VARIOUS FREQUENT ITEM SET BASED ON DATA MINING TECHNIQUE

Manisha Kundal[1], Dr. Parminder Kaur[2]

[1]*Research Scholar, Computer Science, GURU Nanak Dev University, Amritsar*
[2]*Assistant Professor, Computer Science, GURU Nanak Dev University, Amritsar*

ABSTRACT - As with the progression of the IT technological innovation, the quantity of gathered information is also increasing. It has led to lots of information saved in information source, manufacturing facilities and other databases. Thus the Data exploration comes into picture to discover and evaluate the information source to draw out the interesting and previously unidentified styles and rules known as organization concept exploration. This document has focused on regular itemset related based aprioi methods. The overall purpose is to find various restrictions of current methods. The regular itemset exploration has found to be crucial and most expensive step in organization concept exploration. Mining regular styles from extensive information source has appeared as an important problem in information exploration and knowledge finding community. This document ends up with suitable future guidelines to improve the regular item set further.

Key Words:-KNOWLEDGE DISCOVERY PROCESS, ASSOCIATION RULES IN DATA MINING, PROPOSED ALGORITHIMS

1. INTRODUCTION

Information exploration is a procedure of finding information from the data source. Information exploration is a procedure that uses a variety of information research tools to discover styles and relationships in data that may be used to make legitimate forecasts. Information exploration is used to obtain information from the information sets. A information finding procedure includes different stages.

- Data cleaning: also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.
- Data integration: at this stage, multiple data sources, often heterogeneous, may be combined in a common source.
- Data selection: at this step, the data relevant to the analysis is decided on and retrieved from the data collection.
- Data transformation: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.
- Data mining: it is the crucial step in which clever techniques are applied to extract patterns potentially useful.
- Pattern evaluation: in this step, strictly interesting patterns representing knowledge are identified based on given measures.
- Knowledge representation: is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

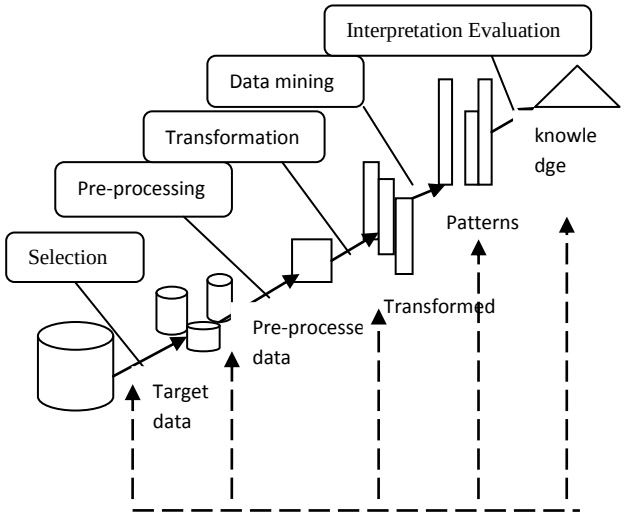


Figure 1: Knowledge Discovery Process [1]

