International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 6 | ISSN : 2456-3307

A Study on Data Mining: Frequent Itemset Mining Methods Apriori, FP growth, Eclat

Dhinakaran D[#1], Dr. Joe Prathap P M[#2]
[#1]Assistant Professor, Department of Computer Science & Engineering, Peri Institute of Technology, Mannivakkam, Kanchipuram, India
[#2]Associate Professor, Department of Computer Science & Engineering, RMD Engineering College, Kavaraipettai, Tiruvallur, India

ABSTRACT

Data mining is described as a process of discovering useful and interesting patterns hidden in huge amounts of data stored in multiple data sources. Data mining is a interdisciplinary field, ranging from Statistics, Database technology, Information recovery, Artificial intelligence, Machine learning, Pattern recognition, Neural networks, Knowledge-based systems, High-performance computing, and Data visualization have had impacts on the growth of data mining. Association rule mining is the core process in the field of data mining. It discover set of frequent items & generates ruleset within huge transaction databases. Data mining and its techniques can be enormously helpful in many fields such as business, education, government, fraud detection, and financial banking, future healthcare and so on. Data mining have a lot of merits but still data mining systems face lot of troubles and hazards. The purpose of this paper is to discuss the basic concepts of data mining, its various techniques , specifically about Frequent Itemset Mining Methods, various challenges, applications and important issues related to data mining.

Keywords : Data Mining, Association Rule Mining, Frequent Itemset Mining, Transaction databases.

I. INTRODUCTION

Data mining - knowledge discovery from data, is a process of extracting interesting patterns or knowledge (non-trivial, implicit, previously unknown and potentially useful)from huge amount of data. Alternative names of data mining are Knowledge discovery in databases (KDD), knowledge extraction, pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc. Data mining is an interdisciplinary field bringing together techniques from Database technology, Statistics, Information retrieval, Artificial intelligence & Machine learning, Pattern recognition, Neural networks, Knowledge-based systems, High-performance computing and Data visualization to address the issue of information[9].

Data mining often involves the analysis of data stored in a data warehouse[20]. Three of the major data mining techniques are association mining which define associationrules for finding frequent pattern among variables. The next is classification which to identify the unknown class label. The next technique is clustering which is to identify the meaningful or useful cluster of data which have similar characteristics. In today's scenario the most active research area covers a association of mining information in transactional dataset leading to the evolution of Association rule mining (ARM). Association rule mining technique have been engaged in applications such as market basket analysis, health care, web usage mining, bioinformatics and prediction. Data Mining is mostly used in various areas. There are number of commercial data mining scheme available nowadays yet there is many competitive situation in this field.

This paper presents association rule mining as a technique for discovering patterns within large amount of transaction data. Section 2 consists of various data mining process, Various algorithms and techniques in section 3, Followed by section 4, in which detailed working of association rule mining and Various algorithms like Apirori, FP-Tree and ECLAT are summarized. In Section 5 we list the challenging issues and at last in Section 6 we list the applications of data mining.

II. DATA MINING PROCESS

Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the essential extraction of hidden, previously unknown and potentially useful information from data in databases. Data mining is actually part of the knowledge discovery process. The following Fig.1 shows Data mining as a step in the process of knowledge discovery. The Knowledge Discovery in Databases process involves few steps starting from raw data collections to useful knowledge modeling [9].

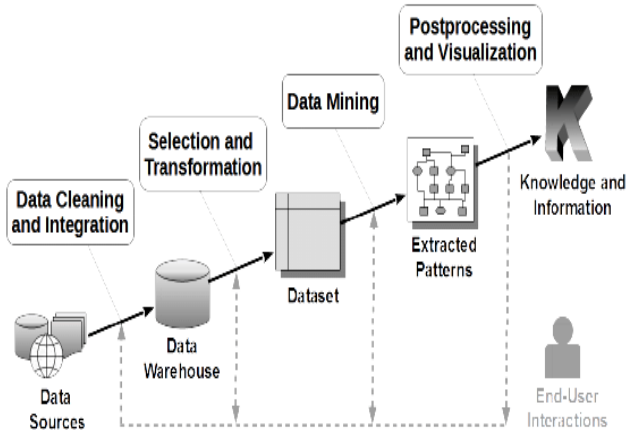

Figure1. Knowledge Discovery in Databases process

Data preprocessing it goes through various process:

Data cleaning: Applied to remove noise and irrelevant data from the database.

