



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

A Study on Different Mining Operations on RFID Dataset

Jagpreet Singh¹
Student, M.Tech. (CSE)
ITM University, Gurgaon

Meghna Sharma²
Asstt. Professor, M.Tech. (CSE)
ITM University, Gurgaon

Abstract: When the raw data is collected from different users and the sources and managed in a centralized warehouse. Then major problem is to identify the most co-related attributes and information from different sources and maintain them at single source. The work also includes to discard the missing value records and to identify the irrelevant attributes from the different data sources. The presented paper is the study of different dataset filtration approaches such as association mining, apriori algorithm etc. The work is here defined for the RFID data. The work also includes the study of different mining operations on large RFID dataset to perform the cleaning and classification.

Keywords: RFID Dataset, Apriori, Data Cleaning, Attribute Selection

I. Introduction

Data mining is about to perform an intelligent analysis on large datasets to maintain the dataset effectively as well as to retrieve the most valuable information from it. It includes the search over the dataset to identify a particular pattern over the dataset. In an organization, most of the management level decisions are taken after the analytical study under data mining approaches. Data mining itself is incorporated with different methodologies to perform different operations on datasets. The operations conducted on a dataset include the data cleaning, dataset segmentation, classification of dataset etc. The operation defined under the data mining includes the statistical, learning and optimization approaches. The objective of these all approaches and operations is to maintain the dataset effectively. Most of these operations also affected by the type of dataset as well as on the size of dataset. To derive the effective results from mining operations it is required to identify the dataset features. The validity and reliability of dataset is an important factor to extract the dataset features[1,2].

To derive the effective results from the dataset accuracy of dataset is required. The reliability or the validity of dataset depends on the source of dataset generation. If the source is reliable, the dataset is reliable whether the obtained information is complete or not. Once the reliability of dataset is verified, the next work is to perform the filtration over the dataset. The filtration process is about to identify the attributes having the valid values i.e. non-zero and true values. Another work in filtration process is the identification of relevant information. The relevancy is identifying by performing the dependency analysis. The existing of a particular value kind in an attribute respective to other attribute time is the way to identify the relevancy [4,5].

Just after the data cleaning, the next work is performing the major data mining operation. The major operations are the actual operation implemented on the dataset such as the classification, clustering, feature extraction etc. Another attribute related to the mining concept is the maintenance of data sessions so that effective information from each session can be retrieved. Multiple user interaction and concurrency are the other features that affect the data mining operations. These operations give the overlapping of data values. The data resides in a dataset having the different priorities levels based on the user type, session time etc. There are number of data mining approaches that are used to conduct the different operations on large datasets. Some of these common approaches are discussed here under.

A. Association Mining

Association is about to identify the interrelation between different attributes as well as attribute values. Association mining is most useful mining approach used as an individual process as well as a stage of some data mining model to improve the accuracy in results. It basically states about the relationship between the data values. It is associated with different mining operations such as data cleaning, classification, filtration approach etc. Association mining can be done horizontally or vertically. The levels of association mining are also considered by defining the number of associated attributes [3][8].

B. Clustering

It is never effectual to process on large dataset at one time. To obtain the quick running results from the large datasets, complete dataset is divided in smaller datasets. There are number of existing clustering approaches to collect the similar dataset in one cluster. Clustering is basically performed on distance based analysis such as Euclidian distance analysis. There are number of existing clustering approaches such as Kmeans approach, Fuzzy CMeans approach, Hierarchical Clustering etc

C. Data Visualization

Visualization is about the representation of the dataset or the results in effective form so that some valuable information can be derive easily from the dataset. To perform the visualization, some graphical forms are generally selected as the picture information is more worth then textual information.

