It attracts more attention because of its wide applicability. Association rule mining is normally perform generation of frequent item sets and rule generation in which many researchers presented several efficient algorithms. This paper aims at giving a theoretical survey on some of the existing algorithms. The concepts behind association rules are provided at the beginning followed by an overview to some of the previous research works done on this area. The advantages and limitations are discussed and concluded with an inference.
Keywords: Data Mining, Association rule, Frequent item sets.

I. INTRODUCTION

Researchers are drowning in data, but starving for knowledge. Since the dawn of the Interneterain 1994, electronic commerce and data are growing at, such an astonish in grate and the companies around the world race to move their business online inorder to position them in the Internet dominated world wide trading. This technology elevation leads to store tremendous volumes of data in Information repositories like datawarehouses, XML repository, relational database etc. The interesting, useful (potentially useful and previously unknown rules and patterns) information can be extracted from these large information repositories. Experts treat Datamining as the essential process of Knowledge Discovery in Database (KDD) [7]. The KDD process is shown in Fig.1. It is also known as extraction of information, data/pattern analysis, data archaeology, data dredging, information harvesting and business intelligence. Frequent item set mining leads to the discovery of associations and correlations among items in large transactional correlational datasets [3]. The traditional algorithms for mining association rules built on binary attributes databases [10]. An efficient algorithm should reduce the I/O operation of the process of mining by means of decreasing the times of database searching [15].

CATEGORIES OF MINING

The two categories of datamining are Descriptive mining and Prescriptive mining. Summarizing or characterizing the universal properties of data in data repository is known as Descriptive mining. Prescriptive mining is to perform inference on existing data, to make predictions based on the past data [20]. Association rule mining, classification and clustering are some of the data mining techniques.

II. CLASSIFICATION OF DATA

data can be classified into different categories based on the mining techniques that are applied in Datamining. Some of them are (a)Relational data, (b)Transactional data, (c)Spatial data, (d)Temporal and time series data and (e)World Wide Web data.

III. RECENT AREAS OF STUDY

Recently the Chinese government gave great importance to the culture industry for economic growth. To analyse the factors about recognition, satisfaction and participation of resident son cultural activities, Apriori association rule mining algorithm was applied on a survey data. The mining results revealed that income, occupation and educational background as the main factors of culture industry. Based on the results, suggestions were given to make decision support to improve the living standard and education background of residents to improve the participation in cultural activities [30].

In sports management, Association rule mining algorithm was applied for a case study on Indian Cricket team; especially mining relationship on the team's performance data in one day international (ODI) matches [19]. This analysis used in determining the factors associated with the match out

come so as to enable the team to frame match winning strategies. In recent years, Evolutionary Algorithm has been broadly accepted in many systematic areas and it derives mechanisms of biotic progression and applies the mini problem solving [18]. The algorithm also applied in the field of Tax inspection excavation, traffic management and network analysis [25].

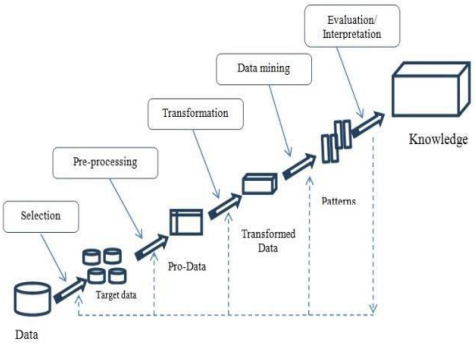

Fig1. KDD Process

IV. ASSOCIATION RULE MINING

Association rule mining discovers the frequent patterns among the item sets. It aims to extract interesting associations, frequent patterns, and correlations among sets of items in the data repositories [9]. For Example, In a Laptop store in India, 80% of the customers who are buying Laptop computers also buy Data card for internet and pendrive for data portability.

The formal statement of Association rule mining problem was initially specified by Agrawal [2]. Let I=$I1,I2,\ldots,Im$ be a set of *m* different attributes, T be the transaction that comprises a set of items such that T$\subseteq$ I, D be a database with different transactions Ts. An association rule is an in sinuation in the form of X$\Rightarrow$ Y, where X, Y$\subset$ I are sets of items termed item sets, and X$\cap$Y=$\varnothing$. X is named antecedent. Y is called consequent. The rule means Ximplies Y.