Data integration: Combine data from several data sources.

Data selection: Appropriate data to the analysis is determined and retrieved from the data source.

Data transformation: Transformation of data in to suitable form for the mining process.

Data mining: It is the most critical process in which various techniques are used to obtain patterns which are potentially useful.

Pattern evaluation: It identifies the accurate interesting patterns based on given measures.

Knowledge representation: It is the final phase which help users to understand and interact with the data mining results with the help of visualization and knowledge representation techniques.

Some of the pre-processing steps combines together. where data cleaning process and the data integration process can be done jointly as a pre-processing stage to build a data warehouse[25]. where data selection process and data transformation process can also be combined where the selection is done on transformed data from the data warehouse.

III. DATA MINING TECHNIQUES

Various techniques like Mining Frequent Patterns, Artificial Intelligence, Associations, Correlations, Regression, Classification, prediction, Clustering etc., are used for knowledge discovery from various data sources [9] .

Data mining techniques are discussed briefly below as:

A. Association

Association mining is one of the familiar data mining technique to scrutinize and to predict customer behavior. Association generate dependency rules which will predict appearance of an item based on appearances of other items using two measures called support and confidence. Association technique is used in applications such as Banking, Catalog design, market basket analysis, health care, web usage mining, bioinformatics and prediction.

Types of association rule

- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule

B. Correlations

Correlation is used to identify the uninteresting Association rules from transactional dataset. The measures like support and confidence are inadequate at filtering out uninteresting Association rules. LIFT is a simple correlation measure to identify positively, negatively and independent correlation from two itemsets.

C. Classification

Classification is the task to analysis data, where a classifier is build to predict categorical labels. Classification classifies data based up on training data set and uses to classify the new data. The data classification method includes learning and classification. In Learning phase the training data are evaluated by classification algorithm. In classification phase the test data be used to forecast the correctness of the classification rules.

If the accuracy is good enough the rules can be applied to the new data. Credit card approval, Target marketing, medical diagnosis, Fraud detections are some of the typical applications of classification.

Types of classification models:

- Decision tree induction
- Tree Pruning
- Bayesian Classification
- Rule-based classification
- Classification by Back propagation
- Support Vector Machines (SVM)
- K-Nearest Neighbor

D. Prediction

Prediction is the task to analysis data, where the model build predicts a continuous-valued function, or ordered value, as disparate to a categorical label. Prediction predicts unknown or missing values. Most widely used method for numeric prediction is Regression. Regression models a relationship between one or more independent or predictor variables and a dependent or response variable. Regression Analysis is a excellent selection when all the predictor variables are continuous-valued.

Types of regression methods

- Linear Regression
- Multiple Linear Regression
- Nonlinear Regression
- Multivariate Nonlinear Regression

E. Clustering

The method of grouping the objects based on the information discovered in the data depicting the objects or their associations. The main task of Clustering is to group the objects of similar type and the objects in one group will be different from the other group. Clustering has the ability to handle with noisy data, with diverse types of attributes and clusters with random shape.