1.Related Work

Archana Singh et al. [1] recommend an organization concept exploration techniques are used to discover the connection between the various product places in the data source. The Apriori criteria is the basic ARM criteria, but it needs so many data source tests to discover regular items. In the suggested criteria an enhanced data framework adjacency matrix is used. Rohiza Ahmad et al. [2] recommend a new binary-based Semi-Apriori strategy that effectively finds the regular itemsets. Comprehensive tests had been performed using the new strategy, as opposed to current Apriori methods.Wei Zhang et al. [3] an improved apriori requirements so known as FP-growth requirements that will help take care of two neck-bottle problems of traditional apriori requirements and has more efficiency than exclusive one. Test results are confirmed, that the requirements has higher discovery efficiency in efficiency time, storage space usage and CPU usage than most present ones like Apriori..Goswami D.N. et al. [4] described three different regular design exploration techniques (Record narrow, Junction and Suggested Algorithm) are given depending on traditional Apriori criteria. In this works a relative research of all techniques on dataset of 2000 deal.Basheer Mohamad Al-Maqaleh et al. [5] suggested an efficient criteria to incorporate assurance evaluate during the process of exploration regular itemsets, which produces assured regular itemsets. This technique has been applied and the trial results show the efficiency and efficiency of the suggested criteria. Saurabh Malgaonkar et al. [6] described that the described system is developed to discover the most regular mixtures of items. Three different methods from organization exploration have been inserted and then best mixture method is used to discover more exciting results. The specialist then is capable of doing the information exploration and removal and lastly determine the result and make appropriate decision. Patel Tushar S. et al. [7] present detail research of methods and talks about some problems of producing regular itemsets from the criteria. The unifying function among the inner working of various exploration methods is researched. The relative research of methods contains factors like different support principles is mentioned. Anitha Modi1 et. al. [8] described the issue of exploration regular itemsets occurs in huge transactional information source where there is need to find organization guidelines among the transactional information for the development of business. The study of various methods for exploration regular itemsets in transactional information source that work on horizontally, straight, estimated and multiple structure datasets is provided. Mihir R. et al. [9] present suggesting method that can be along with Apriori criteria and decreases storage space needed to shop applicant and the performance time by reducing CPU time. The concept of gate is purposed based on support value to reduce the performance efforts and overall storage space space needed to shop applicant generated during checking of dataset. Damor Nirali N. et al. [10] described a new means for producing regular itemsets using regular itemset shrub (FI-tree). Also explain the example of new technique and its result research using bottles dataset. The performance time of purposed technique is better evaluate to SaM technique. Deepak Vidhate et at. [11] explain about ideas of organization exploration, statistical design growth for Multilevel Connection Criteria (MRA) and Execution & Result Research of MRA and efficiency evaluation of MRA and Apriori algorithm. M.M Sufyan Beg et at. [12] recommend an different to Apriori algorithm's trimming phase is suggested. This

substitute is portrayed as a purification phase. Ashika Gupta et at. [13] present the research concentrates on web use finding and particularly keeps an eye on running across the web utilization cases of sites from the server log information. The connection of storage space and time utilization is in evaluation by means of Apriori specifications and enhanced Regular Design Place specifications. Jugendra Dongre et at. [14] described the Apriori criteria is one of the most popular information exploration strategy for finding regular product places from a deal dataset and obtain organization guidelines. The apriori criteria on simulated information source and discovers the organization guidelines on different assurance value.Enrique Lazcorreta et at. [15] current mainly investigates the process of discovering company recommendations in this kind of big data source and of modifying them into user-adapted recommendations by the two-step personalized Apriori technique. Starting results have confirmed that it is possible to run this style in web sites of technique dimension. Zhang Yongliang et at. [16] described the cost, protection and aggressive of data migration, a allocated organization concept exploration criteria based on matrix known as DARMO is put forward for some special allocated programs. This criteria has some features such as high level of parallelism, less data source checking, less interaction expense and low complexness. Lastly, the complexness, similar price, speedup and scalability of the criteria is examined, and efficiency of the criteria is confirmed by example research and trial simulator. Zhou Zhiping et at. [17] suggested matrix organizing catalog criteria greatly enhanced the data exploration performance and scalability.

3. ASSOCIATION RULES IN DATA MINING

3.1 Frequent Item sets

Frequent designs are designs such as product places, sub series, or substructures generally exist in the real world databases. Frequent designs works an essential role in discovery companies, relationships, and many other interesting relationships among details. Moreover, it helps in details classification, clustering, and other details discovery tasks as well. Thus, frequent design discovery has become an important info discovery task.

The feedback of frequent product places exploration is a deal data source, a minimum assistance limit. For the outcome of frequent product set, first find the frequent product set, then estimate assistance count each product set by checking the deal data source. There have been a lot of excellent methods developed for getting frequent product places in very large data source.

3.2 Association rule mining

Company concept exploration or association exploration is determined, is the process of finding association or connections between data source products. It find the application in industry container research (MBA).Market specialist would be interested in recognized frequently bought products, so that organization can follow display space management and efficient sales techniques.

