iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 2 **Issue:** II **Month of publication:** February 2014

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

A Novel Approach of Association Mining with Apriori Algorithm to Extract the Frequent Itemsetset in OOD

Rekha Gulia
Student, DITM Gannaur

Abstract: Most Of the organization have major requirement of Distributed Centralized database system. Each user presents at specification location access the dataset according to his requirement. It means only the partial information is required for each database location. Because of this instead of provide the complete database information to each user only the selected information is provided, this concept is called partitioning. The partitioning can be horizontal or the vertical. Some times because of the security reasons it is required to hide the sensitive information from user, in such also the partition is performed. The presented work is an intelligent approach that will perform the analysis on the user data requirement and accordingly the partition will be performed and the data information present to the user. In this proposed approach we are providing a dynamic approach to create dynamic partitions of the database on the basis of their frequent requirement. The proposed approach will perform an association rule mining along with Apriori Algorithm. The proposed algorithm will return the significant results in efficient Time.

Keywords: - Frequent, Apriori, Association, Partitions, Mining

I INTRODUCTION

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

A Data Warehouse is the collection of large dataset driven from different data centers. Data warehousing is a practical solution for providing strategic information. All the analytical Decision about the organization is taken based on warehouse dataset. It is helpful to the management to take high level decisions. According to this data warehouse contains dataset respective to subject not according to operational data. Data Warehouse is of different types defined as under (i) Enterprises Data Warehouse (ii) Data Mart (iii) Virtual Warehouse.

The ideal starting point is a data warehouse containing a combination of internal data tracking all customer contact coupled with external market data about competitor activity. Background information on potential customers also provides an excellent basis for prospecting.

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks.

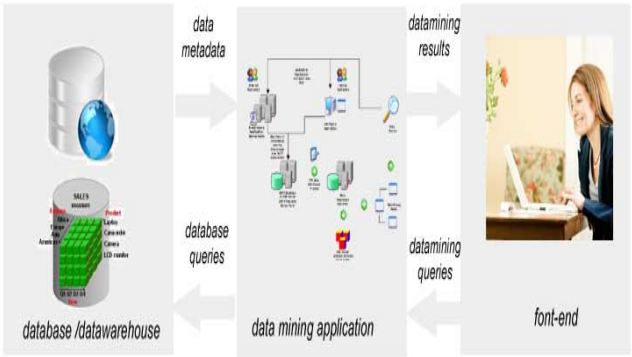


Figure 1: Integrated Data Mining Architecture

- **Anomaly detection** – The identification of unusual data records, that might be interesting or data errors that require further investigation.

II LITERATURE SURVEY

Claudio Lucchese performed a work,**"** Fast and Memory Efficient Mining of Frequent Closed Itemsets**".** His paper presents a new scalable algorithm for discovering closed frequent itemsets, a lossless and condensed representation of all the frequent itemsets that can be mined from a transactional database. Presented algorithm exploits a divide-and-conquer approach and a bitwise vertical representation of the database, and adopts a particular visit and partitioning strategy of the search space based on an original theoretical framework, which formalizes the problem of closed itemsets mining in detail. The algorithm adopts several optimizations aimed to save both space and time in computing itemset closures and their supports[1].B.Murugeshwari performed a work on database extension to hide the sensitive information. It includes the process of minimizing the information while maintaining the privacy as well to save the required information. In this approach the cryptographic approach is used for the authentication [2]. In year 2007 Oman Abdul provide the solution to perform the knowledge hiding by using the polynomial sanitization algorithm. The author used the experimental evaluation approach to perform the related work. The work will be implemented on complex dataset as well as applications [3]. E. Ansari performed a work,**"** Distributed Frequent Itemset Mining using Trie Data Structure**".** The Author proposed a new distributed trie-based algorithm (DTFIM) to find frequent itemsets. This algorithm is proposed for a multi-computer environment. In second phase Author added an idea from FDM algorithm for candidate generation step. Experimental evaluations on different sort of distributed data show the effect of using this algorithm and adopted techniques[4].

