Finding an Efficient Approach for Generating Frequent Patterns in Large Database

Ms. Deepika Fole[1], Asst Prof. Chaitali Choudhary[2]

Abstract— **Data mining is the process of finding interesting pattern from different data sets, which is used in market basket analysis, cross marketing, fraud detection. Data collection and storage technology has made it possible for organizations to accumulate huge amounts of data at lower cost. Exploiting this stored data, in order to extract useful and actionable information, is the overall goal of the generic activity termed as data mining. Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. Frequent pattern mining is extensively used in market basket analysis. Each record consists of the transaction of individual customer towards different products at different level. There is need for a monitoring and effective suggestion mechanism for merchant to develop a strategy for attracts customers to their shop. The proposed methodology uses the FP (Frequent pattern) growth algorithm for customers' behavior which describe a hierarchical scheme to provide cross-level suggestions for merchant. FP Growth uses divide and conquer mechanism. It requires less time and no candidate generation as compare to traditional Apriori algorithm. This system can help a merchant to get multiple levels of abstract suggestions in place of single level frequent pattern mining results**

Index Terms— **Data mining; Cross level frequent pattern; market basket analysis; FP growth, Association Rules.**

I. INTRODUCTION

Data mining has appeared as a branch of study directed at developing tools and techniques for the extraction of large amount of data, in order to obtain current, valuable, significant and absolute existing information [1]. From many years frequent pattern mining has been an important area in the field of data mining. A remarkable development in this field has been made and various efficient algorithms have been designed to find frequent patterns in a transactional database. Market basket analysis was first pattern mining concept proposed [2]. It is about finding association among items bought in a market. This concept used transactional databases and other data repositories in order to find association's casual structures, interesting correlations or frequent patterns among different set [3]. Items, sequences or substructures that present in database transactions with a user identified frequency is called frequent pattern. In an itemset having frequency greater than or equal to minimum threshold will be considered as a frequent pattern [4]. There are various real world applications in which frequent pattern mining can be used. For example in market basket analysis for selling different product, for promotion rules, text searching, wireless sensor networks, other applications that require observation of user environment carefully that are subject to critical conditions or hazards such as gas leak, fire explosion.

In this paper a frequent pattern tree (FP-tree) structure method is proposed. In it an extended prefix-tree structure is used which stores compressed and crucial information for frequent patterns and for mining purpose we can employ FP-growth for complete set of frequent patterns which is used by pattern fragment growth. The FP-growth method provides scalable outputs for mining long and short frequent patterns and gets more efficient results than the Apriori algorithm, and there are various new frequent pattern mining methods which are slower than FP-growth method. This paper shows that cross-level frequent pattern mining can give many suggestions to merchant which helps them to develop a strategy. Furthermore, it combines the methods of multilevel technique and frequent pattern mining which offers a new opportunity in the field of data mining.

There are three main purpose of this research:

A. It will integrate the multilevel hierarchy information encoded technique and frequent pattern mining method to give suggestions for merchant;

B. Conducting some experiments to calculate the proposed methodology using real-world data and realize what customers want;

C. Create a group of some representative cases to understand different requirements of the customers.

The remainder of this paper is organized as follows: Section 2 describes the problem statement of the research field. Section 3 describes milestones of association rule mining algorithm, Section 4 include comparative analysis of those techniques which is describe in section 3, Section 5 describes proposed methodology and finally we conclude this paper in section 6.

II. PROBLEM STATEMENT

There are a large number of research papers are present, which analyses association rule mining over transactional databases and contains single and multidimensional data sets. The existing proposals do not extract the hidden knowledge on cross level data. The proposed work of this paper addresses the issue of cross level data sets very effectively by generating cross level frequent pattern for high dimensional association rules.

A. *Single level frequent pattern*

Manuscript received
*Ms. **Deepika Fole**, Computer Science & Engineering, RCET Bhilai Bhilai, India.*
***Asst Prof Chaitali Choudhary**, Computer Science & Engineering,RCET Bhilai, India.*

ISSN: 2278 – 1323

Let I = {I1, I2, I3….Im} be a set of items and Ds be a transaction database Ds= {T1, T2, T3….Tn}.