The two significant basic measures of association rules are *support*(*s*) and *confidence*(*c*). Since the database is enormous in size, users concern about only the frequently bought items. The users can predefine thresholds of support and confidence to drop the rules which are not souseful. The two thresholds are named *minimal support* and *minimal confidence* [20].

Support(*s*) is defined as the proportion of records that contain X $\cup$ Y to the overall records in the database. The amount for each item is augmented by one, whenever the item is crossed over indifferent transaction in database during the course of the scanning.

V. LITERATURE SURVEY

This section presents a survey on Association rule mining algorithms. Agrawal*etal*. [2] introduced the AIS (Agrawal, Imielinski, Swami) algorithm for mining association rules. It focuses on improving the quality of databases along with the required functionality to process queries and consequent association rules are generated.

For example it only generate rules like X$\cap$ Y$\Rightarrow$ Z but not those rules as X$\Rightarrow$ Y$\cap$ Z.

TID	List of items
	I_1,I_2,I_5
T110	I_2,I_4
T120	I_2,I_3
T130	I_1,I_2,I_4
T140	I_1,I_3
T150	I_2,I_3
T160	I_1,I_3
T170	I_1,I_2,I_3,I_5
T180	I_1,I_2,I_3
T190	I_1,I_2,I_5,I_6
*(a)*Actual Database	

Support sum of XY

Support(XY) = _____

Overall records in the database D

Confidence(c) *is* defined as the proportion of the number of transactions that contain X□□ Y to the overall records that contain X, where, if the ratio outperforms the threshold of confidence, an association rule X$\Rightarrow$ Y can be generated.

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

Co association rules, if the confidence of the

Items	Count number
I_1	7
I_2	8
I_3	6
I_4	2
I_5	3
I_6	1
(b)C_1	

Large 1Items
I_1
I
I_3
I_5
(c)L_1

Items	Countnumber
I_1,I_2	5
I_1,I_5	3
I_2,I_5	3
I_2,I_4	2
I_2,I_3	4

D(C2)

association rule X⇒ Y is 80 percent, it infers that 80 percent of the transactions that have X also comprise Y together, like wise to confirm the interestingness of the rules specified minimum confidence is also predefined by users. Association rule mining is to discover association rules that fulfill the predefined minimum support and confidence [1]. The problem is subdivided into two sub problems. The first one is to find the item sets which existences surpass a predefined threshold, usually called frequent item sets. The next one is to generate association rules from large item sets with the limitations of minimal confidence. If one of the large

Large2 Items
I_1,I_2
I_1,I_5
I_2,I_5
I_2,I_3
I_1,I_3
(e)L_2

(b)L1

C1

Items	Count number
I_1,I_2,I_5	3
I_1,I_2,I_4	1
I_1,I_2,I_3	2
I_1,I_2,I_6	1
I_2,I_3,I_5	1
I_1,I_3,I_5	1
…	…
(f)C_3	

Items	Count number
I_1	7
I_2	8
I_3	6
I_4	2
I_5	3
I_6	1

Items	Countnumber
I_1,I_2	5
I_1,I_3	4
I_1,I_5	3
I_2,I_3	4
I_2,I_5	3
I_3,I_5	1
…	…

Table - 1:AISProcess

Large2Items
I_1,I_2
I_1,I_5
I_2,I_5
I_2,I_3
I_1,I_3

In the AIS process, the actual database was scanned many times to get the frequent item sets. Table1 (b) shows the support count of each individual item accumulation during the first pass. Suppose the *minimal support* threshold is 30%, large one item was generated as shown in Table1 (c). Based on that item *I*4 and *I*6 are removed .From frequent 1-items, candidate2-items are generated as mentioned in the Table1 (d). The process iterates until the generating candidate item sets or frequent item sets becomes empty.

