ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 1, January 2014

A Survey on Association Rule Mining

T. Karthikeyan[1] and N. Ravikumar[2]

Associate Professor, Department of Computer Science, PSG College of Arts and Science, Coimbatore, India[1]

Research Scholar, Department of Computer Science, Karpagam University, Coimbatore, India[2]

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 performed in generation of frequent itemsets 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 itemsets

I. INTRODUCTION

Researchers are drowning in data, but starving for knowledge. Since the dawn of the Internet era in 1994, electronic commerce and e-data are growing at, such an astonishing rate and the companies around the world race to move their business online in order to position them in the Internet dominated worldwide trading [4]. This technology elevation leads to store tremendous volumes of data in Information repositories like data warehouses, 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 Data mining 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 or relational 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].

1.1.1 Categories of Mining

The two categories of data mining 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.

1.1.2 Classification of Data

The data can be classified into different categories based on the mining techniques that are applied in Data mining. 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.

1.1.3 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 residents on 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 outcome 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 them in problem solving [18]. The algorithm also applied in the field of Tax inspection excavation, traffic management and network analysis [25].

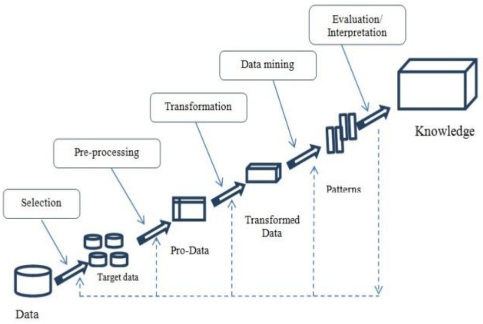

Fig. 1. KDD process

II. ASSOCIATION RULE MINING

Association rule mining discovers the frequent patterns among the itemsets. 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 pen drive for data portability.

The formal statement of Association rule mining problem was initially specified by Agrawal [2]. Let I = $I_1, I_2, \dots, I_m$ be a set of *m* different attributes, T be the transaction that comprises a set of items such that T ⊆ I, D be a database with different transactions Ts. An association rule is an insinuation in the form of X ⇒ Y, where X, Y ⊂ I are sets of items termed itemsets, and X ∩ Y = ∅. X is named antecedent. Y is called consequent. The rule means X implies 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 pre-define thresholds of support and confidence to drop the rules which are not so useful. The two thresholds are named *minimal support* and *minimal confidence* [20].

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

$$\text{Support (XY)} = \frac{\text{Support sum of XY}}{\text{Overall records in the database D}}$$

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 ⇒ Y can be generated.

$$Confidence(X/Y) = \frac{Support\ (XY)}{Support\ (X)}$$

Confidence is a degree of strength of the association rules, if the confidence of the association rule X ⇒ Y is 80 per cent, it infers that 80 per cent of the transactions that have X also comprise Y together, likewise to confirm the interestingness of the rules specified minimum confidence is also pre-defined by users. Association rule mining is to discover association rules that fulfil the pre-defined minimum support and confidence [1]. The problem is subdivided into two sub problems. The first one is to find the itemsets which existences surpass a predefined threshold, usually called frequent itemsets. The next one is to generate association rules from large itemsets with the limitations of minimal confidence. If one of the large itemsets is L_k, $L_k = \{I_1, I_2, \dots, I_{k-1}, I_k\}$, then association rules are generated with those itemsets. Checking the confidence with the rule $\{I_1, I_2, \dots, I_{k-1}\} \Rightarrow \{I_k\}$, it can be decided for interestingness. By deleting the last items, the other rules are created in the antecedent and placing it to the consequent, then the confidences of the new rules are checked to decide the interestingness. The processes iterated till the antecedent becomes empty. The main sub problem can be two folded into *candidate large itemsets* generation process and *frequent itemsets* generation process. Those itemsets whose support exceeds the support threshold called as *large* or *frequent itemsets*, those itemsets that are expected to be large or frequent are known *candidate itemsets*. An efficient model has classification rules with high confidence and large support [28].

