Cross Level Frequent Pattern Mining Using Dynamic Programming Approach

Ms. Deepika Fole[1], Asst Prof. Omprakash Dewangan[2]

***Abstract*— Data mining is a way of generating useful pattern from the large amount of datasets and this pattern is utilized in market basket analysis, detecting fraud activity, cross level marketing. Data gathering and storage facility make it available for organizations to build up enormous amounts of data at low cost. 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 a 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) tree using a dynamic approach algorithm for customers' behavior which describes a hierarchical scheme to provide cross-level suggestions for merchants. FP tree uses divide and conquer mechanism. It requires less time and no candidate generation as compare to the traditional Apriori algorithm. A lot of experiments about the performance on datasets are conducted. The results show that the algorithm proposed is to give better result at cross level in place of single level frequent pattern mining for a customer as well as the merchant. The experimental results verify the compactness and the efficiency of mining shown by the proposed method.**

Keywords*— 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, for obtaining current, valuable, significant and absolute existing information. From many years frequent pattern mining has been an important area in the field of data mining [1].

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 the first frequent pattern mining conceptualized 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. On an item set having a 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 various products, for promotion rules, searching for text, wireless sensor networks, other applications which require observation of user surrounding consciously that are subject to critical conditions or hazards such as leaking of gas and fire explosion.

In this research a frequent pattern structure method is proposed. In it an extended 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 research 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 salient ambitions of this fact-finding work:

- It merges multilevel hierarchy intelligence encoded method and frequent structure mining technique. This gives various ideas for shopkeepers;
- We are performing some test in a laboratory for calculating the proposed methodology with the help of data sets and find what customers actually want;
- Build a group of some useful cases for understanding the various requirement of the customers;

The remainder of this paper is organized as follows: Section 2 describes the literature review. Section 3 describes the problem statement of the research field. Section 4 describes proposed methodology, Section 5 includes experiment and result and finally we conclude this paper in section 6.

II. LITERATURE REVIEW

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. 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 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 item

ISSN: 2278 – 1323

sets. 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.The Apriori Dynamic programming approach overcomes the problem of computational overhead which is occurring in the traditional Apriori algorithm.

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. 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].

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 item sets and all infrequent item sets 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 Had loop [30], Association using neural network[31] and Association rule mining for clustering[32], [33], [34] are seen in the literature.

III.PROBLEM STATEMENT

The main motive of this step is to identify the problem area and explain the problem in general term. The supermarket is most important application area for data mining because it collects different and large amount of data on history of customer shopping, good consumption, service and transportation. Improvement in bar code technology has made it possible for the retail industry to collect and keep large amount of data which is called as basket data. This type of market basket databases consists of the huge amount of transactional database records. Each record consists of the detail of each item purchased by customer on a single purchase round. Market Basket Analysis becomes a key factor of success in the competition of market merchants. Market basket provides a manager with knowledge of customers and their purchasing behavior which gives a high potential to their supermarket. Recent market research gives a suggestion that in-store stimuli, such as product display, shelf-space allocation, have a great influence upon customer buying behavior.

In base paper the frequent patterns for Market Basket Analysis are restricted to the same layer i.e. the single level and base paper methodology is Apriori Dynamic programming. If merchant want to add more stores in one place this methodology cannot help him. This methodology gives suggestion for only a single store in one place. If there is a retail shop in a place, it can generate the suggestion for only retail items that purchase together by customer. If we add one or more shop in one place like retail, electronics, clothes etc. Apriori dynamic programming methodology cannot help the merchant for suggestion that which type of items customer purchase together, if customer buys products from different stores and how to become more effective there shop from other shops. The methodology which is in base paper will not be extending to cross level or multilevel because in this we focus on calculating large candidate 1-itemset and 2-itemset. We can discover more frequent candidate item set like 3-itemset, 4-itemset, up to K-item set with Dynamic Programming approach. Mapping of more than 2-itemset candidate in cross level may be possible with transaction ID. This possibility helps us to find out reaming frequent item set.