Definition: The association rule which is written as A=>B and it is true in Ds, having support s and confidence c. Support s is defined as how many percentage of transactions are present in Ds, that should contain both A and B (AUB), in transaction Ds. Confidence c is the percentage of transactions in Ds, containing A that also contains B [5].

Support (A=>B) = P (AUB)

Confidence (A=> B) = P (B|A) =P (AU B)| P (A)

The problem is to discover all associations' rules that fulfill minimum support and minimum confidence constraints defined by user.

B. *Mining High Dimensional Association rules*

A high dimensional association rule can be of the form

Year (2003)? Location (Mumbai)? Buy (Beer)? Buy (Diaper)?
Cars (2) {sup=30%, conf=80%}

This rule means that in the year 2003 customers in *Mumbai* who buy *beer* and buy *diaper* together and having 2 cars with support 30% of total transactions and those customers in *Delhi* who buy *beer* have a confidence or probability of buying *diaper* together and having 2 cars of 80% are generated as association rules.

Mining these kinds of rules from different data sources are time consuming. It requires a common data representation known as high dimensional data set [6].

III. MILESTONES OF ASSOCIATION RULE MINING ALGORITHMS

A. AIS Algorithm

The AIS (Agrawal, Imielinski and Swami) algorithm was the first algorithm which proposes a method for the generation of association rule by Agrawal et al. in 1993 [2]. It deals with the quality of datasets together by using the necessary functionality to process decision support queries.

The main drawback of the AIS algorithm is that it generates too many candidate itemsets which requires more space and time. Also, this algorithm needs too many scans for the stored database for the generation of large itemsets [7].

B. Apriori Algorithm

There is one most important development in the field of data mining after the AIS algorithm and it was renamed as Apriori [8]. Apriori is an important development in the history of association rule mining. This algorithm was first suggested by Agrawal and Srikant in 1994. Apriori algorithm overcomes the limitations of AIS algorithm which generate too many candidate itemsets and most of them are infrequent. In Apriori two main processes is done: First is the generation of candidate itemsets, in which scanning of transactional database is done and support count of each itemset is calculated. Second is the generation of frequent Itemsets, this is a pruning phase in which prunes those itemsets which has a support count less than minimum threshold. This process is repeated until frequent itemset or candidate itemset becomes empty as in the example shown in Figure 1. The transaction database is scanned first time for candidate itemset which consist one item set and support count is calculated. After these 1-candidate itemset are pruned by eliminating those itemsets that has an item count less than the user-defined threshold (in example threshold=30%). In second phase database is scanned again for a generation of 2-candidate Itemsets which consist of two items, then again pruning is done according to Apriori property [9]. According to a priori property every sub 1-Itemset of 2 frequent Itemsets must be frequent. In example, the process is repeated for fourth scan of database and after that it will be empty.
There are two drawbacks of this algorithm First is generation of the complex candidate itemset which requires large memory and tremendous execution time and the second problem it requires lots of database scans for candidate generation[10].

