AVL Tree Based Approach for Efficient Mining

Renju Samuel , V.Ulagamuthalvi , P.Asha

***Abstract* -In the modern world the process of identifying the required items from a large dataset leads to a large amount of time. The concept of Data Mining is normally used to isolate the required item from a large database. In this paper a new form of isolating the data from a database is used. The traditional Apriori Algorithm is used to generate the Association rules by obtaining frequent itemsets of 1 combination, 2 combinations and so on based on the minimum support count. This study aims to propose a modified Apriori algorithm in which generation of Association rules is by using a Boolean Matrix representational scheme. At first the data representation takes place via the generation of an AVL tree which is used to balance the given data. Then the Boolean Matrix gets generated from the AVL tree and then association rules get generated from the matrix. Its shows clearly that the computation time has improved for the modified algorithm when compared to the traditional Apriori Algorithm**

Index terms: Association Rule, AVL Tree, Boolean Matrix, Data Mining, Frequent Itemset.

———————————— ◆ ————————————

.

1 - INTRODUCTION

The process analysing data from different perspectives and categorising it into useful information. It allows the users to analyse the database from different angles. Technically data mining can be viewed as the process of finding pattern among sets of fields among relational databases. Data mining is used in business to discover pattern, trends, relationships, anomalies in order to make better decisions. Apriori algorithm is designed to operate on transactional databases. It identifies the frequent individual items in the database and extends them to larger item sets as long as those appear sufficiently often in the database.

A "bottom up" approach where frequent subsets are extended as one item at a time and groups are tested against the data is been used by Apriori algorithm. Number of items in the item sets are counted using breadth-first search and hash tree mechanisms in Apriori algorithm. Association rule which highlights the general trends in data bases can be determined by frequent sets which results from Apriori algorithm. The popular method for discovering interesting relations between variables in large databases is called association rule. It identifies strong rules discovered in databases using different measures. Association can be used to discover regularities between products in large scale transaction of data's in supermarkets. For example, the rule

{Milk, Coffee powder} ⇨ {Sugar}

From the sales data of a supermarket would indicate that if a customer buys milk and coffee powder together is likely to buy sugar also. Such information's can be used as the basis for decisions about business and marketing activities such as pricing and product placements. Association rule has many applications including Bio informatics, Continuous production, Web usage mining and intrusion detection. It doesn't usually consider the order of the item either within a transaction or across the transaction. Apriori algorithm uses two rules they are

1) Find all the frequent sets, each of the frequently occurring at least as frequently

as minimum support rule which is predetermined.
2) Generate association rules from the frequent item sets.

Mining association rule is converted to find the frequent item sets. But traditional Apriori algorithm has two drawbacks they are

1) Many input –output operations are needed to scan databases frequently. It needs to scan the transaction database as the same number as the elements of the frequent item sets.
2) When analysing the databases we get large number of tables, which also requires a lot of time and memory (storage space).

To overcome these drawbacks we propose modified Apriori algorithm which uses data representation schemes. The algorithm is improved through AVL-tree based scheme which provides more reduced time and condensed rule than the traditional Apriori method. Using this method it is able to get last frequent item sets and generates association rule depending on the data representation and Boolean matrix. In Apriori algorithm we need to access the database frequently which consumes both time and memory space. In modified Apriori algorithm which is based on data representation and Boolean matrix reduces this overhead.

2 - RELATED WORK

Agarwal R and R Srikant suggested two new algorithms named as Apriori and Apriori Tid are used, for discovering all association rules between items in a large database. Finally they compared these algorithms to previously known algorithms such as AIS and SETM and the new implemented algorithms are much more effective than the existing algorithms [1].

Mannila, H.H Toivenn and A. Verkamo proved that an algorithm that uses all existing information between database, and the extra step they have done is that it passes to avoid checking the coverage of duplicate itemsets . It provides a better execution time when compared to the previous results and it is simple to implement. Concluded in a way that sampling is an efficient technique for finding rules of this type and that algorithms working in the main memory can be used to obtain very good approximations [2**].**

Rakesh Agrawal and Manish Mehta and John Shafer and Ramakrishnan had used a way to discover patterns present in very large databases other than to verify that a certain pattern exists in that database. The result obtained produced high performance and also linear scaling in real life databases. Also they had provided a completeness property that promises all the similar patterns present in the database of certain type will be identified [3].

Hassan M. Najadat, Mohammed Al-Maolegi, Bassam Arkok have used several groups of transaction, and various minimum support count values in the database. And they have applied the same values to both the normal and improved Apriori algorithm. The conclusion obtained is that the new improved Apriori algorithm had improved the time consumption by 67.38%. Also another result is that it is more efficient when compared to the normal apriori algorithm [4]. t

Bay and Bac Le have presented a new form of Apriori algorithm for obtaining the frequent itemsets and to generate the association rules. They scan the database only once in order to generate the transaction ids to obtain the frequent itemsets. A tree structure is used to obtain these results. They omit the duplicate itemsets in order to obtain the results of generating the association rules at a faster phase [5].