Types of Clustering methods

- Partitioning Methods
- Hierarchical methods
- Density based methods
- Grid-based methods
- Model-based methods

IV. ASSOCIATION RULES

Association rule mining is one of the major technique of data mining for knowledge discovery within large transaction databases. An association rule of the form A => B , where A and B are sets of items in a transaction database, so that $A \cap B = \phi$. The rule indicates there is a high probability that whenever all items from set A appear in a transaction, the items of set B will also appear. In general we refer to the left side as the antecedent and to the right side as the consequent. Support and Confidence are the two measures which defines the rules [1]. where support represents the probability that contains both A and B, where else confidence denotes the probability that a transaction containing A also contains in B. The conditional probability, for support and confidence is,

support(A => B) =P(A∪B)

confidence (A => B) =P(B/A)

Rules that fulfill both a minimum support threshold (min sup) and a minimum confidence threshold (min conf) are called strong rules[18]. Table 1 shows the Market-Basket Transactions.

TID	ITEMS
1	Milk, Tea, Sugar
2	Curd, Sugar, Honey, Tea
3	Butter, Milk, Coffee
4	Milk, Honey
5	Milk, Tea, Sugar, Coffee
6	Milk, Sugar, Coffee

Table 1 - Market-Basket Transactions

Example of Association Rules:

{Milk} ==>{Sugar},
{Tea, Sugar} ==>{Milk}

Rule Evaluation:

- **Support (S):** Transactions that contain both A and B.
  Example: {Tea, Sugar} ==>{Milk}
  S=σ({Tea, Sugar, Milk}) / T
  S= 2/6 = 33%
- **Confidence (C):** Measures how often items in A also contains in B .
  Example: {Tea, Sugar} ==>{Milk}
  C = σ ({Tea, Sugar, Milk }) / σ ({Tea, Sugar})
  C = 2/3 = 67%
- **Frequent Itemset:** Generate all itemsets whose Support ≥ min_support threshold.
- **Association rule:** Generate strong association rules from each frequent itemsets, where having the Confidence ≥ min_confidence threshold.

Consider if the minimum support is 30% and minimum confidence is 50%, then {Tea, Sugar} ==>{Milk} is an example for strong association rule.

A. Apriori Algorithm

Apriori algorithm is the most frequently used algorithm for mining frequent itemsets, proposed by Rakesh

Agrawal and Ramakrishnan Srikant in 1994[1]. Support and Confidence are the two measure defines the association rule. Apriori algorithm returns an association rule if its support and confidence values are above the threshold values.

Apriori algorithm is to make multiple passes over the database. Apriori utilize an iterative approach known as a breadth first search, where k -item sets are used to explore (k+ 1) itemsets. First, the set of frequent 1 - itemsets are found by scanning the database to accumulate the count for each item, and collecting those items that satisfies minimum support. The resultant set is specified as L_1. Then, frequent 1-itemset L_1 is used to find the set of frequent 2-itemsets L_2, which is used to find L_3, and so on, until no more frequent k-item sets can be found. The finding of each Lt requires one complete scan of the database[24].

Pseudo-code :

C_k: Candidate itemset of size k
L_k: frequent itemset of size k

L_1 = {frequent items};
for (k = 1; L_k !=ϕ; k++) **do begin**
C_{k+1} = candidates generated from L_k;
for each transaction t in database do
increment the count of all candidates in C_{k+1} that are contained in t
L_{k+1} = candidates in C_{k+1} with min_supp
end
return $U_k L_k$;

Example: Consider if the minimum support is 2, then the above Fig. 2 represents the frequent itemsets of various levels. First, the Candidate itemset C_1 are recognized by scanning the database D, then the set of frequent 1 - itemsets are identified by accumulating the count of each item, and collecting those items that persuade minimum support. The resultant set is specified as L_1. Then, frequent 1-itemset L_1 is used to generate the Candidate itemset C_2 . Then the set of frequent 2 - itemsets are found by scanning the database D to accumulating the count of each itemset, and collecting those itemset that persuade minimum support. The resultant set is specified as L_2. Then, frequent 2-itemset L_2 is used to find L_3.

Two phases:

1) Generate phase:

Candidate (k+1) itemset is generated using K-itemset, this phase creates C_k candidate set.