Let $I=\{I1, I2, \ldots, Im\}$ be a set of itemset, T be deal that contains a set of products such that $T \subseteq I$, D be a data source with different deal information. An organization concept is an effects through $X \Rightarrow Y$, where $X, Y \subset I$ are places of products known as product places, and $X \cap Y = \emptyset$. X is known as antecedent while Y is known as major, the concept indicates X indicates Y.

There are two essential primary activities for company recommendations, assistance and guarantee. Since the databases is huge and clients problem about only those regularly purchased items, usually boundaries of assistance and guarantee are pre-specified by clients to drop those recommendations that are not so interesting or useful. The two boundaries are known as little assistance and little guarantee respectively. Support of an company idea is identified as the percentage/fraction of details that contain $X \cup Y$ to the depend of details in the databases. Assurance of an company idea is identified as the percentage/fraction of the wide range of transactions that contain $X \cup Y$ to the depend of details that contain X. Assurance is a assess of durability of the company recommendations. In typical, a set of items is known as items set. The wide range of items in items set is known as the length of items set. Item locations of some length k are usually known as k product locations.

Usually, an organization guidelines exploration criteria contains the following steps:

- The set of applicant k-item places is produced by 1-extensions of the huge (k- 1) Product places produced in the past version.
- Support for the applicant k-item sets are produced by a successfully omit the data source.
- Product places that do not have the lowest support are removed and the staying Product places are known as large k-item places.[1]

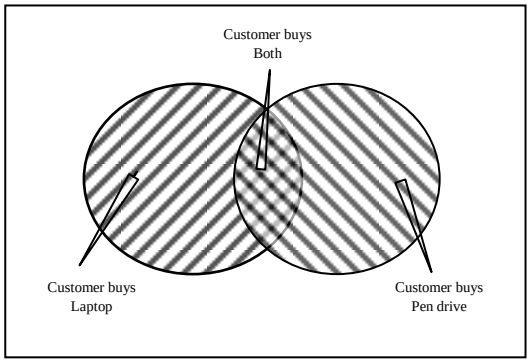


Figure 2: Association Rules [1]

3.3 Frequent Itemset mining

With the improve in data source a huge variety of regular itemsets are produced so there is a need of strong association rules. Trimming insignificant styles is the major process in regular pattern exploration that lead to the finding of methods for regular itemset exploration. Frequent itemsets exploration is the process of finding the styles that regularly occurs. When the threshold value is small, a huge variety of regular itemsets are produced. It prunes the itemsets that are not occurring regularly in the data source. It helps to save the storage space for storing huge data source by pruning the insignificant itemsets. It also improve the speed of getting insignificant itemsets by decreasing the variety of data source tests.
Regular itemsets play an essential part in many information exploration projects that try to like styles from data source, such as organization guidelines, connections, classifiers, groups and many more of which the exploration of organization guidelines is one of the most popular problems. The unique inspiration for searching organization guidelines came from the need to evaluate so called grocery store deal information, that is, to analyze client actions with regards to the purchased products. Association guidelines explain how often items are purchased together. For example, an organization concept "laptop & Printing device (80%)" declares that four out of five customers that purchased laptop also purchased printer. Such guidelines can be useful for choices concerning product costs, special offers, store structure and many others.

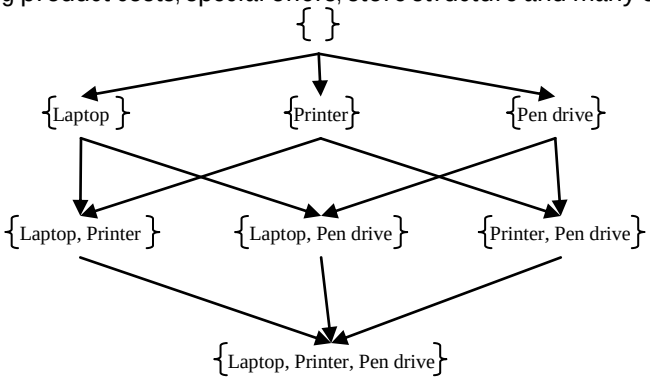


Figure 3: Frequent Itemsets [4]

