ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 3, Issue 9, September 2015**

# Association Rule Generation in Data Streams using FP-Growth and APRIORI MR Algorithms

Dr. S. Vijayarani, R. Prasannalakshmi

Assistant Professor, Dept. of Computer Science, Bharathiar University, Coimbatore, Tamilnadu, India

M.Phil Research Scholar, Dept. of Computer Science, Bharathiar University, Coimbatore, Tamilnadu, India

**ABSTRACT:** Data stream is used for handling dynamic databases in which data can be arrived continuously, limitless and its size are very large. This situation has created a problem, i.e. to perform the mining process in these database, the existing data mining algorithms are not suitable. In order to perform mining task in data streams there is a need for development of new algorithms and techniques. By using this new algorithms and techniques we can able to perform various data mining tasks, i.e. clustering, classification, frequent pattern mining and association rule mining in data streams. Association rule mining is used to find the association between the data items which are exist in the databases. Even though, the traditional algorithms are not suitable for data streams, this paper concentrated on how to perform association rule generation task in data streams using traditional algorithms in order to find the drawbacks as well as comparing the performance of the traditional algorithms. Frequent Pattern Tree Growth algorithm and APRIORI Map/Reduce algorithms are used for generating association rules in data streams. Performance measures used in this work are execution time and number of association rules generated. From the experimental results we come to know that the performance of FP-Tree Growth algorithm is more efficient than APRIORI Map/Reduce algorithm.

**KEY WORDS**: Association Rules, FP-Tree Growth Algorithm, APRIORI Map/Reduce Algorithm, Rapid Miner tool, Tanagra tool.

## I. INTRODUCTION

The data stream is continuous arrival of data and this data is normally dynamic in nature and its size is very huge. In this situation, it is not possible to perform data mining tasks with the traditional algorithms since those algorithms are suitable for static data bases. Hence, data stream mining needs the development of new algorithms and techniques to perform the data mining tasks. Association rules are described by finding the frequent pattern, links, relationship and the related structures among the data objects in the databases. There are two important steps in association rule mining; first step is to find the frequent data items and second step is to generate association rules from the frequent data items. This work has compared two different types of association rule mining algorithms; they are APRIORI Map/Reduce algorithm and Frequent Pattern Tree Growth algorithm. [15] The main objective of this work is to find the drawbacks of the existing algorithm when it is applied to the data streams and also finding the efficiency of the two existing algorithms APRIORI Map/Reduce and Frequent Pattern Tree Growth.

The paper is organized as follows. Section 2 provides the related works. Proposed methodology and the traditional association rule algorithms are given in Section 3. Section 4 described the experimental results. Conclusion is given in Section 5.

## II. RELATED WORK

***Vijayarani S et al.,*** [17] described frequent item-set mining in data streams. Authors have used ECLAT and RARM algorithms for generating the frequent item-sets. The dataset was divided into five windows with different threshold values. The performance factors used are number of frequent items generated and execution time. From the results, the authors observed that the performance of ECLAT is more efficient than RARM.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 3, Issue 9, September 2015**

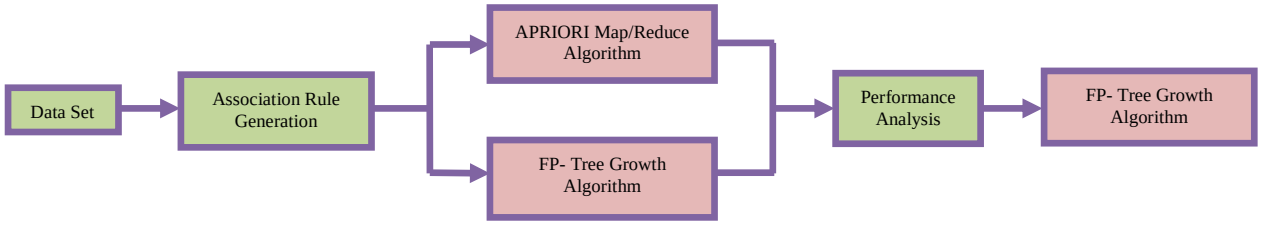
***Vijayarani S et al.,*** [18] performed a comparative analysis of traditional association rule mining algorithms for data streams. The algorithms are APRIORI, APRIORI PT (Prefix Tree) and APRIORI MR (Map/Reduce). Different sizes of windows and threshold values are used and it concluded that APRIORI MR has produced good results than other algorithms.

***Charu C. Aggarwal,*** [4] presented the detailed information about data streams. He discussed how to relate variant data mining technologies to data streams for supportive and unknown data extraction. He also discussed data stream clustering, data stream classification, association rule mining algorithm in data stream and frequent pattern mining.