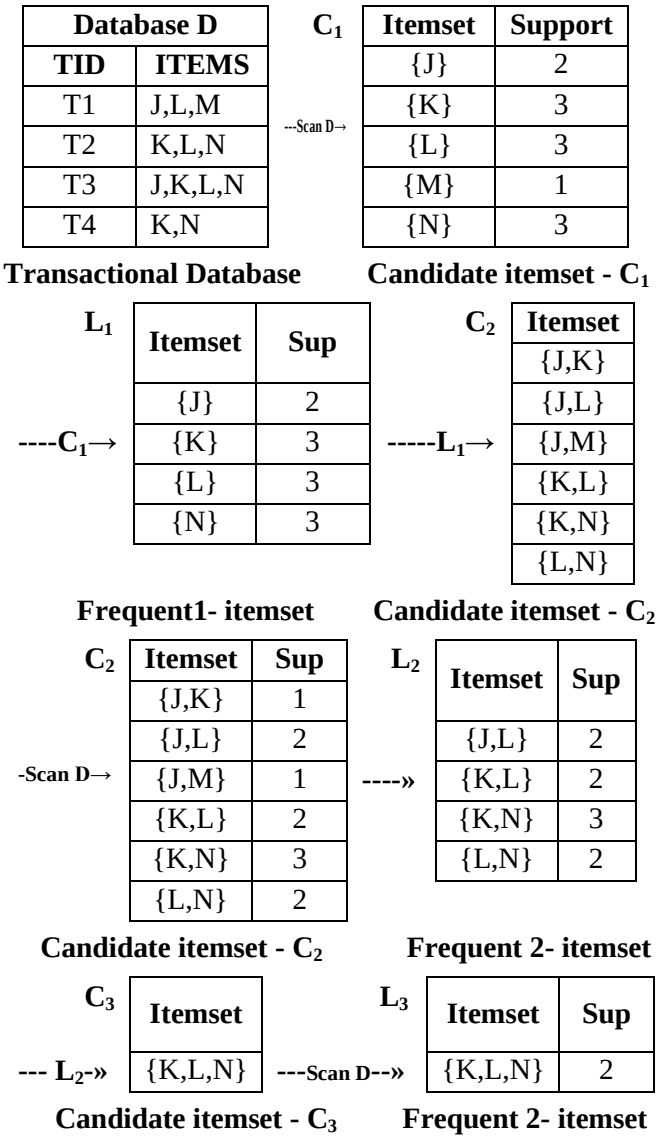

Figure 2. Process of Apriori Algorithm

2) Prune phase:

Candidate set is reduced to generate large frequent itemset using min_supp as the pruning parameter. This phase produces L_k large itemsets.

Advantages:

a) Easy implementation.
b) Due to the property of pruning, itemsets left for further support checking remain less.

Disadvantages:

a) At each one level of processing Apriori algorithm generates an enormous number of candidate itemset.
b) It scans the complete database multiple times.
c) It has a complex candidate generation process that uses most of the time, space and memory.

d) It is costly to determine the support of the candidate sets for each transaction in the database.

B.FP-Tree

FP Stands for frequent pattern, is used in the progress of association rule mining. A FP tree is a kind of prefix tree which allows to discover frequent item set omitting the candidate item set generation [24].

FP-Tree algorithm overcome the problem found in Apriori algorithm. Item set Mining is possible without candidate generation [5] and takes only two scan over the database, FP-Tree was found to be faster than the Apriori algorithm.

FP-Tree is a twostep process:

Step 1: Construct a dense data structure named as FP-tree via 2 passes.
Step 2: Extracts frequent item sets directly from the FP-tree.