3.4 Need of Mining Frequent Itemsets

Data source has been used in business control, govt management, medical and technological innovation information control and many other important programs. With the intense development of information, exploration information and knowledge from large data source has become one of the major difficulties for information control and exploration group.
The regular product set exploration is inspired by problems such as industry container research. In a industry container data source a set of products bought by client in a deal. An organization concept excavated from industry container data source

declares that if some products are bought in deal, then it is likely that some other products are bought as well. Finding all such guidelines is useful for directing future sales special offers and store structure. The problem of exploration regular product sets are basically, to discover all guidelines, from the given transactional data source D that have assistance greater than or similar to the user specified lowest assistance.[4]

4.PROPOSED ALGORITHM

Algorithm: Proposed Algorithm
Input: Data Sets
Output: Frequent Itemsets
Step1 input dataset and min threshold value
Step2 calculate the length of the longest item set in the data set
Step3 Apply binary search to find the frequent item set
Low=length of shortest item
High=length of longest itemset
Mid= (low+high)/2

While (low <= high)

1. If we get the itemset at mid level which is greater than threshold value.
then low = mid+1
Else high = mid-1
Step4 Exit

Example to compare Apriori Algorithm with Proposed algorithm:

The following illustrations describe the difference between apriori criteria with suggested criteria. In this example same data source is taken to evaluate both the methods. The suggested criteria decrease the number of goes to find regular itemsets. As apriori criteria have high complexness and suggested criteria decrease the complexness.

Example of Apriori Algorithm:

Step 1: Consider the transactional data source having variety of products with their transactional id's.

TID	List of Item_IDs
T100	I1,I2,I5
T200	I2, I4
T300	I2, I3
T400	I1,I2,I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1,I2,I3,I5
T900	I1,I2,I3

Step 2: Determine the variety of situations of 1-itemsets. Then make the applicant C1.

Step 3: Check out and Evaluate the applicant with lowest assistance depend. Then make the L1 with the 1- itemsets that fulfill the lowest assistance i.e 2.

Scan D for count of each Candidate

C_1

Itemset	Sup_count
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

Compare Candidate with sup_count with minimum supp_count

L_1

Itemset	Sup_count
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

Step 4: Join L_1 with L_1 to generate the candidate set with 2- itemsets, C_2.

Step 5: Calculate the support count of each candidate in C_2

Step 6: Create L_2 after comparing the candidates with minimum support count.

Generate C_2 candidates from L_1

C_2

Item set
{I1,I2}
{I1,I3}
{I1,I4}
{I1,I5}
{I2,I3}
{I2,I4}
{I2,I5}
{I3,I4}
{I3,I5}
{I4,I5}

Scan D for count of each

C_2

Item set	Sup_count
{I1,I2}	4
{I1,I3}	4
{I1,I4}	1
{I1,I5}	2
{I2,I3}	4
{I2,I4}	2
{I2,I5}	2
{I3,I4}	0
{I3,I5}	1
{I4,I5}	0

Compare Candidate with sup_count with minimum sup_count

L_2

Item set	Sup_count
{I1,I2}	4
{I1,I3}	4
{I1,I5}	2
{I2,I3}	4
{I2,I4}	2
{I2,I5}	2

Step 7: Join $L_2 \bowtie L_2$ to generate the candidate set with 3- itemsets, C_3.

$C_3 = L_2 \bowtie L_2 = \{\{I1,I2\},\{I1,I3\},\{I1,I4\},\{I1,I5\},\{I2,I3\},\{I2,I4\},\{I2,I5\}\} \bowtie$