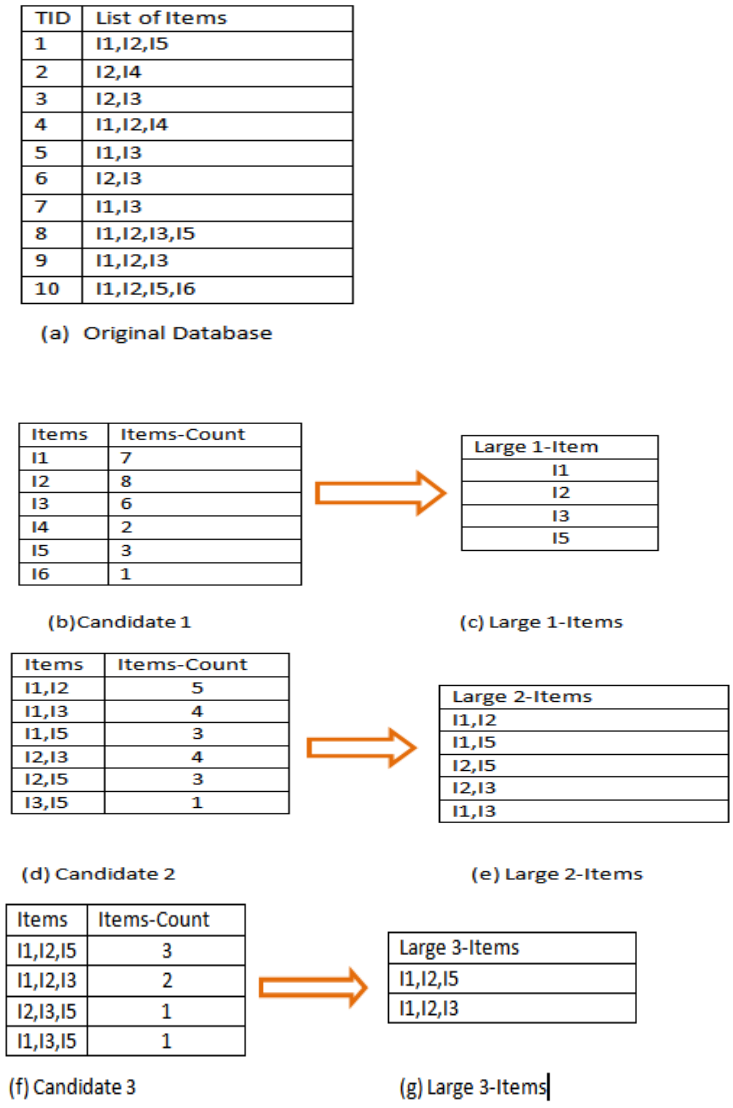

Fig 1: Apriori Process for Mining Patterns

ISSN: 2278 – 1323

C. Apriori Dynamic programming

Apriori Dynamic Programming approach is used to find frequent or large candidate 1-itemset and candidate 2-itemset. In this only one database scan for both frequent candidate 1-itemset and 2-itemset is required. In contrast Apriori requires separate scan for each frequent itemsets. This method contain two main part, first one is OccurrenceCount() Which calculate the occurrence count of 2-itemset and store it. Second one is Frequent_Count() which checks weather the count is frequent or not[14]. And it gives frequent 2-itemsets. Fig 3 shows Structure of Count_table, which represent two dimensional array and is rotated at 45 degrees to the right side. Count_table shows the variation in i and j from 1 to the total number of items that present in database [11].

T_{ID}	Item set
T_1	I_1, I_2, I_5
T_2	I_2, I_3
T_3	I_1, I_5
T_4	I_2, I_4, I_5
T_5	I_1, I_2, I_4
T_6	I_4, I_5

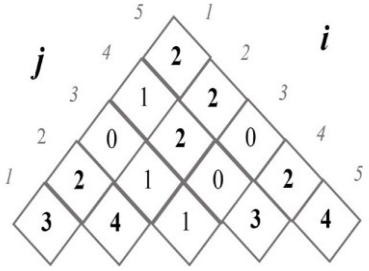

Fig: 2 Transactional database Fig 3: Count_Table layout

Figure 2 shows the transactional database. Figure 3 shows that the apex of the triangle, i.e. Count_table [1] [5] which hold the values "2", which represents the occurrence of itemsets {I1, I5} in all transactions. In the table, last row represents, the occurrence count of candidate 1-itemset in transactions, where both Index (i=j) in same table. Every item set consists of two values, for example {I1, I4}, in this first value indicate ith index and second value indicate jth index in Count_table. The Apriori Dynamic programming approach overcomes the problem of computational overhead which is occurring in the traditional Apriori algorithm. This method works in a bottom up manner. This method gives frequent 2-itemset without generation of candidate itemsets. It reduces one database scan. Count_table is optimized and efficient so that we can traverse through whole table and it is reusable. Without the scanning of database again, one can re-generate frequent candidate itemset at support count. If we can use effective data structure, then overall performance will be improved [11].

