International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Techniques of OLAP and Association Rule Mining

Gunwanti R. Bawane[1], Prarthana Deshkar[2]

[1]PG Student, Dept. of Computer Science and Engineering, Yeshwantrao Chavan College of Engineering, Nagpur (MS), India

[2]Professor, Dept. of Computer Science and Engineering, Yeshwantrao Chavan College of Engineering, Nagpur (MS), India

Abstract: ***OLAP is a multidimensional view of complete data in the data store used for multidimensional analysis. It is the most practical approach used in the data warehouse for analytical process of large data and provides tools for analytical and statistical analysis of data. While Association rule learning is a popular and researched method for discovering interesting relations between variables in very large databases. It is intended to identify strong rules discovered in databases using different measures of interestingness. Apriori algorithm is used to discover the association rules and can be used to integrate with the OLAP systems.***

Keywords: OLAP, Association rule, Apriori algorithm, data cube.

1. Introduction

OLAP provides a multidimensional view of data in a data store which is used to analyze the data through any dimension of user interest. OLAP can be used in multidimensional view of complete data and in analysis of such data which provide analytical modeling tools to generate aggregated data. This can be used in trend analysis and statistical analysis of the data. It is an approach for answering multi-dimensional queries and runs the queries swiftly. OLAP enables end-users to perform analysis of data in multiple dimensions, thereby providing the insight and understanding they need for better decision making. OLAP is also a data summarization tool that is used for data analysis through queries fired by the user and returns back the desired result. OLAP uses multidimensional data model as a base which consists of fact table and dimensions tables, where fact is a numeric and dimensions are the measures used in the table. OLAP provides many operations which can be performed on the data cube as per the user need which include roll-up, drill-down, slicing, dicing, etc. MOLAP, ROLAP, HOLAP are some of the models of OLAP. A data cube consists of three dimensional data and provides the view of data which can be analyzed through any dimension. OLAP is an on-line analytical processing tool provided by data warehouses for the interactive analysis of multidimensional data of varied granularities, and it processes all kinds of analysis operations based on the data stored in data warehouse according to the users' questions and hypothesis and returns the analytical result back to users in comprehensible form. Its definition can be generalized as Fast Analysis of Shared Multidimensional Information [11].
Association rule mining is the technique which is used to find the associations between the data item in a data and is used for generating strong association rules between the items. The association rules generated is useful in making the market decisions which are useful in revenue generation. The rules generated are completely user dependent because the system provides the item sets in which the user is interested. Consider an example of store in a market and the user is interested to find the rules which are based on sales of bread, butter and milk. The system will provide the rule:

{Bread, butter} → {Milk}

The rule states that if the user buy bread and butter the user also tend to buy milk. These rules are useful in market basket analysis. These rules are generated by using some threshold values of Support and Count which are defined by the user. The rules generated which are above the threshold values are only considered to be the rules of interest. Association rules are a group of objects in the database which associated with the relationship between the rules. It is widely used in data mining [5]. There are many algorithms used for association rule but Apriori algorithm is the most traditional and widely used algorithm for rule generation.

Traditional Apriori algorithm has many drawbacks:

a) Generates a lot of candidate itemsets. So, there is a increase in the storage count of these candidate itemset which leads the algorithm to be inefficient.
b) The algorithm performs the repetative scan of data base every time the itemsets are generated in every iteration.

The Apriori algorithm is implemented by applying new techniques to increase its efficiency and new algorithm generation is classified into two categories: first is searching database by iterations and the second one is higher to lower dimensions [8].

2. Related Work

The most traditional and widely used algorithm used for association rule mining is the Apriori algorithm. As discussed in the previous section the algorithm has many shortcomings, so new techniques were applied on the algorithm to increase its efficiency.

Apriori_cube algorithm in [1] is discussed in detail and the advanced algorithm is implemented in order to increase the efficiency of the algorithm and integrate the OLAP systems and the association rule mining. The algorithm implemented used the concept hierarchy and used the drilling operations to adjust the hierarchy of dimension. The algorithm used the hash technique to optimize the algorithm. Detail explanation of step by step execution of the Apriori algorithm on the database of a store in studied in the paper [2] in which

frequent itemset are generated and the rules are generated from the frequent itemset. A novel approach was adapted by the authors in [3] which improved the Apriori algorithm and removed its drawback of data consistency by solving the problem of duplicate data. Data cleaning is done at the early stage to reduce the candidate sets by using the cleaning algorithm. The approach used focuses on finding the association on filtered dataset instead of whole data set which resulted in an efficient algorithm than the traditional Apriori algorithm. The space issue of the apriori algorithm is reduced by using the data compression techniques and developed an algorithm which is more efficient and manages space by avoiding duplication [4]. Applying association rule mining to power demand-side management can supply references to optimizing decision-making program by power-supply companies. [6] introduce the principle of advancing power demand-side management measure accepted by user using association rules, and discuss the feasibility of choosing load-adjusting electric-power terminal based on multi-dimensional association rule mining. [7] Paper describes a multi-dimension data cube model; put forward the formalized definition of the generalized association rule; and concluded two categories mining strategies of creating multi-level frequent itemset mining algorithms. GenHibFreqeh which was suitable for data cube, this algorithm used the abstract level among the itemsets adequately to decrease the number of the candidate item set which needs counting to improve the efficiency of this algorithms, we put forward the Algorithms of mining of the Generalized Association Rule Based on Data Cube (GenerateLHSs-Rule) which can decrease the creating of the redundant rules efficiently. A new algorithm for mining association rules in data cube was implemented in [9] which was framework on inter-dimension association rules from data cube. They used inter-dimensional association meta-rule which allows user to target the mining process in a particular portion of the data cube. The algorithm adapts the traditional Apriori algorithm which is a bottom-up algorithm. In [10] a model of mining association rule oriented data cube is presented by the authors. It focuses on construction of data cube using the facts and the dimension table.