IV. METHODOLOGY

This section describes a method for mining frequent patterns in customer behaviour. In this section we will combine the

ISSN: 2278 – 1323

multilevel association and FP tree algorithm for cross level suggestion which denotes a hierarchical scheme. Using this method a merchant will get a multiple level of abstract ideas in place of just single level suggestions. Figure 4.1 shows the diagrammatic view of the proposed methodology. In this we mainly used FP (Frequent Pattern) tree for generating frequent pattern from transaction database. From review, we can conclude that the FP tree method is suitable for cross level suggestion because it requires a number of scans and reduces search space than the rest of the algorithm.

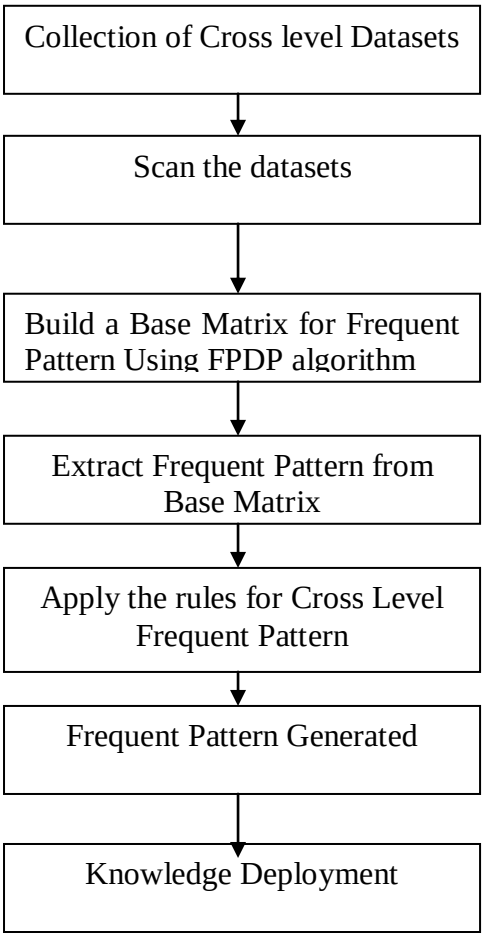

Figure 1: Proposed Methodology

Step 1: **Collection of Cross Level Data**: Data assembly includes collecting of data, in this thesis for testing purpose market data set is used which is collected from a website for frequent item set repository. It consists of 200 transactions at cross level for mining purpose. The dataset consist of list of purchased items separated by space. There is 200*18 matrix is present to show the transaction table. Mainly there are three levels, which are worked in cross level format. We have to find out the association among these three level item sets. The format of the dataset in each box contains three digit numbers because our tree for cross-level frequent pattern is on three levels. 1 super parent followed by 9 parent and then child node which contains three digit data sets.

Step 2: **Scan the Database**: In this basically data transformation technique is used. In this step, we generate id for each transaction in the transaction table. There are 200 transactions are present in the transaction table. Therefore, we generate 200 transaction id (Tid), one for each transaction. We are generating transaction id because with the help of this we can easily identify unique item. For the unique identification purpose, we are using the prime number concept. We give each child node a prime number. For generating transaction id we are multiplying those prime numbers which are present in a transaction. After that multiplying item set, we have a unique id for each and every transaction presents in the data sets. This is called as Data Transformation Technique.

Algorithm 1: Data Transformation Technique
Input: Transaction of the different customer or Transaction Datasets.
Output: Unique Transaction Id for Each Transaction
Procedure:
Step 1: Scan the Database.
Step 2: Give prime number pr for each item ir in each transaction. Then Transaction ID (Tid) is computed by

$$\text{Tid}=\prod_{j}^{k} pr$$

Where T= (tid, X), X= $\{i_j,\ldots,j_k\}$ Item in the transaction.
If there are no more transaction, then GOTO step 3
Step 3: Stop.

For simplification purposes of understanding the data transformation method, we are examining it by an example. Let item set I = {111, 112, 121, 122, 211, 212, 221, 222, 231, 232, 311, 312, 321, 322} and the database, DB, be the first two columns of table 1 with 7 transactions. As shown in the fourth column of Table 1, DB can be presented by TVs, which very smaller than original transactions.

