Volume 2, Issue 9, September 2012 **ISSN: 2277 128X**

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A study on Various Data Mining Approaches of Association Rules

Rachna Somkunwar*

Computer Department, Nagpur University

India

***Abstract*— Data mining is considered to deal with huge amounts of data which are kept in the database, to locate required information and facts. Innovation of association rules among the huge number of item sets is observed as a significant feature of data mining. The always growing demand of finding pattern from huge data improves the association rule mining. The main purpose of data mining provides superior result for using knowledge base system. Researchers presented a lot of approaches and algorithms for determining association rules. This paper discusses few approaches for mining association rules. Association rule mining approach is the most efficient data mining method to find out hidden or required pattern among the large volume of data. It is responsible to find correlation relationships among various data attributes in a huge set of items in a database. Studying Apriori algorithm, it is a illustration of an enhanced association rule mining algorithm, which supports to avoid the replication of same items. This paper discusses an enhanced version of Apriori algorithm that is concentrated on four characteristics namely, First data preparation and chooses the desired data, second produce itemsets that decides the rule constraints for knowledge, third mine k-frequent itemsets using the new database and fourth produce the association rule that sets up the knowledge base and offer better results. Another approach discussed in this paper are the HASH MAPPING TABLE and HASH_TREE tactics used to optimize space complexity and time complexity.**

Keywords—Data Mining, Association rules, Apriori algorithm, HASH_TREE, HMT

I. INTRODUCTION

Data mining techniques can be categorized according to the objectives they follow and the results they offer, which obtains computer as a tool and makes use of the skill and knowledge significance to comprehend and explain the problem. Various data mining techniques such as, decision trees, association rules, and neural networks are already presented and become the point of attention for several years. Association rule mining technique is the most efficient data mining technique to search hidden or desired pattern among the huge amount of data. It is responsible to get correlation relationships among various data attributes in a large set of items in a database. A huge quantity of interesting relevance or related association across the itemsets has been determined by association rules mining. A typical example of the association rules mining is the market basket analysis. Association rules research assists to find the relationship among different products (items) in transaction databases and to find out the customer buyer behaviors, such as the purchase of a commodity impact on other goods. The results can be applied to goods shelf layout, storage arrangements, and classification of user according to buying patterns. Association Analysis is the detection of hidden pattern or condition that occurs frequently together in a given data. Association Rule mining techniques finds interesting associations and correlations among data set. An association rule is a rule, which entails certain association relationships with objects or items, for example the interrelationship of the data item as whether they occur simultaneously with other data item and how often. These rules are computed from the data and, association rules are calculated with help of probability. It has a mentionable amount of practical applications, including classification, XML mining, spatial data analysis, and share market and recommendation systems. This rule measure with support to ensure every dataset treated equally in classical model. The perception of association rule mining suggests the support confidence level outline and condensed association rule mining to the discovery of frequent item sets. Rule support and confidence are two measures of interestingness. Association rules are regarded as appealing if a minimum support and a minimum confidence threshold is satisfied. Boolean association rule mining is more extensively used than other kinds of association rule mining. Association rule mining procedure can be finished in four steps. First data preparation and pick the required data, second produce itemsets that determines the rule constraints for knowledge, third mine k frequent itemsets using the new database and fourth produce the association rule that sets up the knowledge base. The paper discusses an algorithm to mine association rules and the support and confidence are studied.

For Apriori algorithm, there are two disadvantages. First, it must scan data sets repeatedly, which may direct to generate a large number of candidate itemsets. Second, the rare information is difficult to dig because the limitation of

algorithm. The competence of Apriori algorithm has a greatly effect on performance and practicality of related data mining system. When there are several elements and minimum support threshold is low, the competence of Apriori algorithm is easy to become a bottleneck of performance. Subsequently, many researchers proposed a lot of research to the Apriori algorithm, and presented optimization algorithms. Optimization methods currently used include partition-based methods, hash-based methods, parallel methods and the use of sampling methods. This paper also discusses that the mapping-table converts the items in the data sets to get the idea of compressing data set. When counting the support, the use of **hash_tree** disperses the matching of candidate itemsets to decrease the time complexity of algorithm.