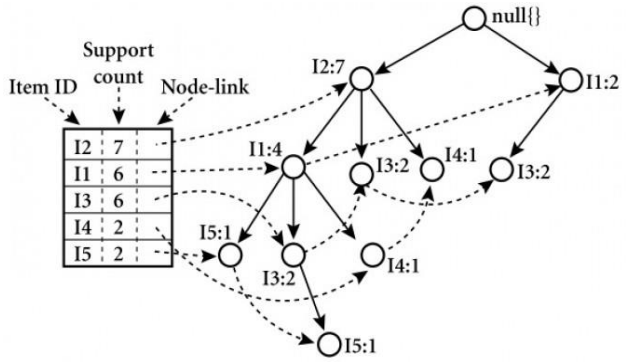

Figure 3. An FP-tree registers compressed, frequent pattern information

Fig. 3 represents the frequent pattern information. FP-Tree uses the tree structure to map the count of each itemset without generating the candidate itemset.

FP-Tree is constructed using 2 passes:

Pass 1:

- Scan the data to find support for all item.
- Eliminate infrequent items.
- Based on the support sort the frequent items in declining order.
- When constructing the FP-Tree use this declining order to share common prefixes.

Pass 2:

- Nodes represents to items and nodes have a counter.
- FP-Growth examine one transaction at once and plots it to a path.
- Determined sequence is used, so paths can lap over when they have same prefix .
  - In this condition, counters are incremented.
- Pointers are conserved between nodes holding the same item, constructing linked lists.
- The various paths that lap over, the higher the density. FP-tree could fit in memory.
- Frequent itemsets are dig out from the FP-Tree.

Advantages:

a) It scales superior in contrast to Apriori algorithm.
b) It entail only two scans of database without any candidate generation which is better than Eclat and Apriori.
c) Used in cases of large problems as it doesn't require generation of candidate sets.

Disadvantages:

a) The resultant FP-Tree is not distinctive for the similar logical database.
b) Due to complex data structure, increases in execution time .
c) It cannot be used in interactive and incremental mining system.
d) Using tree structure creates complexity.
e) FP tree may not fit in main memory.

C.ECLAT

ECLAT is a algorithm used for mining frequent itemsets, ECLAT was proposed by Zaki, Parthasarathy. Both Apriori algorithm and FP-Tree algorithm uses horizontal data set-up, but ECLAT uses vertical data set-up. Where ECLAT algorithm transforms the horizontal data format into the vertical data format for the given transactional data set of TID-itemset[23].

ELCAT process is similar to Apriori algorithm which works in a iterative pattern. It uses tree structure named as the Tidset. The algorithm Searches in a Depth first search manner. ECLAT produces a large set of rule set equivalent to that of Apriori algorithm, but it does not generates candidate set.

TID	Items
1	Milk, Tea, Sugar
2	Sugar, Honey,
3	Milk, Coffee
4	Milk, Honey
5	Milk, Tea, Sugar,
6	Milk, Sugar,

Item set	TID set
Milk	1,3,4,5,6
Tea	1,2,5
Sugar	1,2,5,6
Honey	2,4
Coffee	3,5,6

Table 2. Horizontal Data Format Table 3. Vertical Data Format

The Table 2 represents the Horizontal Data Format and Table 3 represents the Vertical Data Format of the Market-Basket Transactions.

Frequent 1-itemsets	
Item	**TID set**
Milk	1,3,4,5,6
Tea	1,2,5
Sugar	1,2,5,6
Honey	2,4
Coffee	3,5,6

Frequent 2-itemsets	
Item set	**TID set**
Milk, Tea	1,5
Milk, sugar	1,5,6
Milk,Honey	4
Milk,Coffee	3,5,6
Tea, Sugar	1,2,5
Tea, Honey	2
Tea, Coffee	5
Sugar, Honey	2
Sugar, Coffee	5,6
Honey, Coffee	-

Table 4. Intersection of 1-Itemset Table 5. Intersection of 2-Itemset

Frequent 3-itemsets	
Item set	**TID set**
Milk, Tea, Sugar	1,5
Milk, Sugar, Coffee	5,6

Table 6. Intersection of 3-Itemset

By Consider the min_Supp=2, Table 4, 5 and 6 represents the Frequent 1-itemset, Frequent 2-itemset and Frequent 3-itemset.