***Kuldeep Malik et al.,*** [13] explained the FP-Growth algorithm and he proposed Enhanced FP-Growth Algorithm. He defined the Enhanced FP- Growth is working without prefix tree and any other complex data structure and he has proved that Enhanced FP-Growth has produced good results than FP-Growth.

## III. PROPOSED ALGORITHM

*The system architecture of this research work is represented in Figure 1.*




**F**ig 1. System Architecture

*A. Dataset*

The connect data set is used in this work. This dataset is available in http://fimi.ua.ac.be/data/connect.dat. It has 67,558 instances and 48 attributes. From this 1K, 2K and 5K instances are used in this work. In data streams, we assume that the nonstop arrival of data is partitioned into many windows with permanent size, i.e. $W_1$, $W_2$, $W_3$..... $W_n$. In this work, we have created five windows$W_1$, $W_2$, $W_3$, $W_4$, $W_5$ with the data set size of 1K, 2K and 5K.

*B. Association Rule Generation*

In order to generate association rules, two traditional data mining algorithms are used. They are,

- ✓ APriori MR – Apriori Map/Reduce Algorithm.
- ✓ FP- Tree - Frequent Pattern Tree Growth Algorithm.

*C. APRIORI MR Algorithm*

Apriori-Map/Reduce algorithm sprints on equivalent Map/Reduce framework. Candidate generation of Apriori Map/Reduce algorithm is prone ($C_{k+1}$) function is to eliminate the non-frequent item set $C_{k+1}$ by reducing the non-frequent item sets cannot be a subset of frequent item sets. Table 1 corresponds to the APRIORI MR algorithm [12].

Table 1. Pseudo Code for APRIORI Map/Reduce Algorithm

Step 1. Map transaction t in a data supply to all Map nodes
Step 2. Each Map node can handle m
Step 3. Now, can use Candidate Map $Cm_1$ = size of 1 is a frequent item set in the node m
Step 4. Reduce and compute candidate generation of $C_1$ and $L_1$ with all $Cm_1$
Step 5. $C_1$ = size one of frequent item sets;
Step 6. Calculate the Min_Support = Num/ total items;
Step 7. Size 1 of frequent item sets Min_Support is $L_1$
Step 8. Loop begins, For (k=1; $L_k$!= 0; k++) do
Step 9. Each mapped node m is represent by $L_k$ . Such as, $L_{mk}$
Step 10. Sort and remove the duplicate item sets
Step 11. Can use, $C_{m\,(k+1)}$ = $L_k$ join_sort $L_{mk;}$

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 3, Issue 9, September 2015**

Step 12. Reduce methods to use the APRIORI Property to compute the $C_{k+1}$ do
Step 13. Each map node m is increment the count of $L_{m\,(k+1)}$ candidates. That are supplied by transaction t
Step 14. End.
Step 15. Now, Can use reduce method to find the $L_{k+1}$ with $L_{m\,(k+1)}$ and Min_Support.
Step 16. Min_Support of frequent item set generated by size of k+1 is $L_{k+1}$.
Step 17. End
Step 18. Return $U_k L_k$;

*D. FP Tree Growth Algorithm*

FP- Tree Growth used the divide and conquers tactics for creating the frequent item sets. FP-growth is primarily used for mining frequent item sets without candidate generation [13]. It consist of one root tags as null, a set of item prefix sub trees as the children of the root, and a frequent item subtitle table. Every node in the item prefix sub tree consists of three fields: item-name, count and node link where item-name records which item the node characterized; count records the amount of transactions signified by the fraction of path reaching this node, the node links to the next node in the FP- tree. Each item in the subtitle table consists of two fields, item name and head of node link, which positions to the first node in the FP-tree transports the item name. The pseudo code of mining on the FP - tree is portrayed in Table 2.

Table 2. Pseudo Code for FP-Tree Growth Algorithm

| Window Size | Threshold | 1000 Ds | 2000 Ds | 5000 Ds | 10,000 Ds |
|---|---|---|---|---|---|
| | | Rules | | | |
| W1 | $\sigma = 25$, $C = 55$ | 18554 | 7473 | 7473 | 7473 |
| W2 | | 19588 | 7676 | 7667 | 7664 |
| W3 | | 7989 | 11112 | 7986 | 7982 |
| W4 | | 10154 | 10154 | 10143 | 10143 |
| W5 | | 9814 | 9814 | 9814 | 9810 |
| W1 | $\sigma = 45$, $C = 55$ | 3000 | 3090 | 3112 | 5553 |
| W2 | | 3003 | 3060 | 3045 | 3078 |
| W3 | | 3689 | 3652 | 3691 | 3634 |
| W4 | | 8465 | 8434 | 8490 | 9456 |
| W5 | | 5497 | 5493 | 5461 | 5449 |