Anita A. Parmar,Udai Pratap Rao, Dhiren R. Patel, address such the problem of sensitive classification rules hiding. The author proposes a blocking based approach for sensitive classification rule hiding. In this approach at first find the supporting transactions of sensitive rules. Then replace known values with unknown values ("?") in those transactions to hide a given sensitive classification rule. Finally the sanitized dataset is generated from which sensitive classification rules are no longer mined. We also discuss experimental results of our algorithm [5]. An evolution framework is presented by Osman Abul, Harun. It work includes the implementation of recent algorithms on different dataset to analyze the results involving the proposed architecture and the problem. The work is also tested with data distortion and runtime requirements, especially for difficult problem instances [6].

As we are defining the proposed approach to estimate the frequent dataset to estimate the user current requirement on the basis of previous utilization of the user. Such kind of approach is called prediction approach. The prediction is another important aspect of data mining. Here some other work on prediction is defined in literature.

Yongyi perform a prediction based work on economic data. It also describes importance that the application of economic data in the data mining algorithm and its application, which combines with the current economic data of national macro-economic indicators, present the data warehouse model structure. In this work data mining solutions on economic data for the application of data mining solution, system architecture, algorithms implementation, and finally discusses the application of data mining algorithms development trends and key technologies in the economic field[7].

In year 2010, Shiguo wang defined an analytical work to analyze the fraud by analyzing the financial data. This work is the intelligent approach that used the Bayesian network, and stack variables etc. It also includes the Regression Analysis to perform the data hiding the detecting effect and accuracy of NN is superior to regression model [8]. Ferenc Bodon performed a work**," A Trie-based APRIORI Implementation for Mining Frequent Item sequences".**The Author investigate a trie-based APRIORI algorithm for mining frequent item sequences in a transactional database. Author examine the data structure, implementation and algorithmic features mainly focusing on those that also arise in frequent itemset mining. In Presented analysis Author take into consideration modern processors' properties (memory hierarchies, prefetching, branch prediction, cache line size, etc.), in order to better understand the results of the experiments[9].

Bin Sheng performed a work,**" Data Mining in Census Data with CART.** Using Data Mining in census data can make full use of these data to provide services for country's social and economic development. Classification is one of the important Data Mining techniques. The Decision Trees of classification analysis can give high accuracy prediction results, and the output results are easy to understand [10]

III PROPOSED WORK

The proposed is about the study and analysis of a large database or the data warehouse to predict the frequent dataset requirement and to provide the necessary information to it. This

concept of providing the selective information to a distributed user is called partitioning. For any database it is never easy to transfer the complete database for each user query to some distributed machine. The partitioning is generally performed in two ways.

Horizontally Partitioning

The horizontal partitioning includes the selection of data requirement of user in terms of key attribute. Example would be having 53 partitions created and storing data of the Sales Fact table as follows - data belonging to first week of the year goes into first partition, data belonging to the second week goes into the second partitions and so on. This example is applicable if the business wants to report daily sales data for the past 1 year only. Incase of reporting needs requiring longer duration, we would create as many partitions. This operation is basically performed by using the Apriori Algorithm in this proposed work.

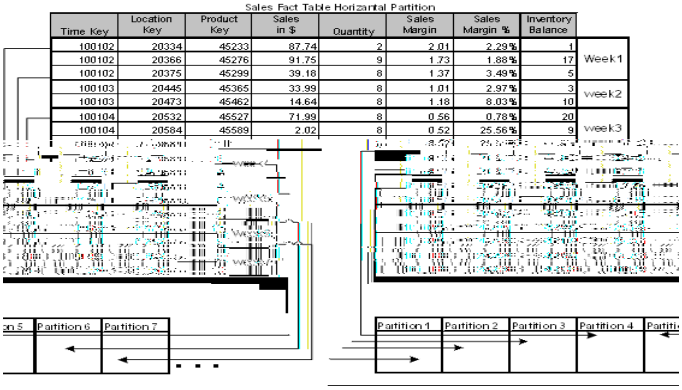


Figure 2: Horizontal Partitioning

Vertical Partitioning

The vertical partitioning includes the selection of attribute requirement for a specific distributed data user. Here we store data of a column or a set of columns of the table into multiple partitions. This would be more appropriate for a table which had some columns which are rarely used. Another way of looking into a vertical partitioning is to **normalization,** splitting a table into multiple tables for avoiding redundancy. To perform this association mining will be performed. The most associated attributed along with the required attributes will be accessed by the user.