II. RELATED WORKS

Algorithms for mining association rules from relational data have been done since long before. Association rule mining was first presented at 1993 by R. Agrawal, T. Imielinski, and A. Swami [1].After that many algorithms have been proposed and developed- Apriori [4] , DHP [5] , and FP growth [6] .The Apriori algorithm [2] employs a bottom-up breadth-first approach to get the large item set. As it was presented to hold the relational data this algorithm cannot be applied directly to mine complex data. Another recognized algorithm is FP growth algorithm. It uses divide-and-conquer approach. First it calculates the frequent items and characterizes the frequent items in a tree called frequent-pattern tree. This tree can also exploit as a compressed database. The association rule mining is done on the compressed database with the help of this FP tree. This denotes that the dataset needs to be inspecting once. Also this algorithm does not need the candidate item set generation. Many modified algorithm and technique has been proposed by different authors. Such as FP- tree and COFI based approach is proposed for multilevel association rules. Here except the FPtree, a new type of tree called COFI- tree is proposed [8]. An Apriori based data mining technique is studied at [9]. In 2000, Attila Gyenesei discussed a significant way of mining association rules for market analysis [16-21]. The Traditional association rule mining algorithms can only be used to data mining problems with categorical attribute. For a data mining problem with quantitative attribute, it is necessary to transform each quantitative attribute into discrete intervals. The first is mining frequent itemsets with Apriori, and then generating association rules according to the frequent itemsets mined [22]. Agrawal discusses this with good example and implementing association rules method in the customer transaction. Apriori is using circulatory generation for searching frequent itemsets that produces (k+1) – itemsets from k – itemsets [20]. LI Pingxiang [25] presented a method explores the database to filter frequent 1-itemsets and then it gets the candidate frequent 2-itemset, 3-itemsets up to n-itemset by evaluating their probabilities in Equation. HUANG Liusheng [26] presented an algorithm BitMatrix, This algorithm is compared with the previously known algorithms, the Apriori and AprioriTid algorithms. The main task [17,18] of every association rule mining algorithm is to find out the sets of items that frequently appear together—the frequent itemsets. The number of database scans required for the task has been reduced from a number equal to the size of the largest itemset in Apriori . By looking at Carlos Ordonez [27], several problems come up when trying to discover association rules in a high dimensional data set. This paper [28, 29], observe and address the problem of constraint-based rule mining in dense data. For example, most association rule miners permit users to set an alternative measure such as lift (Berry and Linoff, 1997; International Business Machines, 1996) or conviction (Brin et al., 1997). Then, this paper pointed out additional measures for identifying interesting rules, including lift and conviction. Daniel kunkle [30] pointed out three possible straightforward solutions. First, mine all frequent generalized itemsets, and then eliminate the non-max ones. Second, mine max frequent itemsets in the ordinary case and third choice is to dynamically browse the lattice of all generalized itemsets.

III. DISCUSSION ON APRIORI ALGORITHM

BASIC TERMINOLOGIES

A.. Data Definition

Data mining is nothing but, the extract of knowledge from databases, which is classification. Data generally by organized in tables that hold a set of complete database environment. A table can be looked like a matrix. Each matrix row symbolizes an instance that is coupled with a test case to analyse. An example of an instance may be a teacher, while all the instances may be a population of college teachers to analyse. Each matrix column symbolizes all the values associated with the domain name.

B. Association Rule

Association rule mining is one of the best discussed models for data mining.

Definition- The discovery of association rules from a transaction database DB, Y={y1, y2 ….. yn} be array of n different element called itemsets in DB, each transaction T in DB is a set of item (i.e. itemsets)

Support

The support of an association pattern is the percentage of task-relevant data transactions for which the pattern is true.