D. Mining frequent pattern using FP-Growth

The FP - growth algorithm is most popular algorithm in the field of data mining, which is used for pattern discovery. It overcomes two drawbacks of Apriori algorithm. FP-tree uses a compact data structure named as the FP - tree is constructed. FP-tree is a prefix tree Structure which stores countable information about frequent patterns. Basically FP algorithm is a two-step approach. At first step database is scanned two times and construction of FP-tree is done. In first scanning of database, data sets are scanned and calculation of support count is done, deletion of infrequent pattern is done from the list and existing pattern is sorted in descending order. In second scanning of the database, construction of FP-tree is done. In a second step, with the help of a FP - tree algorithm, extraction of frequent pattern is made from a FP - tree. It is done with the help of conditional FP-tree. For frequent pattern base, conditional FP-tree is constructed and frequent pattern are extracted from conditional FP-tree [12].

Conditional FP tree and Conditional Pattern Base uses node link property and prefix path property. Figure 4(a) shows conditional pattern base for each item in the table. Figure 4(b) shows construction of conditional tree for I5. For frequent items, construction of conditional FP tree is done in pattern base. When construction of FP tree is done then we can extract frequent pattern from FP tree as given in Table 1 [10]. The FP growth algorithm has three advantages: First, it scans database only two times and it also decreases computational cost Second, generation of candidate itemsets is removed in FP tree algorithm. Third, it reduces the search space by using divide and conquers method. The FP growth algorithm has one disadvantage. It cannot used in incremental mining because when new Transactions are added to the database, FP tree requires updating in data sets and whole process are repeated after this [13,14].

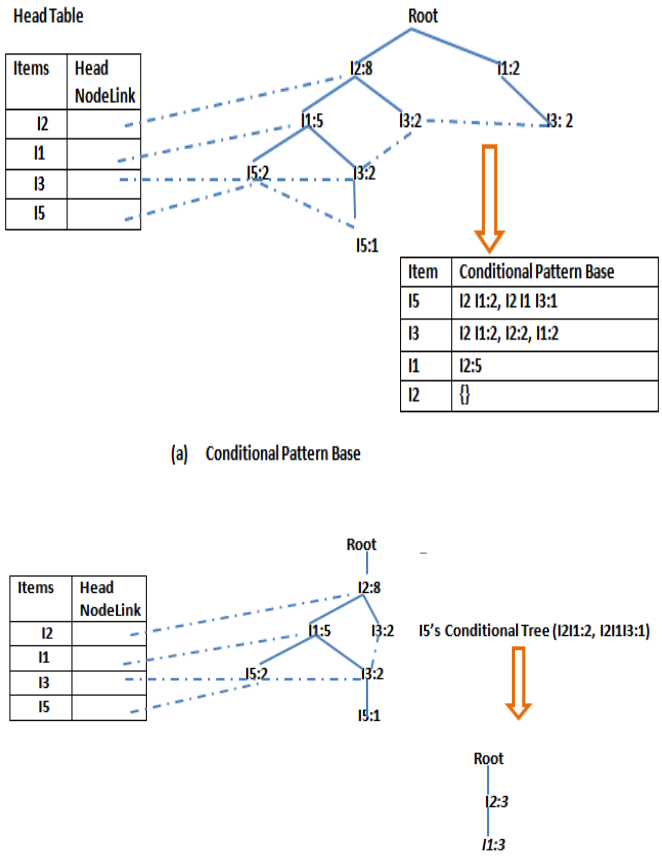

Fig 4: Conditional Pattern Base & Conditional FP Tree

Table 1: Conditional FP Tree and Associated Frequent Patterns

ISSN: 2278 – 1323

Item	Conditional Tree	Frequent Pattern
I5	{(*I2 I1*)}| *I5*	*I2* I5, *I1* I5, *I2 I1 I5*
I3	{(*I2 I1*)}| *I3*	*I2* I3, *I1* I3, *I2 I3 I5*