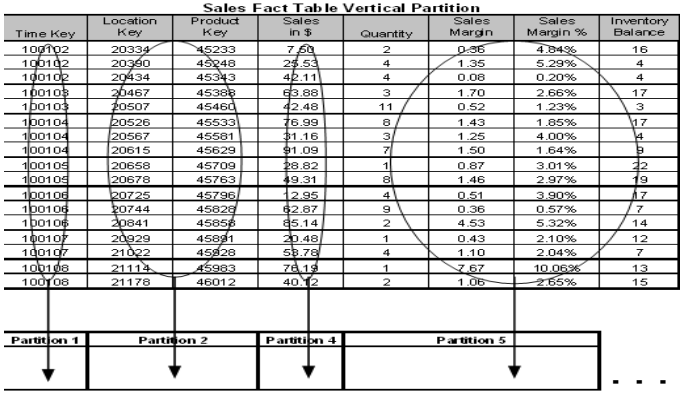


Figure 3: Vertical Partitioning

In this proposed work Apriori algorithm is used for selecting items to sanitize among the transactions supporting sensitive item sets. In this work at first dataset classification will be performed. All the valid modifications related to the sensitive rules, the non sensitive rules, and the spurious rules that they can affect when applied. After that heuristic methods will be applied to increases the number of hidden sensitive rules, while reducing the number of modified entries.

At first the basic approri will be implemented to reduce the support of the sensitive item sets by deleting a set of supporting transactions. After that the approach will modifies the database dynamically and instead of deleting, the supporting transactions by removing some items until the sensitive item sets are protected. Finally it aggregates the previous two by using the first approach to identify the sensitive transactions and the second one to remove items from these transactions, until the sensitive knowledge is hidden.

The base algorithm used in this proposed work is the apriori algorithm.

Apriori Algorithm

Apriori is the best-known algorithm to mine association rules. It uses a breadth-first search strategy to count the support of itemsets and uses a candidate generation function which exploits the downward closure property of support.

- Usefulness of a rule can be measured with a minimum support threshold.
- This parameter lets to measure how many events have such itemsets that match both sides of the implication in the association rule.
- Rules for events whose itemsets do not match boths sides sufficiently often (defined by a threshold value) can be excluded
- Database D consists of events T_1, T_2,… T_m, that is D = {T_1, T_2,…, T_m}
- Let there be an itemset X that is a subregion of event T_k, that is $X \subseteq T_k$

- The support can be defined as

$$sup(X) = \frac{|\{T_k \in D \mid X \subseteq T_k\}|}{|D|}$$

- This relation compares number of events containing itemset X to number of all events in database.

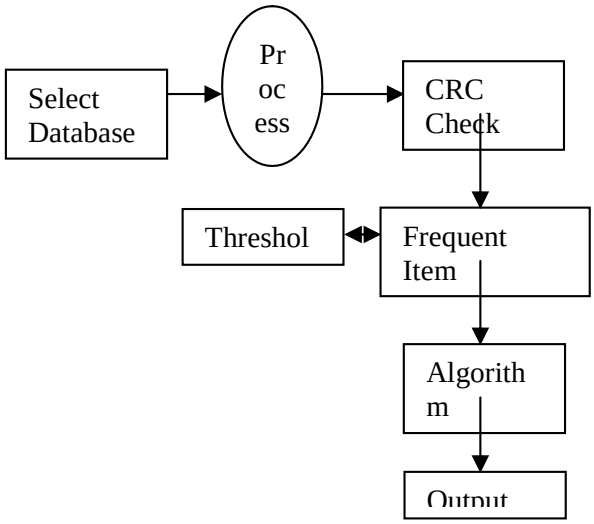


Figure 4: Frequent Dataset Selection

In figure 4, the basic algorithmic process is defined. At first the dataset is selected and some basic operation is performed on to it. This process includes the cleaning and the filtration process. Once the filtration performed it will check for data base error to deduce the invalid entries. Now the frequent dataset is analyzed. After analysis a threshold value will be decided and only the dataset values that will follow rule will be retained with the dataset. After this Apriori Algorithm will be implied to find the frequent dataset.