Integration of OLAP and Association rule mining can be done by applying the association rule mining on the multidimensional data structure by making use of Apriori algorithm. Apriori_cube algorithm is based on the improved Apriori algorithm. Generation algorithm is based on the data cube, in which corresponding predicate set support count has been recorded in each box. Thus as long as the count-by-value data cube sweep seek their support, if its support is greater than the minimum support, the grid corresponding set of predicates (these items are from a different dimension to the dimension members) is a requirement for frequent predicate sets [1].

APRIORI_CUBE ALGORITHM

As discussed above the traditional Apriori algorithm has many drawbacks and to integrate it with the OLAP systems it is necessary to get the efficient rule which should consider the multidimensionality and hierarchical structure of the database. Apriori_cube algorithm is used to integrate the Association rule mining and the OLAP technology. According to the algorithm process can be seen, if the data cube is poor, only a few item set is frequent, especially when the number of the item set increase, frequent item sets are very few. The Apriori_cube algorithm by using Apriori properties of candidate sets of clip, greatly reduce the length of the candidate set, measure the level of reducing the time and improve the algorithm performance. If the data cube is very tight, the connection and cut time spent will be very long [1]. Traditional Apriori_cube algorithm has drawbacks:

a) It is necessary to consider the dimensional hierarchies before generating the rules in order to avoid boring and contrary rule generation.
b) Apriori rules generate a lot of candidate item sets. It increases the cost of storage and count.
c) Rule generated are redundant rule so the speed of rule generation decreases.

Improved Apriori_cube algorithm tries to make the algorithm more efficient by making use of several techniques to overcome its drawbacks. These include:

a) Use of concept hierarchy tree in mining process to adjust dimension of the hierarchy and to determine the level of dimension hierarchies. If the dimension hierarchy is too high then drilling operations can be used to reduce the dimension. Similarly if the dimension is low then drilling operation can be used to adjust the dimension accordingly.
b) Use of hash technology is used to optimize the algorithm by making small candidate itemsets. Apriori algorithm scans the database every time the candidate itemsets are generated so the key to optimize the algorithm used is to generate smaller set of itemset by removing unnecessary itemsets. Hash method is on the table between storage location and the key to establish a certain corresponding functional relationship. Each key corresponds to a single storage location in the structure: Address = Hash(Rec.key). During search the first function for the key code table is calculated and the function values as a storage location. After comparing the table items of these positions if the key is equal then search is successfully completed. In the table items are kept in accordance with the same function calculation storage location, and stored at this location. The hash function used is the compressed image function.
c) The new improved algorithm uses the nature of the frequent item sets through the depth of recursion method of generating frequent subset to accelerate the process of generating association rules. This also reduces the redundancy of association rule generation.

3. Conclusion

Apriori algorithm is discussed in detail in the above sections. The algorithm has some limitations but the study of various techniques and methods will help in increasing the efficiency of the algorithm. OLAP system and association rule mining can be integrated to provide a system which considers the multidimensionality of the database and can provide a system which generates the result which is more useful for the users for analyzing the database and make decisions. Apriori_cube algorithm can be used to integrate OLAP system and association rule mining.

References

[1] Yingjie Wang, Lili Yu, Qinrun Wen, Congli Liu, "Improved multi-level association rule in mining algorithm based on a multidimensional data cube", Consumer Electronics, Communications and Networks (CECNet), 3rd International Conference, 2007.
[2] Lugendra Dongre, Gend Lal Prajapati, S. V. Tokekar, "The Role of Apriori Algorithm for Finding the Association Rules in Data Mining", Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014.
[3] Chanchal Yadav, Shuliang Wang, Manoj Kumar, "An Approach to Improve Apriori Algorithm Based On Association rule Mining", Computing, Communications and Networking Technologies (ICCCNT), 2013.
[4] Du Ping, Gao Yongping, "A New Improvement of Apriori Algorithm for Mining Association Rules", International Conference on Computer Application and System Modeling, 2010.
[5] Libing Wu, Kui Gong, Yanxiang He, Xiaohua Ge, Jianqun Cui, "A Study of Improving Apriori Algorithm", Intelligent Systems and Applications (ISA), 2010.
[6] Yun-Yan Li, "Application of Association Rules Mining in Power Demand-Side Management", Seventh International Conference on Machine Learning and Cybernetics, Kunming, 2008.
[7] Zhang Hong, Zhang Bo, Kong Ling-Dong,Cai Zheng-Xing, "Generalized Association Rule Mining Algorithms based on Data Cube", Eighth ACIS International Conference, 2007.
[8] Dongme Sun, Shaohua Ten, Wei Zhang, Haibin Zhu, "An Algorithm to Improve the Effectiveness of Apriori", Cognitive Informatics, 6th IEEE International Conference, 2007.
[9] Riadh Ben Messaoud, Omar Boussaid, Sabine Loudcher Rabaseda, "Mining Association Rules in OLAP Cubes", Innovations in Information Technology, 2006.
[10] Hong Shi, Ji-Fu Zang,Lian Zheng, "Mining Association Rule Oriented Data Cube And Its Application", First International Conference on Machine Learning and Cybernetics, Beijing, 2002.