E. The process of generating frequent pattern using multi-level association mining

In real life, various applications are used. It is difficult to find, strong association rules between data sets at low or primitive level of abstraction in the multi-dimensional functionality. Strong association rules which is generated at a higher level May be common sense to some candidate, but is can be difficult for others. Multilevel association rule mining is used to mine strong association rule between intra and inter different levels of abstraction [15].

At starting level, mining association rules are loses detailed information and it will show only specific rules without Capable of getting inner part of the rule. Data mining should also be present for mining association rules at Different levels of abstraction. Each transaction can be encoded, which is based on dimension and levels in association rules [16].

In multilevel association rule mining, each item in the database is identified by using a concept hierarchy. In hierarchy, mining will be happen at multiple levels. There might be no rule that matches at the constraints at the lowest levels. At high levels rules can be general. In multiple level associations, a top-down approach is used in which the support count is same or varies from level two levels (Support will be reduced from higher level to lower levels) [17].

Numerous methods have been discussed for association rule mining [18], [19], [20] and [21]. This paper [22] presents an efficient method, Prefix-span algorithm, for placing products on shelves in supermarkets. This method mines all sequential patterns from a customer transaction database. From this, the products are assigned to shelves based on these sequential orders of mined patterns. This algorithm uses a pattern growth methodology which finds sequential pattern using in two steps. In the first step, mining of the sequence of the product categories is done and then products are placed on shelves according to sequence order of mined patterns. In the second step, again for each category patterns are mined using Prefix-span algorithm and then reorganize the products under the category by combining the profit calculation on mined patterns.

In this paper [23], filtration approach is proposed, which is used alternately to the pruning method used in Apriori algorithm. This new method can generate optimum numbers of candidate k-frequent itemsets and all infrequent itemsets are eliminated.

Granular computing association rule [24], Rapid association rule mining[10], Equivalence Class Transformation algorithm[10], Associated Sensor Pattern Mining of Data Stream[10], Positive and negative association[25][26], Pattern Growth approach, Agent association rule[27], Critical Relative Support (CRS) to mine critical least association rules[28], Maximal and closed frequent pattern mining algorithms[29], Boolean Matrix with Hadloop [30], Association using neural network[31] and Association rule mining for clustering[32], [33], [34] are seen in the literature.

IV. COMPARATIVE ANALYSIS

Algorithm Name	Search Type	Number of Passes/Scans	Data Structure
AIS[5]	Breadth First Search	K+1	List
Apriori[8]	Breadth First Search	K+1	Hash Tree + Hash Table
Apriori Dynamic	Bottom up Method	2	Count Table
FP-Growth[29]	Divide and Conquer	2	FP-tree
Multi-Level Association	Top-down Method	1	Table

Table 2: Comparative Analysis

V. PROPOSED METHODOLOGY

Many ideas come from different abstraction levels and single level frequent pattern mining, which requires more effort for mining transaction database. This paper describes a method for mining frequent patterns in customer behavior. In proposed methodology we are trying to combine some user domain knowledge and effective interesting measures in the processing step which reduce the number of scan and memory requirement. So the methods will be applied as efficiently as possible to improve any business process. It will be denote cross level suggestion in a hierarchical scheme. Using this method a merchant will get a multiple level of abstract ideas in place of just single level suggestions.

VI. CONCLUSION

We have comparatively analyzed various frequent pattern mining algorithms like Apriori, FP Growth, AIS, and multilevel in mining association rule and data mining. We compared these methods using similar datasets to identify their edge characteristic feature. Major problem in this subject were how to avoid complex candidate generation process, execution time and memory requirements for large datasets and found FP-tree perform well as compared to other algorithms. So in proposed methodology we will combine the multilevel association with FP-tree algorithm for getting more efficient result at cross level.