The Apriori algorithm calculates rules that express probabilistic relationships between items in frequent item sets For example, a rule derived from frequent item sets containing A, B, and C might state that if A and B are included in a transaction, then C is likely to also be included. An association rule states that an item or group of items implies the presence of another item with some probability. Unlike decision tree rules, which predict a target, association rules simply express correlation.

The combined approach of complete algorithm is presented as

1. Start
2. Load a sample of records from the database that fits in the memory.
3. Define the RuleSet in terms of associated columns to identify Frequent Dataset
4. Based on this ruleset Apply the Apriori algorithm On Filtered Dataset to find the frequent item sets with the minimum Support. Suppose A is set of the frequent item set generated by Apriori Algorithm.
5. Find AssociationValue of Each Column According to RuleSet.
6. Let Nvalue is number of records that follow the same RuleSet Represent it as the Support value identified as the frequent dataset
7. If Nvalue > MinSupport
   {
   Present it as Result Recordset
   }
8. Divide All Records in Seprate Partitions for Each Rule.
9. If the desired number of generations is not completed, then go to Step 3.
10. Stop

Table 1: Proposed Algorithm

After applying above dataset we get most necessary dataset for a user. The system will perform a user respective intelligent partitioning.

IV CONCLUSION

In the traditional approach the database extensions are generated on complete original database but this approach will not work where the clients requirements are dynamic. It means the client's frequent required database is changed. In this proposed approach we are providing a dynamic approach to create dynamic partitions of the database on the basis of their frequent requirement.

REFERENCES

[1] Claudio Lucchese," Fast and Memory Efficient Mining of Frequent Closed Itemsets**"**, Ieee Transaction On Knowledge And Data Engineering.

[2] B.Murugeshwari, Dr.K.Sarukesi, Dr.C.Jayakumar, "An Efficient method for knowledge hiding Through database extension", 2010 International Conference on Recent

Trends in Information, Telecommunication and Computing, 978-0-7695-3975-1/10 © 2010 IEEE.

[3] Osman Abul Maurizio Atzori Francesco Bonchi Fosca Giannotti, "Hiding Sequences", 1-4244-0832-6/07/ ©2007 IEEE.

[4] E. Ansari," Distributed Frequent Itemset Mining using Trie Data Structure"©2009 IEEE.

[5] Anita A. Parmar,Udai Pratap Rao, Dhiren R. Pate, "Blocking based approach for classification Rule hiding to Preserve the Privacy in Databas", 2011 International Symposium on Computer Science and Society, 978-0-7695-4443-4/11 © 2011 IEEE.