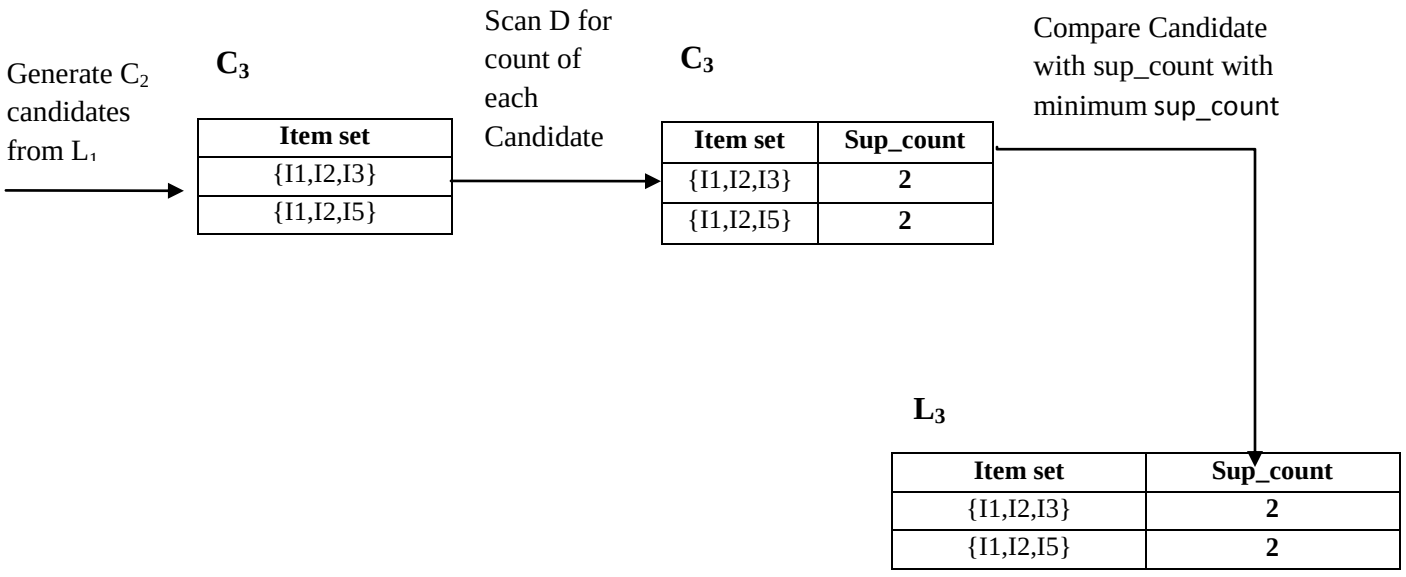
$\{\{I1,I2\},\{I1,I3\},\{I1,I4\},\{I1,I5\},\{I2,I3\},\{I2,I4\},\{I2,I5\}\} = \{\{I1,I2,I3\},\{I1,I2,I5\},\{I1,I3,I5\},\{I2,I3,I4\},\{I2,I3,I5\},\{I2,I4,I5\}\}$

Prune use the Apriori Property[]: All nonempty subset of a frequent itemset must also be frequent.

- The 2-item subsets of {I1,I2,I3} are {I1,I2},{I1,I3}and{I2,I3} and these 2- item subset are members of L_2 means they are frequent, therefore keep {I1,I2,I3} in C_3.
- The 2-item subsets of {I1,I2,I5} are {I1,I2},{I1,I5}and{I2,I5} and these 2- item subset are members of L_2 means they are frequent, therefore keep {I1,I2,I5} in C_3.
- The 2-item subsets of {I1,I3,I5} are {I1,I3},{I1,I5}and{I3,I5} and these 2- item subset are not members of L_2 means they are not frequent, therefore remove {I1,I2,I3} from C_3.
- Check up to all 2-item subsets .

Step 8: Calculate the support count of each candidate in C_3.

Step 9: Create L_3 after comparing the candidates with minimum support count.



Step 7: Join $L_3 \bowtie L_3$ to generate C_4 having 4-itemsets. The join result in {{I1, I2, I3, I5}}, itemset {I1, I2, I3, I5} is pruned because its subset {I2, I3, I5} is not frequent. Therefore, $C_4 = \emptyset$ and the algorithm terminates and find all the frequent itemsets.

Example of Proposed Algorithm:

Step 1: Consider the transactional database having number of items with their transactional id's.

TID	List of Item_IDs
T100	I1,I2,I5
T200	I2, I4
T300	I2, I3
T400	I1,I2,I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1,I2,I3,I5
T900	I1,I2,I3

Step 2: Set the minimum threshold value.

Step 3: Calculate the length of longest itemset.