REFERENCES

[1] Kantardzic M, "Data Mining: Concepts, Models, Methods and Algorithms", New Jersey: Wiley, 2003.

[2] J. Han and M. Kamber, "Data mining: concepts and techniques,"Morgan Kaufman Publishers, 2012.

[3] Sourav S. Bhowmick Qiankun Zhao, "Association Rule Mining: A Survey," Nanyang Technological University,Singapore, 2003.

[4] Jiawei Han · Hong Cheng · Dong Xin · Xifeng Yan, "Frequent pattern mining: current status and future Directions," Data Mining Knowl Discov, vol. 15, no. I, p. 32, 2007.

[5] Shengwei Yi, Tianheng Zhao, Yuanyuan Zhanga, Shilong Ma, Zhanbin Che, "An effective algorithm for mining sequential generators", Advanced in Control Engineering and Information Science Elsevier, 2011.

[6] K.Prasanna, M.Seetha, "Mining High Dimensional Association Rules by Generating Large Frequent K-Dimension Set", International Conference on Data Science & Engineering, IEEE, 2012.

ISSN: 2278 – 1323

[7] Djoni Haryadi Setiabudi, Gregorius Satia Budhi, I Wayan Jatu Purnama, Agustinus Noertjahyana, "Data Mining Market Basket Analysis' Using Hybrid-Dimension Association Rules, Case Study in Minimarket X", International Conference on Uncertainty Reasoning and Knowledge Engineering, IEEE, 2011.

[8] Ozgur Cakira, Murat Efe Aras, "A recommendation engine by using association rules", Elsevier, 2012.

[9] Feri Sulianta, Imelda Atastina, Thee Houw Liong, "Mining Food Industry's Multidimensional Data to Produce Association Rules using Apriori Algorithm as a Basis of Business Strategy", International Conference of Information and Communication Technology, IEEE, 2013.

[10] Shamila Nasreen, Muhammad Awais Azam, Khurram Shehzad, Usman Naeem, Mustansar Ali Ghazanfar "Frequent Pattern Mining Algorithms for Finding Associated Frequent Patterns for Data Streams: A Survey", The 5th International Conference on Emerging Ubiquitous Systems and Pervasive Networks, Elsevier, 2014.

[11] Dharmesh Bhalodiya, K. M. Patel, Chhaya Patel, "An Efficient way to Find Frequent Pattern with Dynamic Programming Approach", IEEE conference, 2013.

[12] Shruti Mishra, Sandeep Kumar Satapathy, Debahuti Mishra and Vinita Debayani Mishra, "An Approach to Frequent Pattern Discovery from Gene Expression Data using PSO Variants", International Conference on Modeling Optimization and Computing, Elsevier, 2012.

[13] M.S.B. PhridviRaj, C.V. GuruRao, "Data mining – past, present and future – a typical survey on data streams", The 7th International Conference Interdisciplinarity in Engineering, Elsevier, 2013.

[14] Jing Guo, Peng Zhang, Jianlong Tan, Li Guo, "Mining Hot Topics from Twitter Streams", International Conference on Computational Science, Elsevier, 2012.

[15] Sun Lianglei, Li Yun, Yin Jiang, "Multi-Level Sequential Pattern Mining Based on Prime Encoding", International Conference on Applied Physics and Industrial Engineering, Elsevier, 2012.

[16] Carlos Roberto Valêncio, Fernando Takeshi Oyama, Paulo Scarpelini Neto, Angelo Cesar Colombini, Adriano Mauro Cansian, Rogéria Cristiane Grat de Souz, Pedro Luiz Pizzigatti Correa, "MR-Radix: a multi-relational data mining algorithm", Springer, 2012.

[17] Esma Nur Çinicioglu, Gürdal Ertek, Deniz Demirer, Hasan Ersin Yoruk, "A Framework for Automated Association Mining Over Multiple Databases" IEEE,2011.