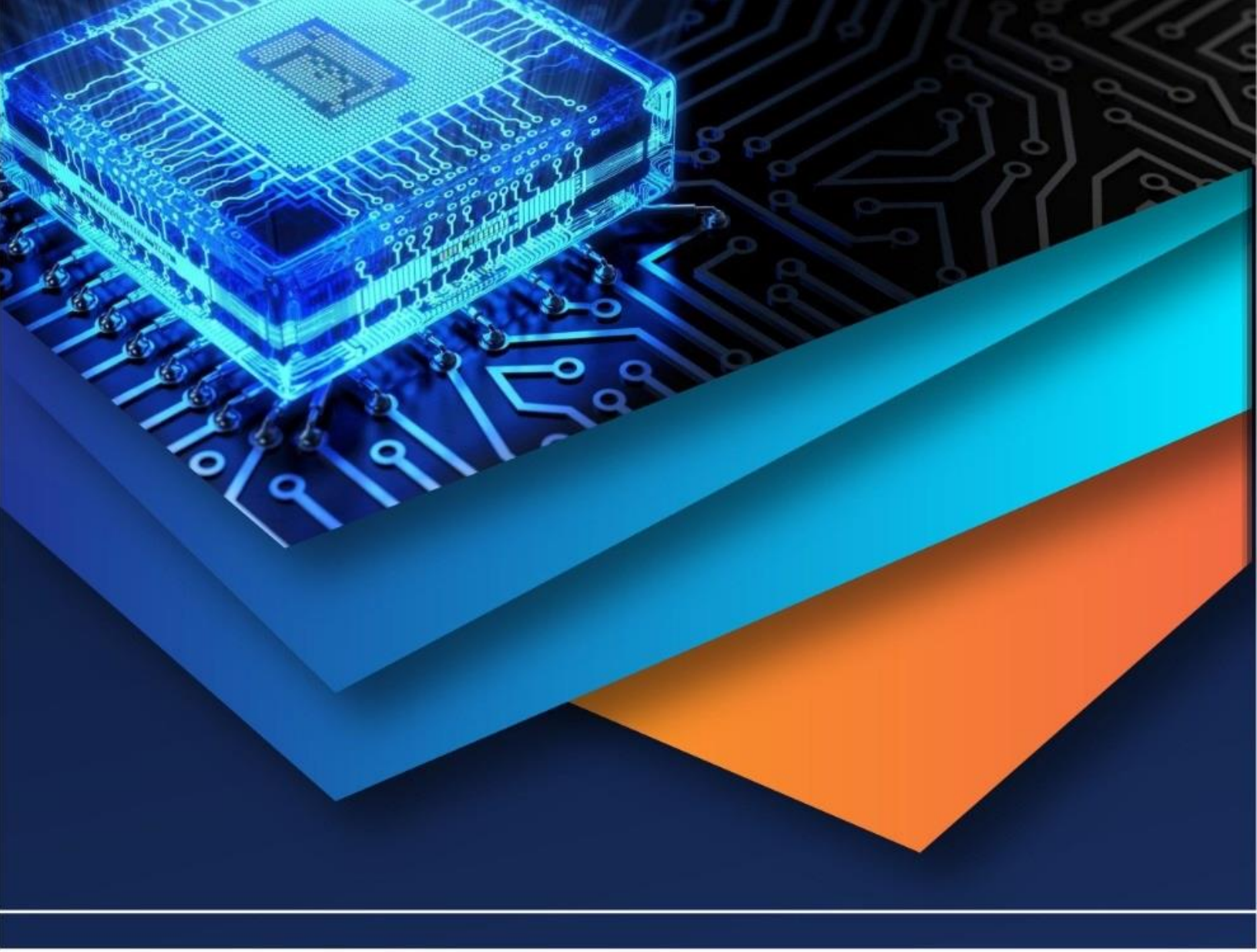
[6] Osman Abul, Harun G¨okc¸e, Ya˘gmur S¸engez [2009], **"**Frequent Itemsets Hiding: A Performance Evaluation Framework", 978-1-4244-5023-7/09.

[7] Yu et al. Yu Huang, Xiao-yi Zhang, Zhen Yuan, Guo-quan Jiang [AUG 2009], "A universal Frequent Pattern Analysis framework based on user model", IEEE, ISECS International Computing, Communication, Control, and Management, Sanya, China.

[8] Shiguo wang," A Comprehensive Survey of Data Mining-based Accounting-Fraud Detection Research", 2010 International Conference on Intelligent Computation Technology and Automation 978-0-7695-4077-1/10© 2010 IEEE.

[9] Ferenc Bodon," A Trie-based APRIORI Implementation for Mining Frequent Item sequences" © 2009 IEEE.

[10] Bin Sheng," Data Mining in Census Data with CART", 20IO 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE) 978-1-4244-6542-2© 2010 IEEE.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)