II. Literature Review

In year 2010, Chien-Chih Yu presented a Framework on customer oriented intelligent decision support system. It was a web based frame work to personalize the B2C web e-services. The work includes the personal management, auction, negotiation, evaluation, planning, collaboration, transactions, payments, feedback and quality control. In this framework, almost all the business oriented functionality is implemented and described for investment and the tourism applications. The work is about to improve the efficiency and the effectiveness of the decision support system in the same area [1]. In same year, an e-learning based decision support system was presented by Marta Zorrilla for E-Learning. The distance course learning and teaching process is processed in this application. The work includes the course description, course assignment, and other course based data mining operations are been handled by the system. The system includes the pattern based model along with probability analysis for the decision making. The work will help the instructor to answer all the student query and to find the outcome of standard process performed. The work also includes the reporting tool to present the work in an effective way [2]. In year 2012, Salma Mahgoub presented an interactive case based learning system for the business intelligence. The work is performed for the teaching system to answer the classroom question and performed a factor based analysis based on real time business application with complex compatibility. The paper describe based on interactive action for student participation for the learning process with user and instructor interactivity. The student can perform the query and the system will automatically perform the decision oriented query handling and provide the results respective to user requirement [3].

In year 2007,, Mohammed N. A. Abdelhakim presented a work on intelligent decision making for the evaluation and selection of educational multimedia. The work is a web based group decision system that will perform a statistical analysis on education provider with continuous evaluation to investigate the requirement for developing the educational application. The work also include the knowledge management along with design and implementation of performance evaluation to present the collect and process the data from instructors, experts and producers and propose a solution for the educational consumers for the evaluation of the system[4]. In year 2008, Suresh Kalahari presented a work on the data mining operation on student driven content analysis while online teaching. The work is presented in the form of a web model integrated with data mining operations to handle the classroom discussions to predict the student faring. The model also provides the feedback on student discussion regarding the topic discussion in class, and a comparative analysis with other topics. The analysis includes the text mining operations relative to the answers submitted by the RFID Data [5]. Another mining based analysis on academic data is performed by J.M.Lauria in year 2012 for the analysis of college student retention. Author has presented an analytical research on academic risk using data mining approaches. The work is presented in the form of methodological framework to develop the query based model to analyze the course management respective to academic records and the classification process is performed to work on selected dataset [6].

In year 2012 Evis Trandafili performed a worm on the information extraction and the evaluation on student profiles based on machine learning. The work includes the student enrolment along with course and performance indicators on the data set. The work includes the complex analysis on large volume of data. It includes the restructuring along with knowledge discovery to analyze the RFID Data and the quality of course delivery. The analytical analysis includes the clustering to perform a selective analysis on dataset [7]. In Year 2010, S Shirgaonkar performed a work," Overview of Real Time Decision Support System". This paper proposes the Enhanced Decision Making Model making efficient rational decision. New types of Decision Support System like Real Time Decision Support System (RTDSS) has been introduced [8]. Andrej Kristofic [9] performed a work on the adaptation of web based educational hypermedia to perform knowledge discovery. The work includes the content extraction and navigation for the predefined set of rules. Other behaviors and the preferences given respective to the behavioral analysis based on the presented rule set. The work includes the exploration of data mining operations to discover the student knowledge behavior along with the prediction prediction architecture [10]

III. Association Mining

Association rule mining is an important technique in the data mining. A major concern today in Association rule mining is to improve the algorithmic performance. Association rule mining is to discover the potential relation between the sets of data items. Many algorithms were defined under association mining they are Eclat algorithm, FP-growth algorithm, GUHA procedure ASSOC, OPUS search.

In the algorithms of association mining, Apriori is the oldest which is offered by Agrawal R in 1993. Apriori is the best one under association mining. It uses a breadth first search technique. Eclat algorithm uses set interaction and is a depth first search technique. In the FP - growth algorithm, FP stands for Frequent pattern, and uses recursive processing approach. GUHA procedure ASSOC uses fast bit operation and is a method for exploratory data analysis. OPUS is an efficient association technique but does not require monotone, anti-monotone constraint.