Step 4: Set high= length of longest itemset and low= length of shortest itemset

Step 5: Calculate mid=(low + high)/2
Set low= 2
and high= 4
mid = (2+4)/2 = 6/2 = 3

Step 6: Create itemset containing number of items= 3 (mid)

Itemset
{I1,I2,I3}
{I1,I3,I4}
{I1,I4,I5}
{I1,I2,I4}
{I1,I2,I5}
{I1,I3,I5}
{I2,I3,I4}
{I2,I3,I5}
{I2,I4,I5}

Step 7: Calculate the support count of generated itemsets.

Itemset	Sup_count
{I1,I2,I3}	2
{I1,I3,I4}	0
{I1,I4,I5}	0
{I1,I2,I4}	1
{I1,I2,I5}	2
{I1,I3,I5}	1
{I2,I3,I4}	0
{I2,I3,I5}	1
{I2,I4,I5}	0

Step 8: The itemset with count greater than threshold value are found.

Step 9: Set low= mid+1
low= 3 +1 =4
high= 4
mid= (4+4)/2 = 8/2 = 4

Step 10 : Create itemset containing number of items= 4 (mid)

Itemset
{I1,I2,I3,I4}
{I1,I2,I3,I5}
{I1,I3,I4,I5}
{I2,I3,I4,I5}

Step 11: Calculate the support count of generated itemsets.

Itemset	Sup_count
{I1,I2,I3,I4}	0
{I1,I2,I3,I5}	1
{I1,I3,I4,I5}	0
{I2,I3,I4,I5}	0

Step 12: The itemset with count greater than threshold value are not found.
high= mid-1
high= 4-1 = 3
low=4
now low>high
therefore terminate the process.

Comparative Analysis

All the methods generate regular itemsets on the reasons for lowest support.

- Apriori criteria is quite effective for market centered research in which dealings are huge but regular products produced is small in number.
- Partition criteria works worse where data source check out time required is less then producing applicants.
- Straight Structure centered methods statements to be quicker than Apriori but require larger storage than horizontally layout centered because they needs to load applicant, data source and TID list in main storage.
- For FP-Tree and H-mine, works better than all mentioned above methods because of no creation of applicant places but the suggestions required to shop kept in storage space need large storage space.

Algorithm	Memory Utilization	Time	Databases
Apriori Algorithm	Require Large Space	More Execution Time	Both Sparse And Dense
DHP	Less Space At Earlier Passes And More Space	Small Execution Time For Small Databases	Medium Databases

Algorithm	At Later Stages		
Partitioning Algorithm	Requires Less Memory	More Execution Time	Large Databases
DIC Algorithm	Variable Memory	Small Execution Time	Medium And Low Databases
Sampling Algorithm	Very Less Amount Of Memory	Small Execution Time	Any Kind Of Database But Not Give Accurate Results
Eclat Algorithm	Requires Less Memory	Small Execution Time	Not Suitable For Small Datasets
FP-Growth Algorithm	Requires More Main Memory	More Execution Time	Medium And Large Databases
H-Mine	Variable Memory	More Execution Time	Both Sparse And Dense

Conclusion and future work

Regular itemset exploration is essential and most expensive step in organization concept exploration. Mining frequent styles from extensive information source has appeared as an important problem in information exploration and information finding group. The significant task found in frequent design exploration is a huge variety of result styles. Various different methods like, apriori, DHP, DIC, Partition, Example, FP-Tree, H-mine etc. have been designed. but these methods are quite effective for market based research in which dealings are huge but frequent items produced is small in variety . Most of them methods execute more intense where information source check out required is less then quicker than Apriori but they cannot control the space complexness. Some modifications works better in heavy datasets but with low support its efficiency degrades for rare datasets.

In near upcoming to improve the efficiency, with various mentioned aspects, an criteria is needed for regular product places. So that it can be used to locate exciting cross-sells and relevant products and many other programs. Therefore in near upcoming we will recommend a new binary search criteria based regular product set strategy to improve the calculations features further.