[18] Yaqiong Jiang,Jun Wang, "An Improved Association Rules Algorithm based on Frequent Item Sets", Advanced in Control Engineeringand Information Science, Elsevier, 2011.

[19] Pornsak Deekumpa, Pitikhate Sooraksa, "Associate Rule Minimization using Boolean Algebra Set Function", The 4th Joint International Conference on Information and Communication Technology, Electronic and Electrical Engineering, IEEE, 2014.

[20] M. Thilagu, R. Nadarajan, "Efficiently Mining of Effective Web Traversal Patterns With Average Utility", International Conference on Communication, Computing, and Security, Elsevier, 2012.

[21] Zailani Abdullah, Tutut Herawan, Noraziah Ahmad, Mustafa Mat Deris, "Extracting highly positive association rules from students' enrollment data", Elsevier, 2011.

[22] George Aloysius, D. Binu, "An approach to products placement in supermarkets using PrefixSpan algorithm", Journal of King Saud University – Computer and Information Sciences, Elsevier, 2013.

[23] Lalit Mohan Goyal, M. M. Sufyan Beg, "An efficient Filteration approach for mining association rule", IEEE, 2014.

[24] Xiaojun Cao, "An Algorithm of Mining Association Rules Based on Granular Computing", International Conference on Medical Physics and Biomedical Engineering, Elsevier,2012.

[25] Kushal Bafna, Durga," Feature Based Summarization of Customers' Reviews of Online Products", 17th International Conference in Knowledge Based and Intelligent Information and Engineering Systems, 2013.

[26] Weimin Ouyang, "Mining Positive and Negative Association Rules in Data Streams with a Sliding Window", Fourth Global Congress on Intelligent Systems, IEEE, 2013.

[27] Wang Xiaohu, Wang Lele, Li Nianfeng, "A Fast Search Algorithm Based on Agent Association Rules", International Conference on Solid State Devices and Materials Science, Elsevier, 2012.

[28] Zailani Abdullah, Tutut Herawan, Noraziah Ahma, Mustafa Mat Deris, "Mining significant association rules from educational data using critical relative support approach", Elsevier, 2011.

[29] Caiyan Dai, Ling Chen, "An Algorithm for Mining Frequent Closed Itemsets in Data Stream", International Conference on Applied Physics and Industrial Engineering, Elsevier, 2012.

[30] Honglie Yu, Jun Wen, Hongmei Wang ,Li Jun, "An Improved Apriori Algorithm Based On the Boolean Matrix and Hadoop", Advanced in Control Engineeringand Information Science, Elsevier 2011.

[31] Agus Mansur, Triyoso Kuncoro, "Product Inventory Predictions at Small Medium Enterprise Using Market Basket Analysis Approach - Neural Networks",ICSMED, Elsevier, 2012.

[32] Mahmoud Houshmand, Mohammad Alishahi, "Improve the classification and Sales management of products using multi-relational data mining", IEEE, 2011.

[33] Dake Zhang, Kang Jiang, "Application of Data Mining Techniques in the Analysis of Fire Incidents", International Symposium on Safety Science and Engineering in China, Elsevier, 2012.

About Authors:

Ms.Deepika Fole received the B.E. degree from Chhattigarh Swami Vivekanand Technical University, Bhilai (C.G.) India in Computer Science & Engineering with Honors in the year 2013. She is currently pursuing M.Tech. Degree in Computer Science Engineering with specialization in Computer Science & Engineering from CSVTU Bhilai (C.G.), India. Her research area includes Data Mining and Artificial Intelligence etc.

Ms.Chaitali Choudhary is currently Assistant professor in Department of Computer science & Engineering RCET, Bhilai (C.G.) India. She completed her B.E and M.Tech. in Computer Science and Engineering Branch. Her research area includes Neural Network, computer Network, Artificial Intelligence etc. She has published many Research Papers in various reputed National & International Journals, Conferences, and Seminars.

ISSN: 2278 – 1323