III. LITERATURE SURVEY

This section presents a survey on Association rule mining algorithms. Agrawal *et al.*[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 ∩ Y ⇒ Z but not those rules as X ⇒ Y ∩ 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

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 1 Items
I_1
I_2
I_3
I_5

(c) L_1

Items	Count number
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
I_1, I_4	1
…	…

(d) C_2

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 1, January 2014

Large 2 Items
I_1, I_2
I_1, I_5
I_2, I_5
I_2, I_3
I_1, I_3

(e) L_2

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

Table - 1 : AIS Process

In the AIS process, the actual database was scanned many times to get the frequent itemsets. Table 1(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 Table 1(c). Based on that item I_4 and I_6 are removed. From frequent 1-items, candidate 2-items are generated as mentioned in the Table 1(d). The process iterates until the generating candidate itemsets or frequent itemsets becomes empty.

Agrawal *et al.*, [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 itemsets from the database. The candidate itemsets are generated first, then the database is scanned to check the actual support count of the corresponding itemsets. In the first scanning, the support count is calculated and the large 1-itemsets are generated by pruning the itemsets falls below the predefined threshold as in Table 2(a) and (b). The processes are executed iteratively until the candidate / frequent itemsets become empty. Apriori is an influential algorithm for mining frequent itemsets for Boolean association rules [16]. Another algorithm AprioriTid [1] is not used the database for counting the support of candidate itemsets after the initiation pass. Rather, an encryption of the candidate itemsets 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 AprioriTid, a hybrid algorithm AprioriHybrid was designed [1]. It uses Apriori in the earlier passes and switches to AprioriTid in the latter passes. AprioriHybrid performs better than Apriori in almost all cases. Based on the outcome of [1], the AprioriHybrid has excellent scale-up properties., opening up the feasibility of mining association rules over very large databases.

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

(a) C_1

Large 1 Items
I_1
I_2
I_3
I_5

(b) L_1

Items	Count number
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
…	…

(c) C_2

Large 2 Items
I_1, I_2
I_1, I_5
I_2, I_5
I_2, I_3
I_1, I_3

(d) L_2

Items	Count number
I_1, I_2, I_5	3
I_1, I_2, I_3	2

(e) C_3

Table - 2 : Apriori Process

Han *et al.* [12][13] worked and designed a tree structure pattern mining algorithm called *FP-Tree algorithm* (Frequent Pattern Tree). The FP-Tree algorithm generates frequent itemsets 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 itemsets online. The algorithm needs, at most, 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 itemsets. Phase-II initially removes all itemsets which are trivially small, i.e. itemsets with *maxSupport* below the last user specified threshold. By rescanning the transaction sequence, Phase-II determines the precise number of occurrences of each remaining itemset and continuously removes the itemsets, 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. Pre-processing is done through *trie Itemset* (TrieIT). *RARM* eliminates second time scanning of database and generate 1-itemsets and 2-itemsets quickly through Support–Oriented Trie Itemset (SOTrieIT) structure. A comprehensive theoretical analysis of sampling technique for association rule mining was presented by Venkatesan *et al.* [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 independent of the number of transactions and the number of items as well.

An extended association rule mining method was proposed by Shuji Morisaki *et al.* [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 *et al.* [4] introduced the generalized disjunctive association rules *(d-rules)* which allow the disjunction of conjuncts to capture contextual inter-relationships among items. The thrifty-traverse algorithm borrows concepts such as subsumption from propositional logic to mine a subset 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 $A \Rightarrow \neg C$ 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. Hamid Reza Qodmanan *et al.* [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. Sunita Sarawagi *et al.* [24] explored various architectural alternatives for integrating mining with RDBMS. Jacky *et al.* [14] studied, how XQuery can be used to extract association rules from XML data. The intrinsic flexibility structure and semantics of XML data bases makes more challenges [22].