Table 4.1: The transaction database DB and its Transaction Values

Transaction	Items	Transformed	Tid (TV)
1	231, 321, 322, 312	23,41,43,37	1500313
2	111, 112, 212, 321, 322, 231	2, 3, 13, 41, 43, 23	31628822
3	322, 321, 111, 212, 231	43, 41, 2, 13, 23	1054274
4	111, 222, 112, 212, 211	2, 19, 3, 13, 11	16302
5	112, 211, 212, 311	3, 11, 13, 31	13299
6	111, 311, 212, 112, 231	2, 31, 13, 3, 23	55614
7	231, 121, 321	23, 5, 41	4715

ISSN: 2278 – 1323

Step 3: **Build a Base Matrix for Frequent Pattern Using FPDP algorithm**

To build a base matrix for frequent pattern, Firstly, we generate unique identification code for parent and super parent. The detail description of this given below:

Generating Id for Parent:
Each child node having their parent at just one layer above, we have to find out there parent ID related to each parent i.e. at level 2 we have to find out parent ID. From data transformation technique we have to know that each child has assigned to a unique prime number. We multiply the child's prime number related to each parent and this is called parent ID for each parent.

Generating ID for Super Parent:
Above the parent level, there is a super parent, which is calculated with the help of parent ID. Each level has a unique parent ID; we multiply this parent id related to each super parent and generate a Super Parent ID for each super parent.
After generating Parent ID and Super Parent ID, we generate Base matrix for frequent Pattern. The algorithm for generating Base Matrix is as follows:

Algorithm 2: FPDP (Frequent Pattern Using Dynamic Programming Approach) Algorithm

Input: Unique Transaction Id for Each Transaction
Output: Association rules for Cross Level Frequent Pattern Mining
Procedure:
Step 1: Scan the Transaction ID for each transaction.
Step 2: For 1st Level (Super-Parent)
(a) Create a unique id for each node. Call it Spar.
(b) Divide unique id of super parent (Spar) by tid
if remainder=0
then set=1 in base matrix
else
set=0
For 2nd Level (Parent)
(a) Create unique id for each node. Call it Par.
(b) Divide unique id of parent (Par) by tid
if remainder=0
then set=1 in base matrix
else
set=0
For 3rd level (item)
(a) Divide tid by the unique id of each item
if remainder=0
then set=1 in base matrix
else
set=0
Step 3: Define Minimum Support (Min_sup) for each level.
if the frequency of item >= Min_sup at each level
then
Add the item in threshold matrix
else
Discard the item.
Step 4: If the frequency of referred item is equals to items appearing in the list,
then
add it into the rules
else
Discard the item.
Step 5: End.

Step 4: Extract Frequent Pattern from Base Matrix

In this step we define minimum support for each level. We count that how many times one particular item appears in a database, if it satisfies the minimum support then we add that particular item in the threshold matrix otherwise we discard that item. This step is shown in algorithm 2 at step 3.

Step 5: Apply the rules for Cross Level Frequent Pattern

We identify rules for cross level frequent pattern in this step. If the frequency of referred item is equal to the items appearing in the list, then add it into the rules. This step is shown in algorithm 2 at step 4.

Step 6: Frequent Pattern Generated

After step 5, frequent pattern for cross level suggestion are generated according to our methodology.

Step 7: Knowledge deployment: Here visualization and knowledge representation techniques are used to present mined knowledge to the users.

V. EXPERIMENT & RESULT

This chapter demonstrates the experiment that we have to perform to evaluate the new scheme. For the evaluation purpose, we have conducted several experiments using the existing data set. Those experiments performed on computer with Intel Pentium CPU 2.10 GHz, 2.00 GB RAM and hard disk 100 GB. For Data Analysis purpose we are using weka too. This algorithm is developed in MATLAB for the unit of measuring the time in seconds.