An Effective Hash Based Algorithm for Mining Association Rules by Hao Y, Ye Li and Cheng Z , they had used Hash function instead of normal joining method. Transactionreduction is done by generating hash function using DHP algorithm. Collision in hash table reduces the effectiveness of hash table. Hash functions are complex to be coded after 2-itemset combination. (DHP- Direct hashing and pruning) [6].

Dhilhan perera, Judy Kay, Irena Koprinska, Kalina Yacef , and Osmar Zaiane have proposed high level view of information from the database, that is

together they pair the patterns of the same group. Their goal was to enable the particular groups and their identities the various operations of the group and the feedback that is provided during the operations. The tool used in order to obtain the result is TRAC. They extract patterns of the strong groups eliminating the weaker groups to get more successful results [7].

Shinji Funjiwara, Jeffrey D. Ullman and Rajeev Motwani had proposed a way to find all implication and redundancy rule based on confidence pruning without support pruning. Dynamic pruning done to the large data set by counting number of rows and avoid the rows containing missing values. Only confidence pruning is considered. They focused only on the reduction in the transaction size and not on efficient rule generation [8].

3 - PROPOSED WORK

Initially the data present in the given database gets sorted out in ascending order by using the AVL tree .The next step is the data discretization of the sorted data. Here the sorted values from the AVL tree gets replaced according to the Index values provided to it. After this particular step the Hash Map table is being generated using the values obtained in the previous step. The Hash Map table consist of transaction id and the List of items present in the list. From these values the Boolean Matrix is obtained. All the places at which the value is present are termed as 1 and the place at which value is not present is termed as 0. This result is formed by using the Hash Map table. And the last step is generation of Association Rules from the Boolean Matrix.

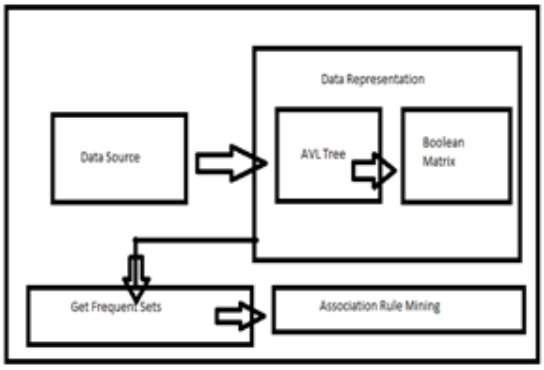

Fig. 1. Architecture Diagram

Data representation transforms source data into an AVL tree which is a balanced structure. Also it reduces the computational time for the process. Also it obtains various multi data types of the represented data from the database such as a hash table, tree and Boolean matrix.

Shown below (Table.1) is an example for the generation of Association rules using the modified approach.

Table 1 : Given Dataset

Transaction ID	List Of Items
T101	8,14
T102	10,12,14
T103	8,12,14
T104	10,14
T105	8,24
T106	12,14
T107	8.12
T108	12,14

Here the given value gets balanced by the AVL Tree from the database (Fig. 2).

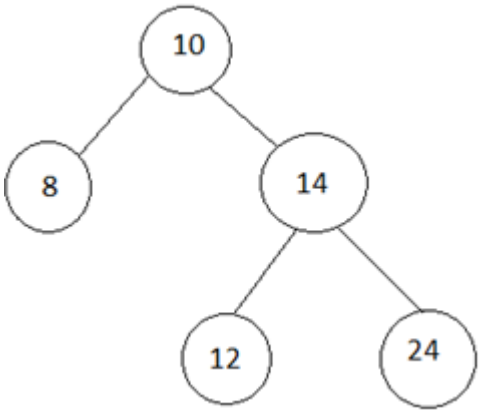

Fig. 2. Generation of AVL Tree

After the balanced values are obtained from the AVL Tree, the data discretization (Table.2) is done by replacing the elements present at the particular level in the database.

Table 2: Data Discretization

1	2	3	4	5
8	10	12	14	24

Next step is the generation of Boolean Matrix (Table.3), it is done by substituting the values which are obtained from the previous step, values are substituted as 1 in the place of element is present and 0 at the place where presence of element is null.

Table 3: Boolean Matrix

	1	2	3	4	5
T101	1	0	0	1	0
T102	0	1	1	0	1
T103	1	0	1	1	0
T104	0	1	0	1	0
T105	1	0	0	0	1
T106	0	0	1	0	1
T107	1	0	1	0	0
T108	0	0	1	1	0

Hence the Boolean Matrix is generated. Now the next step is to find the frequent itemsets from the obtained Boolean Matrix

$$\text{Computed Weight} = \frac{\text{Count 1's in Column}}{\text{Number Of Rows}} >= S$$

Here S is the minimum support count given by the user

Table 4: Generation of Frequent Itemsets

	1	2	3	4	5
T101	1	0	0	1	0
T102	0	1	1	0	1
T103	1	0	1	1	0
T104	0	1	0	1	0
T105	1	0	0	0	1
T106	0	0	1	0	1
T107	1	0	1	0	0
T108	0	0	1	1	0
COUNT	4	2	5	4	3
Count / Total Number of Rows	4/8 = 0.5	2/8 =.25	5/8 =.625	4/8 =.5	3/8 =.375

Here all 1's are counted in each column and it is added and divided with the total number of rows. Let us take the minimum support count (S) as 0.5, so all values which are equal to and greater than equal to 0.5 is taken into consideration. Rest all values are rejected (Table.4).