Items	Countnumber
I_1,I_2,I_5	3
I_1,I_2,I_3	2

Table - 2: Apriori Process
C2eC3

Agrawaletal., [1] presented an improved algorithm named Apriori for Association rule mining in 1994 and found more efficient. It employs a different candidate generation method and a new pruning technique. In Apriori, there are two processes to find out all the large item sets from the database. The candidate item sets are generated first, then the database is scanned to check the actual support count of the corresponding item sets. In the first scanning,

the support count is calculated and the large1item sets are generated by pruning the item sets falls below the predefined threshold as in Table 2 (a) and (b). The processes are executed iteratively until the candidate/frequent item sets be come empty. Apriori is an influential algorithm for mining frequent item sets for Boolean association rules [16]. An other algorithm Apriori T id[1] is not used the database for counting the support of candidate item sets after the initiation pass. Rather, an encryption of the candidate item sets are used in the previous pass is employed. In later passes, the size of encoding can become much smaller than the database. Hence it is saving much reading effort. Combining the best features of Apriori and Apriori T id, a hybrid algorithm Apriori Hybrid was designed [1]. It uses Apriori in the earlier passes and switches to Apriori T id in the latter passes. Apriori Hybrid performs better than Apriori in almost all cases. Based on the outcome of [1],the Apriori Hybrid has excellent scale-upproperties., opening up the feasibility of mining association rules over very large databases.

Hanetal .[12] [13] worked and designed a tree structure pattern mining algorithm called FP-Tree algorithm (Frequent Pattern Tree). The FP- Tree algorithm generates frequent item sets by scanning the database only twice without any iteration process for candidate generation. The first one is FP- Tree construction process and the next one is generation of frequent patterns from the FP-Tree through a procedure called FP-growth.

Christian Hidber [8] presented Continuous Association Rule Mining Algorithm (CARMA), a novel algorithm to compute large item sets online. The algorithm needs, atmost, two scans of the transaction sequence to produce all large itemsets. During the first scan-Phase-I, the algorithm continuously constructs a lattice of all potentially large item sets. Phase-II initially removes all item sets which are trivially small, i.e. item sets with *max Support* below the last user specified threshold. By rescanning the transaction sequence, Phase-II determines the precise number of occurrences of each remaining item set and continuously removes the item sets, which are found to be small.

A different association rule mining algorithm *Rapid Association Rule Mining* (RARM) [5], uses an efficient tree structure to represent the original database and avoids candidate generation process. Preprocessing is done through *trie Item set* (TrieIT). *RARM* eliminates second time scanning of database and generate 1-item sets and 2-item sets quickly through Support–Oriented Trie Item set (SO Trie IT) structure. A comprehensive theoretical analysis of sampling technique for association rule mining Was presented by Venkatesan *etal* .[27] to assess the quality of the solutions obtained by sampling and showed that the sampling based technique can solve the problems using a sample whose size is in dependent of the number of transactions and the number of items as well.

An extended association rule mining method was proposed by Shuji Morisaki *etal.* [23] that take advantage of interval and ratio scale variables, instead of simply replacing them into nominal or ordinal variables. The rule describes the arithmetic characteristic of quantitative variables in the consequent part composed with related metrics and typical statistics can be revealed as rules. Amit A.Nanavati *etal* .[4] introduced the generalized disjunctive association rules*(d-rules)* which allow the disjunction of conjuncts to capture contextual interrelationships among items. The thrifty-traverse algorithm borrows concepts such as subsumption from propositional logic to mine a sub set of such rules in a computationally feasible way.

It is essential to minimize the harmful impacts as well as maximize possible benefits in the mining process. Negative association rules such as $A \square\square \neg C$ plays an important role in decision making because as $A \Rightarrow \neg C$ can reveal that C (Which may be a harmful factor) rarely occurs when A(which may be a beneficial factor) occurs [29].

Sanat Jain*et.al*.[21] describes a *Genetic Algorithm* (GA) for efficient mining of positive and negative association rules in databases using genetic operators and fitness function assignment. HamidRezaQodmanan *etal.* [11] discussed the application of multi objective genetic algorithm and proposed a method based on genetic algorithm without taking the minimum support and confidence into account. SunitaSarawagi *etal.* [24] explored various architectural alternatives for integrating mining with RDBMS. Jacky *etal.* [14] studied, how X Query can be used to extract association rules from XML data. The intrinsic flexibility structure and semantics of XML databases makes more challenges [22].

Bhatnagar [6] worked to find association rules in distributed databases that can process the databases at their specific sites by swapping required information between them and get similar results that would have been attained if the databases were merged. A protocol was suggested [26] for protected mining of association rules in parallel distributed databases. Olmezogullari *etal.* [17] analysed online association rule mining over big data. It can create more exclusive rules with higher throughput and much lower latency than offline rule mining.

VI. DISCUSSION

In a theoretical study, it is hard to find common factors among the algorithms due to their variable structural aspects. Hence the unique features are

taken for discussion besides the advantages and limitations of some algorithms that are analysed in the review section.