Bhatnagar [6] worked to find association rules in distributed data bases that can process the data bases 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 *et al.* [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.

IV. 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 itemsets. Hence this algorithm requires multiple scans on the whole database. It also generates too many candidate itemsets 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 Table 1 and Table 2, it is apparent that the number of candidate itemsets generation reduces intensely.

The tree structure design of FP-Tree algorithm breaks the bottlenecks of Apriori series algorithms such as complex candidate generation process and multiple scanning. Due to the frequent pattern mining technic, the frequent itemsets 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 interactive 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 large 1-itemsets and 2-itemsets quickly without scanning the database for the second time. But it requires more memory.

V. 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. and Srikant, R. Fast Algorithms for Mining Association rules. Proc. 20th VLDB conference, Santiago, Chile, 1994.

[2] Agrawal, R., Imielinski, T. and Swami, A. Mining association rules between sets of items in large databases. In Proceedings of the ACM SIGMOD International Conference on Management of Data, 207-216, 1993.

[3] Al-Maqaleh, B.M. and Shaab, S.K. An Efficient Algorithm for Mining Association Rules Using Confident Frequent Itemsets. 3rd

International Conference on Advanced Computing and Communication Technologies, 90-94, Rohtak, 6-7 April 2013.

[4] Amit A. Nanavati, Krishna, I. and Chitrapura Sachindra Joshi Raghu Krishnapuram. Mining Generalised Disjunctive Association rules. ACM 1-581 13-436-3/01/0011, 2001.

[5] Amitabha Das, Wee Keong Ng and Yew Kwong Woon. Rapid Association Rule Mining. ACM, 1-58113-436/01/0011, 2001.

[6] Bhatnagar, S. Algorithm for finding association rules in distributed databases. 2nd IEEE International Conference on Parallel Distributed and Grid Computing, 915-920, Solan, 6-8 Dec, 2012.

[7] Chen, M.S., Han, J. and Yu, P. S. Data mining: an overview from a database perspective. IEEE Trans. On Knowledge and Data Engineering, pp 866 – 883, 1996.

[8] Christian Hidber. Online Association rule mining. SIGMOD '99 Philadelphia PA. ACM 1-58113-084-8/99/05, 1999.

[9] Dhanabhakyam, M and Punithavalli, M. A Survey on Data Mining Algorithm for Market Basket Analysis. Global Journal of Computer Science and Technology, Vol. 11 issue 11, version 1.0, 2011.

[10] Gosain, A. and Bhugra, M. A comprehensive survey of association rules on quantitative data in data mining. IEEE Conference on Information & Communication Technologies, 1003-1008, JeJu Island, 11-12 Apr 2013.

[11] Hamid Reza Qodmanan, Mahdi Nasiri and Behrouz Minaei-Bidgoli. Multi objective association rule mining with genetic algorithm without specifying minimum support and minimum confidence. ELSEVIER – Expert Systems with Applications, Vol. 38, Issue 1, pp. 288-298, January 2011.

[12] Han, J. and Kamber, M. Data Mining: Concepts and Techniques. Morgan Kanufmann, San Francisco, CA USA, 2001.

[13] Han, J., Pei, J. and Yin, Y. Mining frequent patterns without candidate generation. ACM SIGMOD Intl. Conference on Management of Data, ACM Press, 1-12, 2000.

[14] Jacky W.W. Wan and Gillian Dobbie. Mining Association rules from XML data using XQuery. Australian Computer Society, 2004.

[15] Li Xiaohui. Improvement of Apriori algorithm for association rules. World Automation Congress, 1-4, Mexico, 24-28 Jun 2012.

[16] Luo Fang. The Study on the Application of Data Mining based on Association Rules. International Conference on Communication Systems and Network Technologies, 477-480, Rajkot, 11-13 May 2012.