Hence from the following table, the values 0.5, 0.625, 0.5 are selected. Its following values are taken inorder to generate the Association Rule from the table

In this example the last frequent itemset are {1, 3, and 4}

The following Rules can be generated

Rule 1: 1^3 →4, **Rule 2**: 1^4 →3

Rule 3: 3^4 →1, **Rule 4**: 1 →3^4

Rule 5: 3→ 1^4, **Rule 6**: 4 → 1^3

Thus the Association Rules are generated from the given dataset using the Modified Apriori Algorithm

4 - CONCLUSION

The following study had provided suitable information such that with the data representation scheme that is followed here reduces the computational time. It also improves the various processes such as the generation of n-frequent itemsets and also the generation of the Association Rules when compared to the normal Apriori Algorithm. The major advantage of using this type is that the entire database needs to be scanned only once inorder to generate the Association Rules. Thus it saves a lot of computational time than the normal Apriori, which has to invoke the database each and every step inorder to generate its Association Rules.

REFERENCES

[1]. **Agarwal R and R Srikant, 1994 , "Fast Algorithms for Mining Association Rules in large databases"** : Jorge B Bocca , Matthias Jarke and Carlo Zaniolo, editors , Proceedings of the 20th International Conference on Very Large Data Bases , VLDB , Santiago , Chile

[2]. **Mannila , H.H Toivenn and A. Verkamo , 1994** : **Efficient algorithm for discovering association rule**s . AAAI workshop on Knowledge Discovery Database Seattle 181-92

[3]. **The Quest Data Mining System : Rakesh Agrawal and Manish Mehta and John Shafer and Ramakrishnan Srikant** , IBM Almaden Research Center San Jose, California 95120, U.S.A 1996.

[4]. **An Improved Apriori Algorithm for Association Rules , Hassan M. Najadat, Mohammed Al-Maolegi , Bassam Arkok**, International Research Journal of Computer Science and Application Vol. 1, No. 1, June 2013, PP: 01 – 08

[5]. **Fast Algorithm for Mining Generalized Association Rules , Bay Vo and Bac Le** , International Journal of Database Theory and Application Vol. 2, No. 3, September 2009

[6]. **Effective Hash Based Algorithm for Mining Association Rules by Hao Y, Ye Li and Cheng Z,** SIGMOD '00 Proceedings of the 2000 ACM SIGMOD international conference on Management of data

[7]. **Clustering and Sequential Pattern Mining of online Collaborative Learning Data , Dhilhan perera , Judy Kay , Irena Koprinska** , Member , IEEE Computer, Society , Kalina Yacef , and Osmar Zaiane , Member , IEEE IEEE Transactions On Knowledge And Data Engineering , Manuscript ID 2011

[8]. **Dynamic Miss Counting Algorithm Finding Implication and Similarity Rules with Confidence Pruning,Shinji Funjiwara, Jeffrey D. Ullman and Rajeev Motwani ,** Data Engineering, 2000. Proceedings. 16th International Conference on 29 Feb 2000-03 Mar 2000 ,501 - 511

[9]. **A study on the Performance of CT-APRIORI and CT-PRO Algorithms Using Compressed Structes for Pattern Mining ,** Journal of Global Research in Computer Science ,Volume 1 , Number 2 September 2010

[10]. **The Quest Data Mining System : Rakesh Agarwal , Manish Mehta , John Shafer and Ramakrishnan Srikanth ,** IBM Almaden Research Centre , San Jose, California 95120, U.S.A 1996

[11]. **Saravanan Suba and Chistopher.t. Article: A Study on Milestones of Association Rule Mining Algorithms in Large Databases .** International Journal of Computer Applications 47(3):12-19, June 2012. Published by Foundation of Computer Science, New York, USA

[12]. **Mining Sequential Patterns: Generalizations and Performance Improvements Ramakrishnan Srikant and Rakesh Agrawal F Srikant**, IBM Almaden Research Center 650 Harry Road, San Jose,CA95120

[13]. **Information Retrieval through Semantic Web, Gagandeep Singh, Vishal Jain ,** International Journal Of Database , Vol 8, No 5june 2012

[14]. **Mining Association Rules Based on Cloud Model and Application in Credit Card Marketing , Yan-li Zhu, Yu-Fen Wang , Shun-Ping Wang , Xiao-juan Guo , Shenzhen,** China , pp.165-168, 2010 Asia-Pacific Conference on Wearable Computing Systems, 2010

Renju Samuel is currently pursuing masters degree program in Computer Science and Engineering in Sathyabama University , India, PH- +91 9952037819. E-mail: renju.samuel@gmail.com

V.Ulagamuthalvi is working as Associate Professor in Computer Science department in Sathyabama University, India PH- +91 9645105144. E-mail: muthalvi73@gmail.com

P.Asha is working as Research Scholar in Computer Science department in Sathyabama University, India PH- +91 9944992337 E-mail: ashapandian225@gmail.com