REFERENCES

[1] Singh, Archana, and Jyoti Agarwal. "Proposed algorithm for frequent item set generation." In Contemporary Computing (IC3), 2014 Seventh International Conference on, pp. 160-165. IEEE, 2014.
[2] Fageeri, Sallam Osman, Rohiza Ahmad, and Baharum B. Baharudin. "A semi-apriori algorithm for discovering the frequent itemsets." In Computer and Information Sciences (ICCOINS), 2014 International Conference on, pp. 1-5. IEEE, 2014.
[3] Wei Zhang, Hongzhi Liao and Na Zhao "Research on the FP Growth Algorithm about Association Rule Mining", IEEE Vol. 1, 2008, Wuhan, pp. 315-318.
[4] Goswami D.N., Chaturvedi Anshu. Raghuvanshi C.S. " An Algorithm for Frequent Pattern Mining Based On Apriori", IJCSE Vol. 2, 2010, pp. 942-947.
[5] Basheer Mohamad Al-Maqaleh and Saleem Khalid Shaab " An Efficient Algorithm for Mining Association Rules using Confident Frequent Itemsets ", IEEE , 2013, Rohtak , pp. 90-94.
[6] Saurabh Malgaonkar, Sakshi Surve and Tejas Hirave, "Use of Mining Techniques To Improve The Effectiveness of Marketing and Sales", IEEE, 2013, Mumbai, India, pp. 1-5.
[7] Patel Tushar S., Panchal Mayur, Ladumor Dhara, Kapadiya Jahnvi, Desai Piyusha, Prajapati Ashish and Prajapati Reecha, "An Analytical Study of Various Frequent Itemset Mining Algorithms", Res. J. Computer & IT Sci., Vol. 1(1), 2013, pp.6-9.
[8] Anitha Modi1and Radhika Krishnan, " Mining Frequent Itemsets in Transactional Database Mining", IJETAE., Vol. 3, 2013, ISSN 2250-2459.
[9] Mihir R. Patel,Dipti P. Rana and Rupa G. Mehta "FApriori: A Modified Apriori Algorithm Based on Checkpoint", IEEE , 2013, Mathura, pp. 50-53.
[10] Damor Nirali N., Radhika Krishnan and Patel Hardik, " A New Method to Mine Frequent Itemsets using Frequent Itemset Tree", Res. J. Computer & IT Sci., Vol.1(3), 2013, ISSN 2320 – 6527, pp. 9-12.
[11] Vidhate, Deepak. "To improve association rule mining using new technique: Multilevel relationship algorithm towards cooperative learning." In Circuits, Systems, Communication and Information Technology Applications (CSCITA), 2014 International Conference on, pp. 241-246. IEEE, 2014.
[12] Goyal, Lalit Mohan, and M. M. Beg. "An efficient filtration approach for mining association rules." In Computing for Sustainable Global Development (INDIACom), 2014 International Conference on, pp. 178-185. IEEE, 2014.
[13] Gupta, Ashika, Rakhi Arora, Ranjana Sikarwar, and Neha Saxena. "Web usage mining using improved Frequent Pattern Tree algorithms." In Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014 International Conference on, pp. 573-578. IEEE, 2014.
[14] Dongre, Jugendra, Gend Lai Prajapati, and S. V. Tokekar. "The role of Apriori algorithm for finding the association rules in Data mining." In Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014 International Conference on, pp. 657-660. IEEE, 2014.
[15] Lazcorreta, Enrique, Federico Botella, and Antonio Fernández-Caballero. "Towards personalized recommendation by two-step modified Apriori data mining algorithm." Expert Systems with Applications 35, no. 3 (2008): 1422-1429.
[16] Zhang, Yongliang, Jie Qin, and Shiming Zheng. "Research on distributed mining algorithm for association rules oriented mass data." In Control Conference (CCC), 2014 33rd Chinese, pp. 492-499. IEEE, 2014.
[17] Zhou, Zhiping, and Jiefeng Wang. "An improved matrix sorting index association rule data mining algorithm." In Control Conference (CCC), 2014 33rd Chinese, pp. 500-505. IEEE, 2014.

BIOGRAPHIES

Manisha kundal is a research Scholar in GNDU Amritsar, Punjab. I have done my bachelor In beant college of engineering technology, Gurdaspur.