We have performed the experiment with different size of the original database and with different support threshold. The following are snapshots of the implementation of the algorithm. The data set used in this experiment is a supermarket dataset. In this experiment we have taken dataset with 200 transactions. All results are ten times performed in our laboratory, then taking mean values of that and shows that our methodology is take 81% less time as compare to Apriori algorithm in fig 5.1. The proposed software will be efficiently generating frequent pattern in cross level with the input data

ISSN: 2278 – 1323

sets. The data set was obtained from the UCI repository of machine learning databases. The data set was donated by Tom Brijs and contains the supermarket data from an anonymous Belgian store. Our database contains 200 different transactions and 783 items. Fig 5.1 represents bar chart of time requirement of individual algorithms. X-axis draws horizontally with the interval of 50 items. And Y-axis draws vertically with time. Following discussion, we refer traditional algorithm as Apriori and proposed algorithm as FPDP (Frequent Pattern tree with Dynamic Programing). First, we take a constant Minimum support count (min_sup) 2% with different items. This will be used for effective improvement in proposed change by reducing two database scan and to utilize memory space to store cross-level datasets.

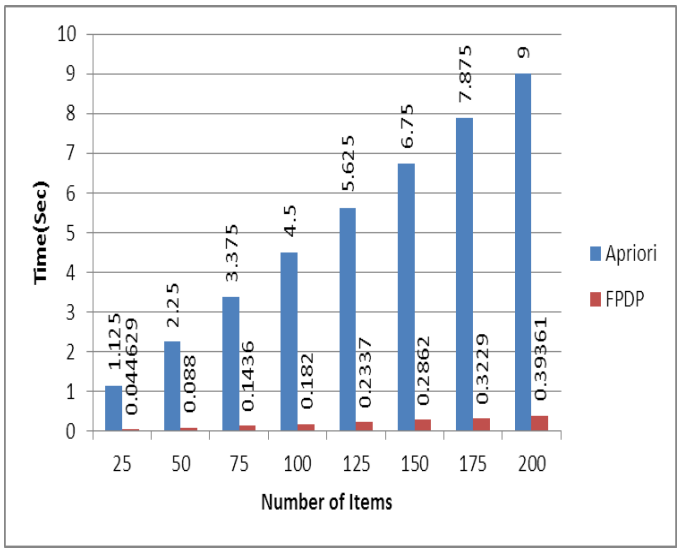

Figure 5.1: Apriori and FPDP Execution time with 2% min_sup

VI. Conclusion

The association rules play a major role in many data mining applications, trying to find interesting patterns in databases. In order to obtain these association rules the frequent sets must be previously generated. We have comparatively analyzed various frequent pattern mining algorithms like Apriori, FP Growth, AIS, and multilevel in mining association rule and data mining. The work presented in the thesis is for cross-level association rule mining. The work can be enhanced to generate multi-dimensional association rules. A tool for generating association rules can be developed. This tool can choose the approach for frequent item sets mining, according to the properties of the dataset to be mined. Therefore, in our methodology we are combining the multilevel association with a FP-tree algorithm for getting more efficient results at cross level.

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, lots of database scan and execution time and memory requirements for large data sets and found the FP-tree perform well as compared to other algorithms.

VII. Future Work

We still focus on calculating for more number of cross levels. We focus on only 3 levels, i.e. child, parent and super parent. We can discover more frequent candidate item set for more level like 4-level, 5-level, up to K-level with the Dynamic Programming approach. Mapping of more than 3-level may be possible with the transaction ID. This possibility helps us to find out reaming frequent item set.

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.

[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.

ISSN: 2278 – 1323

[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 Engineering and 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 Prefix Span 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 on 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.

Mr.Omprakash Dewangan is currently Assistant Professor in Department of Computer science & Engineering RCET, Bhilai (C.G.) India. He completed his M.Sc. and M.Tech. in Computer Science and Engineering with specialization in Computer Technology from CSVTU Bhilai (C.G.). His research area includes Data Mining, Image processing, Artificial Intelligence etc. He has published many research papers in various reputed National & International Journals, Conferences, and Seminars.

ISSN: 2278 – 1323