The AIS Algorithm focus on improving quality of Database and process the decision support queries. The databases were scanned many times to get the frequent item sets. Hence this algorithm requires multiple scans on the whole database. It also generates too many candidate item sets and needs more memory.

Apriori algorithm is more efficient than AIS during the candidate generation process. It reduces the computation, I/O cost and memory requirement because of the new pruning technique. By comparing Table1 and Table2, it is apparent that the number of candidate item sets generation reduces intensely.The tree structure design of FP-Tree algorithm breaks the bottle necks of Apriori series algorithms such as complex candidate generation process and multiple scanning. Due to the frequent pattern mining technic, the frequent item sets are generated with only a couple of scans and eliminate the candidate generation procedure. Hence it is faster than Apriori algorithm. But on the other side it is difficult to use in an inter active mining system and not suitable for incremental mining. RARM method uses the tree structure to represent the original database and avoids candidate generation process. It is much faster than FP-Tree algorithm since it generates large1-itemsetsand2-item sets quickly without scanning the database for the second time. But it requires more memory.

VII. CONCLUSION

The algorithmic aspects of association rule mining are reviewed in this paper and observed that a lot of attention was focused on the performance and scalability of the algorithms, but not adequate attention was given to the quality (interestingness) of the rule generated. The above discussed algorithms may be enhanced to reduce the execution time, complexity and improve the accuracy. It is concluded that, in the association rule mining process, further attentiveness is needed in designing an efficient algorithm with decreased I/O operation by means of reducing the spells of database scanning. This kind of approach may be lead to various architectural alternatives in future and these methods are very useful in Data Mining to minimize the harmful impacts and maximizing the possible benefits.

REFERENCES

[1] Agrawal, R., et al "Mining association rules between sets of items in large database". In: Proc. of ACM SIGMOD'93, D.C, ACM Press, Washington, pp.207-216, 1993.

[2]. Agarwal, R., Imielinski, T., Swamy, A. "Mining Association Rules between Sets of Items in Large Databases", Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data,pp. 207-210, 1993.

[3]. Srikant, R., Agrawal, R "Mining generalized association rules", In: VLDB'95, pp.479-488, 1994.

[4] Agrawal, R., Srikant, R, "Privacy-Preserving Data Mining", In: proceedings of the 2000 ACM SIGMOD on management of data, pp. 439-450, 2000.

[5] Lindell, Y., Pinkas, B, "Privacy preserving Data Mining", In: Proceedings of 20th Annual International Cryptology Conference (CRYPTO), 2000.

[6] Kantarcioglu, M., Clifton, C, "Privacy-Preserving distributed mining of association rules on horizontally partitioned data", In IEEE Transactions on Knowledge and Data Engineering Journal, IEEE Press, Vol 16(9), pp.1026-1037, 2004.

[7] Han, J. Kamber, M, "Data Mining Concepts and Techniques". Morgan Kaufmann, San Francisco, 2006.

[8] Sheikh, R., Kumar, B., Mishra, D, K, "A Distributed k- Secure Sum Protocol for Secure Multi-Site Computations". Journal of Computing, Vol 2, pp.239-243, 2010.

[9] Sugumar, Jayakumar, R., Rengarajan, C "Design a Secure Multi Site Computation System for Privacy Preserving Data Mining". International Journal of Computer Science and Telecommunications, Vol 3, pp.101-105. 2012.

[10] N V Muthu Lakshmi, Dr. K Sandhya Rani ,"Privacy Preserving Association Rule Mining without Trusted Site for Horizontal Partitioned database", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, pp.17-29, 2012.

[11] N V Muthu lakshmi, Dr. K Sandhya Rani, "Privacy Preserving Association Rule Mining in Horizontally Partitioned Databases Using Cryptography Techniques", International Journal of Computer Science and Information Technologies(IJCSIT), Vol. 3 (1) , PP. 3176 – 3182, 2012.

[12] Goldreich, O., Micali, S. & Wigerson, A. ,"How to play any mental game", In: Proceedings of the 19th Annual ACM Symposium on Theory of Computing, pp.218-229, 1987.

[13] Franklin, M., Galil, Z. & Yung, M.,"An overview of Secured Distributed Computing". Technical Report CUCS- 00892, Department of Computer Science, Columbia University.