ECLAT process:

1) Get Tidlist for each item.
2) Tidlist of {x} is exactly the list of transactions containing {x}.
3) Intersect Tidlist of {x} with the Tidlists of all other items, resulting in Tidlists of {x, y},{x, z},....
4) Repeat from 1 on {x}-conditional database.
5) Repeat for all other items

Advantages:

a) Less memory usage.
b) Scanning the database is not needed to get the support of (k+1) itemsets, for k>=1.
c) ECLAT is quicker compared to Apriori and FP-Tree.

Disadvantages:

a) ECLAT needs additional time for intersection when Tidlist is abundant.
b) Takes more space and time to store candidate set when Tidlist is large.

V. CHALLENGES IN DATA MINING

Data mining have a lot of merits but still data mining systems face lot of troubles and hazards.

- Privacy Preservation
- Scalability
- Distributed Data and Operations
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality
- Network Setting

VI. APPLICATION OF DATA MINING

Data mining and its techniques can be enormously helpful in many fields such as

- Banking
- Future Healthcare
- Market Basket Analysis
- Research Analysis
- Marketing
- Government and Defense
- Manufacturing and Production

VII. CONCLUSIONS

In this paper we attempt to give an insight to various process, techniques, issues and applications related to Data mining. Under the Association rule mining technique of data mining various algorithms namely Apriori, FP-Tree and ECLAT algorithms are studied. This review would be useful to researchers to focus on the Association rule mining techniques of data mining. However association rule mining is still in a position of penetrating and improvement. There are still various critical issues that need to be studied for identifying useful association rules.

VIII. REFERENCES

[1]. Rakesh Agrawal, Imielinski, T. and Swami, A. "Mining association rules between sets of items in large databases", International Conference on Management of Data, 207-216, 1993.

[2]. Savasre A., Omienciski E., and Navathe S., " An efficient algorithm for mining association rules in large databases, International conference on VLDB, pp. 432-444, 1995.

[3]. Rakesh Agrawal and Ramakrishnan Srikant, "Fast algorithms for mining association rules", Proc. of 20th Int. Conf. on Very Large Data Bases, Vol. 1215, pp. 487-499, 1994.

[4]. Christian Hidber. "Online Association rule mining", SIGMOD '99 Philadelphia PA. ACM 1-58113-084-8/99/05, 1999.

[5]. Jiawei Han, Jian Pei and Yiwen Yin, "Mining frequent patterns without candidate generation", Association for Computing Machinery Special Interest Group on Management of Data, Vol. 29, Issue 2, pp. 1-12, 2000.

[6]. Andrew Kusiak, Jeffrey A. Kern, Kemp H. Kernstine, and Bill T. L. Tseng, "Autonomous Decision-Making: A Data Mining Approach", IEEE Transactions On Information Technology In Biomedicine, VOL. 4, NO. 4, 2000.

[7]. Luigi Lancieri, Member, IEEE, and Nicolas Durand "Internet User Behavior: Compared Study of the Access Traces and Application to the Discovery of Communities" IEEE Transactions On Systems, Man, And Cybernetics, VOL. 36, NO. 1, 2006.

[8]. C.S. Selvai, A.Tamilarasi, "Association Rule Mining With Dynamic Adaptive Support Thresholds for Associative Classification", International Conference on Computational Intelligence & multimedia Application (ICCIMA'07),Vol. 2, pp. 76-80, 2007.

[9]. J. Han, M.Kamber, "Data Mining Concepts and Technique", Second edition, Morgan Kaufmann Publishers, pp. 1-40, 2008.

[10]. Fayyad, Usama, G.P.Shapiro, P Smyth, "From Data mining to Knowledge Discovery in Databases", Fayyad, pp. 12-17, 2008.