[17] Olmezogullari, Erdi, Ari and Ismail. Online Association Rule Mining over Fast Data. IEEE International Congress on Big Data (BigData Congress), 110-117, Santa Clara, CA, USA, 27th Jun – 2nd Jul 2013.

[18] Peddi Kishor and Sammulal Porika. Literature Survey on Association Rule Discovery in Data Mining. International Journal of Computer Science and Management Research, Vol. 2, Issue 1, Jan 2013.

[19] Raj, K.A.A.D. and Padma, P. Application of Association Rule Mining: A case study on team India. International Conference on Computer Communication and Informatics, 1-6, Coimbatore, 4-6 Jan 2013.

[20] Qiankun Zhao and Sourav S. Bhowmick. Association rule mining : A Survey. Technical Report, CAIS, Nanyang Technological University, Singapore, No. 2003116, 2003.

[21] Sanat Jain and Swati Kabra. Mining and Optimization of Association Rules using effective algorithm. International Journal of Emerging Technology and Advanced Engineering, Vol. 2, Issue 4, April 2012.

[22] Sasikala, D. and Premalatha, K. Mining association rules from XML document using modified index table. International Conference on Computer Communication and Informatics, Coimbatore, 4-6 Jan 2013.

[23] Shuji Morisaki, Akito Monden, Tomoko Matsumura, Haruaki Tamada and Ken-ichi Matsumoto. Defect Data Analysis Based on Extended Association rule mining. Fourth international workshop on Mining Software Repositories. IEEE Computer Society, 2007.

[24] Sunita Sarawagi, Shiby Thomas and Rakesh Agrawal. Integrating Association rule mining with Relational Database Systems: Alternatives and Implications. ACM 089791-995-5/98/006, 1998.

[25] Suriya, S., Shantharajah, S.P. and Deepalakshmi, R. A Complete survey on association rule mining with relevance to different domain. International Journal of Advanced Scientific and Technical Research, Issue 2, Vol. 1, Feb 2012.

[26] Tassa, T. Secure Mining of Association Rules in Horizontally Distributed Databases. IEEE Transactions on Knowledge and Data Engineering. Issue 99, 1, 07 Mar 2013.

[27] Venkatesan T Chakaravarthy, Vinayaka Pandit and Yogish Sabharwal. Analysis of Sampling Techniques for Association rule mining. ACM 978-1-60558-423-2/09/0003, 2009.

[28] Viet Phan Luong and Lif. Mining normal and abonormal class-association rules. IEEE 27th International Conference on Advanced Information Networking and Applications. 968-975, Barcelona, 25-28 Mar 2013.

[29] Xindong Wu, Chengqi Zhang and Shichao Zhang. Efficient Mining of both Positive and Negative Association rules. ACM 1046-8188/04/0700-0381, 2004.

[30] Zhengui Li and Renshou Zhang. The association rule mining on a survey data for culture industry. International Conference on Systems and Informatics, Yantaj, 19-20 May 2012.

BIOGRAPHIES

Prof. Thirunavukarasu Karthikeyan received his graduate degree in Mathematics from Madras University in 1982. Post graduate degree in Applied Mathematics from Bharathidasan University in 1984. Received Ph.D. in Computer Science from Bharathiyar University in 2009. Presently he is working as an Associate Professor in Computer Science Department of P.S.G. College of Arts and Science, Coimbatore. His research interests are Image Coding, Medical Image Processing, Data Mining and Software Engineering. He has published many papers in national and international conferences and journals. He has completed many funded projects with excellent comments. He has contributed as a program committee member for a number of international conferences. He is the review board member of various reputed journals. He is board of studies member for various autonomous institutions and universities.

N. Ravikumar is a Research Scholar of Karpagam University. He received his Master degree in Computer Applications from Bharathiar University in 2001 and received his M.Phil Degree in Computer Science from Bharathiar University in 2005. He has participated in Some National Conferences and workshops and presented papers in Data Mining