Other categories of association rule mining were contrast set learning, weighted class learning, high-order pattern discovery, K-optimal pattern discovery, generalized association rules, quantitative association rules, interval Data association rules, maximal association rules, sequential pattern mining, sequential Rules. Association rule mining was used to find association and interesting association between the data in the large data set.

Basically, these association rule mining algorithm were defined as class association rules (CAR) mining. This CAR comprises of three steps:

- Rule generation: rules are generated from the dataset.
- Rule ordering: classifiers are used to order the rules. Ordering can be done according to many criteria's. Some of which are used in a priori are supported, confidence, minimum support threshold, rule length, etc.
- Classification: In this step, according to the rule specified the data are classified.

The need for finding association and patterns arose when the market felt the need to learn about customer behavior of purchasing, so as to improve the sale, enhance business, attract customer, and to maximize the profit. So, in 1993, R. Agrawal et.al felt the need to design a new algorithm which was Apriori Algorithm, for finding associations among the dataset. [1]

For generating rules the constraint defined which it must satisfy were: syntactic constraint and support constraint. Pruning was based on the count of remaining tuples on the pass. The benefit of pruning was that it saves both memory and measurement effort. Pruning was referred as a remaining tuples optimization. Those rules which satisfy both minimum transactional support s and minimum confidence c were found. An improved algorithm for the generation of fast rules of mining has been discussed in 'fast algorithms for mining association rules' by R. Agrawal and R. Srikant in 1994. Execution time was improved as the number of dataset increases. [7]

Then in 1995, R. Agrawal presented a new algorithm for mining association rules in large dataset. It reduces CPU overhead and in this Partitioning Algorithm was defined. [11] In 'Mining Sequential Patterns', the whole process of mining was sorted into five phases: sort phase, Litemset phase, transformation phase, sequence phase and maximal phase. [12] The work done by R. Agrawal in 'Mining Sequential Patterns', had certain limitations, first is the absence of time constraint, then rigid definition of transaction and last is absence taxonomies. So to overcome all these a new approach GSP algorithm was formed which consists of two phases: join phase and prune phase.

Some new algorithms were designed to help reduce the size of candidate set. This included hash-based technique, partitioning, sampling and use of the vertical data format. Apriori algorithm is one of the classic algorithms designed to operate on databases containing transactions. It is one of the oldest algorithms in association rule mining defined by Aggarwal et al. In 1993. It is one of the popular approaches to derive rules and patterns from the dataset. It is a seminal algorithm that uses candidate generation for finding frequent item sets [13][14].

Apriori is easy and simple to implement approach. Before going into the details of the algorithm we would define some common terminologies associated the algorithms:

Itemset- let me- $I_1, I_2, \dots, I_n$, is a set if items i.e. me. Itemset is the collection of items in the database. If there are m items in the database, then there will be 2^m possible itemset.

Transaction- It is a database entry that contains details of all the items. The transaction is denoted by T and T is a subset of me. $T = \{I_1, I_2, \dots, I_n\}$

Database- It is a set of transactions. $D = \{T_1, T_2, \dots, T_n\}$

Support- Support measures the transaction which has item set that measure both sides of implication in association rule. Denoted by s .

SUPPORT ($A \rightarrow B$) = No. of Transaction containing both A and B.

It corresponds to statistical significance. If support is not large enough means rule is not worth consideration or is less preferred.

Minimum Support- It is minimum threshold, which should be satisfied with an item to be the frequent item in the dataset. Denoted by Mins.

Confidence- measure of rule's strength or its certainty. Denoted by c.

CONFIDENCE ($A \rightarrow B$) = No. of Transaction containing both A and B / transaction in A

Frequent Itemset- The item set which satisfies the criteria of being greater or equal to minimum support are frequent item set. It is denoted by L_i , i = itemset. If itemset does not satisfy the criteria it is non-frequent itemset.

Candidate Itemset- items which are to be used for processing are candidate itemset. Denoted by C_i , i =itemset. Apriori generates candidate itemset of length $n+1$ from itemset of length n .