[11]. PM Joe Prathap, V Vasudevan," Pay per view-a multimedia multicast application with effective key management", International Journal of Mobile Network Design and Innovation, Volume 3, No. 2, pp. 82-92, 2009.

[12]. Yanthy W., Sekiya T., Yamaguchi K., "Mining Interesting Rules by Association and Classification Algorithms", International Conference on Frontier of Computer Science and Technology, pp. 177-182, 2009.

[13]. K.Srinivas B.Kavihta Rani Dr. A.Govardhan "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks" Srinivas et al. / (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 02, 250-255, 2010.

[14]. Yoshimasa Tsuruoka, Makoto Miwa, Kaisei Hamamoto, Jun'ichi Tsujii and Sophia Ananiadou, "Discovering and visualizing indirect associations between biomedical concepts", Bioinformatics, Vol. 27, Issue. 13, pp. i111-i119, 2011.

[15]. Jiban K Pal "Usefulness and applications of data mining in extracting information from different perspectives" Annals of Library and Information Studies Vol. 58, pp. 7-16, March 2011.

[16]. Rahul Isola, Rebeck Carvalho, and Amiya Kumar Tripathy, "Knowledge Discovery in Medical Systems Using Differential Diagnosis, LAMSTAR, and k-NN ", IEEE Transactions On Information Technology In Biomedicine, VOL. 16, NO. 6, NOVEMBER 2012.

[17]. V. Vaithiyanathan, K. Rajeswari, R. Phalnikar, and S. Tonge, "Improved apriori algorithm based on selection criterion," in Computational Intelligence Computing Research (ICCIC), IEEE International Conference on, pp. 1-4, Dec 2012.

[18]. Suriya, S., Shantharajah, S.P. and Deepalakshmi, R. , "A Complete survey on association rule mining with relevance to different domain", International Journal of Advanced Scientific and Technical Research, Issue 2, Vol. 1, Feb 2012.

[19]. Sasikala, D. and Premalatha, K, "Mining association rules from XML document using modified index table", International Conference on Computer Communication and Informatics, Coimbatore, 4-6 Jan 2013.

[20]. Kale Sarika Prakash, PM Joe Prathap, "Bitmap Indexing a Suitable Approach for Data Warehouse Design", International Journal on Recent and Innovation Trends in Computing and Communication ISSN, Volume 3, No. 2, pp. 2321-8169, 2015.

[21]. X. Li, V. Ceikute, C. S. Jensen, and K. Tan, "Effective online group discovery in trajectory databases," IEEE Trans. Knowl. Data Eng., vol. 25, no. 12, pp. 2752-2766, 2013.

[22]. Varun G Menon, PM Joe Prathap, " A Review on Efficient Opportunistic Forwarding Techniques used to Handle Communication

Voids in Underwater Wireless Sensor Networks", Advances in Wireless and Mobile Communications, vol. 10, No. 05, pp. 1059-1066, 2016.

[23]. J. Heaton, "Comparing dataset characteristics that favor the apriori, eclat or FP-growth frequent itemset mining algorithms," in SoutheastCon2016. IEEE, mar 2016.

[24]. W. Altaf, M. Shahbaz, and A. Guergachi, "Applications of association rule mining in health informatics: a survey," Artificial Intelligence Review, vol. 47, no. 3, pp. 313-340, may 2016.

[25]. Kale Sarika Prakash, PM Joe Prathap, " Efficient execution of data warehouse query using look ahead matching algorithm ", International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT) , pp. 384-388, 2016.

[26]. Chirag A. Mewada, Rustom D. Morena, "Model using Improved Apriori Algorithm to generate Association Rules for Future Contracts of Multi Commodity Exchange (MCX)", International Journal of Advanced Research in Computer Science, Volume 8, No. 3, ISSN No. 0976-5697,March - April 2017.