Confidence

Confidence is defined as the measure of certainty or trustworthiness associated with each discovered pattern

The Apriori algorithm uses a bottom-up breadth-first approach to find the large item set. In Apriori, in each iteration (or each pass) it creates a candidate set of large itemsets, counts the number of occurrences of each candidate itemset, and

then decides large itemsets based on a pre-determined minimum support. In the first iteration, Apriori simply scans all the transactions to count the number of occurrences for each item. Minimum support and minimum confidence are two important indicators of association rules. The algorithm uses apriori principle to generate candidate k-itemsets from frequent (k-1)-itemsets, and prunes candidate itemsets. Through the support counting, get candidate k-itemsets. Then the candidate k-itemsets generate frequent (k-1)-itemsets, so back and forth, until the frequent itemsets can not be produced. All the frequent itemsets will get association rules according to rules generation.
The following section discusses that the mapping-table converts the items in the data sets to get the purpose of compressing data set. When calculating the support, the utility of **hash_tree** disperses the matching of candidate itemsets to decrease the time complexity of algorithm. At the same time, **fp-growth** algorithm only scans transaction data sets twice.

IV. DISCUSSION ON THE BUILDING PROCESS OF HMT

Basic Terms

HMT - The data sets of characters in the file mapped to the integer, to enhance the matching effectiveness, and decrease memory footprint.
hash_ tree - Utilize hash function to build a special data structure for candidate itemsets.
hash_ node - The node of hash tress, it has branch_ node and leaf_ node
branch_ node - Be used for connect leaf node
leaf_ node - Be used for connect bucket
bucket - It has itemsets inside HMT(HASH MAPPING TABLE) is a One-dimensional mapping table of the key-value pairs, which bases on hash function, it is intended at compress the data sets, to decrease memory footprint. When the HMT is recognized, it can query the mapped value according to the terms, or in turn. Its main value is a string type of item, and its value is integer. Association rules mining for itemsets is all handling its main value.

Building process

1) Take the data set line by line, according to the list separator to separate the items.
2) Inquire the HMT whether it has that item.
3) If the item already present, ignores it; or else, adds it into the HMT through the hash.
The above process can incorporate with the compression of data set, the incorporated process is following:
1) Take the data set line by line, according to the list separator to separate the items.
2) Inquire the HMT whether it has that item.
3) If the item already present, ignores it; or else, adds it into the HMT through the hash.
4) At the same time, include the compressed data into the memory map of the data set.

Processing and analysis HMT

In the Apriori algorithm, when utilize the HMT to compress the data sets, the processing of itemsets have become the treatment of integer data. That incorporate the comparison of itemsets, HASH computer, the generation of subsets etc. When compress the itemsets, it will expend some handle time, but the time of support counting in Apriori is more than the compress time.

V. CONCLUSIONS

This paper discusses algorithms used in mining the training data set, which can discover implicit and potential useful knowledge from large preprocessed databases. This paper discusses an enhanced version of Apriori algorithm. Another approach discussed here is the HASH MAPPING TABLE tactic used to optimize space complexity and time complexity.

REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami.. *Mining association rules between sets of items in larged databases,* In Proceedings of the 1993 ACM SIGMODInternational Con nference on Management of Data, pages 207-216, Washington, DC, May 26-281993.
[2] R. Agrawal and R, Shrikanth, *Fast Algorithm forMining Association Rules*,Proceedings Of VLDB conference, pp 487 – 449, Santigo, Chile, 1994.
[3] J. Han and M. Kamber,*Data Minning: Concepts and Techniques* Morgan Kaufman, San Francisco, CA,2001.
[4] J. S. Park, M.-S. Chen, and P. S. Yu, *An effective Hash-Based Algorithm for Mining Association Rules* ,Proceedings of the ACM SIGMOD, San Jose, CA, May1995, pp. 175-186.
[5] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur, *Dynamic Itemset Counting and Implication Rules for Market Basket Data,*, Proceedings of the ACM SIGMOD, Tucson, A AZ,May 1997, pp. 255-264.
[6] J. Han, J. Pei, and Y. Yin, *Miniing Frequent Patterns without Candidate Generation,* Proceedings of th e ACM SIGMOD, Dallas, TX, May 2000, pp. 1-12.