Association rule- the rule basically used to find an association between two items or which is used to associate/ relate two items is association rule. Let $A \subseteq I$ and $B \subseteq I$, and $A \cap B = \emptyset$, the rule can be $A \rightarrow B$.

Apriori approach finds frequent itemset from a transaction database and then derives association rule from them. For the generation of frequent itemset, it uses candidate generation step. When finding frequent itemset, the itemset with a frequency greater or equal to the minimum support threshold is taken.

Whereas to generation association rules confidence greater or equal to minimum confidence is taken.

Candidate generation in Apriori algorithm follows two steps: Join step and Prune step.

In Join step, the union of two frequent itemset of size n , suppose A_n and B_n , which have first $n-1$ element in common is taken.

$$J_{n+1} = A_n \cup B_n$$

In Prune step, all the itemset of size n in J_{n+1} are checked whether they are frequent or not, those which pass the requirement generate C_{n+1} .

Join step as follows Prune step in apriori algorithm. Prune step is further classified as into Subset prune phase and Interest prune phase. Pruning techniques are categorized as Remaining tuple optimization and Pruning function optimization.

Association rules are found which satisfies the criteria of minimum transaction support and minimum confidence.

IV. Conclusion

In this paper, the data mining operations are been explained respective to RFID dataset. The main stress is here given to the data cleaning operations and the apriori algorithm. The work includes the different aspects of association mining and the apriori algorithm.

V. References

- [1]. R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules Between Sets of Items in Large Databases", In *Proceedings of the ACM SIGMOD Conference on Management of data*: 207-216, May 1993
- [2]. Joseph McKendrick, "Big Data, Big Challenges, Big Opportunities: 2012 IOUG Big Data Strategies Survey", IOUG, Sept 2012
- [3]. R. Srikant, R. Agrawal, "Mining Quantitative Association Rules in Large Relational Tables", The 1996 ACM SIGMOD International Conference on Management of Data, Monreal, Canada, pp-1-12, June, 1996
- [4]. "Big Data for Development: Challenges and Opportunities", Global Pulse, May 2012
- [5]. Nigel Wallis, "Big Data in Canada: Challenging Complacency for Competitive Advantage", IDC, Dec 2012
- [6]. Ivanka Valova, Monique Noirhomme, "Processing Of Large Data Sets: Evolution, Opportunities And Challenges", Proceedings of PCaPAC08
- [7]. Neha Saxena, Niket Bhargava, Urmila Mahor, Nitin Dixit, "An Efficient Technique on Cluster Based Master Slave Architecture Design", Fourth International Conference on Computational Intelligence and Communication Networks, 2012
- [8]. Edmon Begoli, James Horey, "Design Principles for Effective Knowledge Discovery from Big Data", Joint Working Conference on Software Architecture & 6th European Conference on Software Architecture, 2012
- [9]. Kapil Bakshi, "Considerations for Big Data: Architecture and Approach", IEEE, 2012
- [10]. Lawrence O. Hall, Nitesh Chawla, Kevin W. Bowyer, "Decision Tree Learning on Very Large Data Sets", IEEE, Oct 1998
- [11]. Trevor Hastie, Robert Tibshirani, Jerome Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction", Springer, 2nd edition, 2008
- [12]. Mr. D. V. Patil, Prof. Dr. R. S. Bichkar, "A Hybrid Evolutionary Approach To Construct Optimal Decision Trees with Large Data Sets", IEEE, 2006
- [13]. Guillermo Sinchez-Diaz, Jose Ruiz-Shulcloper, "A Clustering Method for Very Large Mixed Data Sets", IEEE, 2001
- [14]. Mehmet Koyuturk, Ananth Grama, and Naren Ramakrishnan, "Compression, Clustering, and Pattern Discovery in very High-Dimensional Discrete-Attribute Data Sets", IEEE Transactions On Knowledge And Data Engineering, Vol. 17, No. 4, April 2005