[7] Qin Ding and gnanasekaranSundaraj, A*ssociation rule mining from XML data*,,Proceedings of the co onference on data mining.DMIN'06
[8] VirendrakumarShrivastava, Dr.parveenkumar and DR. K.R.pardasani, *FP-Tree and COFI Based App roach for Mining of Multiple Level Association Rules in Large database* , IJCSIS, International Journal of Computer Science and Information Security,Vol.7 No. 2,2010
[9] Irina Tudor, *Association rule mining as a data mining technique*, ,BULETINULuniversitatii Petrol-Gaze din Ploiesti,Vol.LX No1/2008,page 49-56.
[10] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur, "Dynamic Itemset Counting and Implication Rules for Market Basket Data,"
[11] C. Silverstein, S. Brin, and R. Motwani, *Beyond Market Baskets: Generalizing Association Rules to Dependence Rules*, Data Mining and Knowledge Discovery, 2(1), 1998, pp 39–68
[12] Qihua Lan,Defu Zhang, Bo Wo , *A new algorithm for frequent itemset mining based on apriori and FP-tree*, Global Congres on Intelligent System,2009.
[13] Jiawei Han, jian pei, and Yiwen Yin, *Mining frequent patterns without candidate generation*, paper id :196, SIGMOD '2000.
[14] N.Balaji Raja, Dr. G.Balakrishnan, A *Model of Algorithmic Approach to Itemsets Using Association Rules.*
[15] Zhiyong Zeng, Hui Yang, Tao Feng, *Using HMT and HASH_TREE to Optimize Apriori Algorithm*, 2011 International Conference on Business Computing and Global Informatization
[16] Hamid Mohamadlou, *A method for mining association rules in quantitative and fuzzy data*, 978-1-4244-4136-5/09, IEEE,pp.453 –458 Year of Publication 2009..
[17] Agrawal R, Srikant R. *Fast algorithms for mining association rules*, In: Proceedings of the 20th VLDB Int'l Conf., pp 487–499,1994.
[18] Agrawal R, Srikant R, *Fast algorithms for mining association rules in large database*, Technical report Fj9839, IBM Almaden Research Center, San jose, CA, jun.1994.
[19] Agrawal R, Imielinski T, Swami A. *Mining association rules between sets of items in large database,* In Proc,ACM SIGMOD, May 1993,pp.207-216.
[20] XindongWu and et al, *Top 10 algorithms in data mining*, Knowl Inf Syst ,Springer-Verlag London Limited ,pp 14:1–37,2008.
[21] Attila Gyenesei, *A Fuzzy approach for mining quantitative association rules*, Technical Report: TUCS-TR-336 Year of Publication, 2000.
[22] Zhu Ming, datamining, University of Science and Technology, china Press, Hefel, pp: 115 – 126, 2002.
[23] Balaji Raja.N and Balakrishnan.G, *Evaluation of Rule Likeness Measures for a Learning Environment*, In: Proceedings of the ICOREM Int'l Conf., pp: 1682 – 1690, 2009.
[24] Ahmed Riadh BABA-ALI, *A Novel Two Level Evolutionary Approach For Classification Rules Extraction*, IEEE Congress on Evolutionary Computation (CEC), pp 3306 – 3313, 2009.
[25] LI Pingxiang, CHEN Jiangping, BIAN Fuling, *A Developed Algorithm of Apriori Based on Association Analysis*, Geo-spatial Information Science ,Vol. 7, Issue 2, pp 108-112, june 2004.
[26] HUANG Liusheng, CHEN Huaping, WANG Xun, CHEN Guoliang, *A Fast Algorithm for Mining Association Rules*, J. Comput. Sci. & Technol., Vol. 15 No. 6, pp 619-624, Nov. 2000.
[27] Carlos Ordonez, Norberto Ezquerra, Cesar A. Santana, *Constraining and summarizing association rules in medical data*, Knowl Inf Syst, 9(3), pp 259-283, 2006.
[28] Berry M.J.A and Linoff, G.S., *Data Mining Techniques for Marketing Sales and Customer Support*, John Wiley & Sons, Inc., 1997.
[29] Brin, S., Motwani, R., Ullman, J., and Tsur, S., *Dynamic itemset counting and implication rules for market basket data*, In Proc. of the ACM-SIGMOD Int'l Conf. on the Management of Data, PP. 255-264, 1997.
[30] Daniel Kunkle, Donghui Zhang, Gene Cooperman, *Mining Frequent Generalized Itemsets and Generalized Association Rules Without Redundancy*, J. Comput. Sci. & Technol., Vol. 23(1), pp. 77-102, JAN. 2008.