Step 1. Input: constructed FP-tree
Step 2. Output: complete set of frequent patterns
Step 3. Method: Call FP-growth (FP-tree, null).
Step 4. Procedure:
Step 5. FP-growth (Tree, α)
Step 6. {
Step 7. if Tree contains a single path P then
Step 8. For each combination do generate patterns β ∪ α with
i. Support = min_sup of nodes in β.
Step 9. Else For each header $a_i$ in the header of Tree do {
Step 10. Generate pattern $\beta = a_i \cup \alpha$ with support = $a_i$.support;
Step 11. Construct β.s conditional pattern base and then β.s conditional
i. FP-tree Tree β
Step 12. If Tree β = null {}
Step 13. Then call FP-growth (Tree β, β)}
Step 14. }

## IV. SIMULATION RESULTS

The section describes the experimental results of **APRIORI MR** and **FP-Tree Growth algorithms**. The work is implemented on Tanagra tool. Totally, five windows $W_1$, $W_2$, $W_3$, $W_4$ and $W_5$ are used in this research work. There are

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2015

four different sizes of datasets; 1K, 2K, 5K and 10K is tested and their results are obtained. Different threshold values are applied for analyzing the results. The performance factors used in this analysis are number of rules generated and execution time.

Table 3.1 APRIORI MR Algorithm for Rule Generation

| Window Size | Threshold | 1000 Ds | 2000 Ds | 5000 Ds | 10,000 Ds |
|---|---|---|---|---|---|
| | | Rules | | | |
| W1 | $\sigma = 25$, $C = 55$ | 18554 | 7473 | 7473 | 7473 |
| W2 | | 19588 | 7676 | 7667 | 7664 |
| W3 | | 7989 | 11112 | 7986 | 7982 |
| W4 | | 10154 | 10154 | 10143 | 10143 |
| W5 | | 9814 | 9814 | 9814 | 9810 |
| W1 | $\sigma = 45$, $C = 55$ | 3000 | 3090 | 3112 | 5553 |
| W2 | | 3003 | 3060 | 3045 | 3078 |
| W3 | | 3689 | 3652 | 3691 | 3634 |
| W4 | | 8465 | 8434 | 8490 | 9456 |
| W5 | | 5497 | 5493 | 5461 | 5449 |

Table 3.1 depicted the number of rules generated using APRIORI MR Algorithm. Two different thresholds values $\sigma = 45$, $C = 55$ are used.

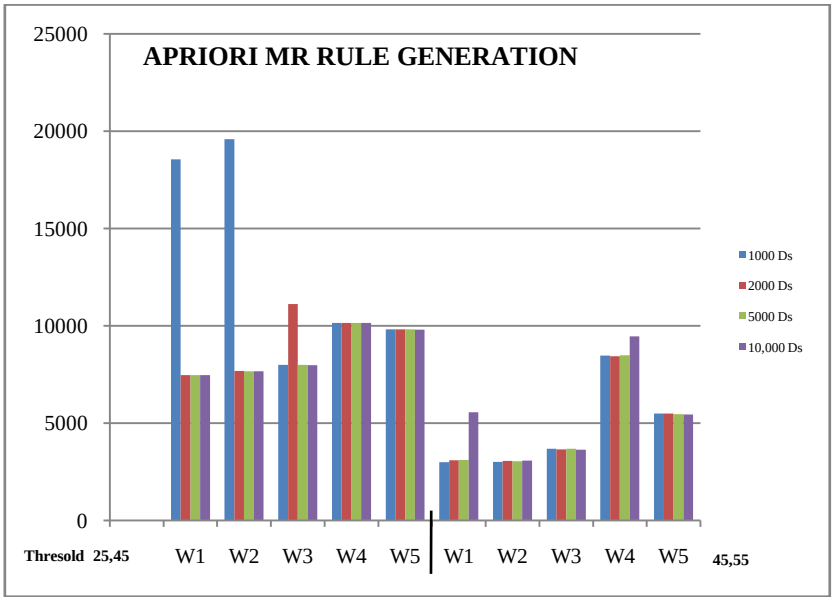



Fig 3.1 Association Rule generation – APRIORI MR Algorithm

Figure 3.1 represented the rule generated by the APRIORI MR algorithm. From this result, it is observed that the minimum threshold value has generated more number of rules.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2015

Table 3.2. Apriori MR Algorithm for Time Computation

| Window Size | Threshold | 1000 Ds | 2000 Ds | 5000 Ds | 10,000 Ds |
|---|---|---|---|---|---|
| | | Time(ms) | | | |
| W1 | $\sigma = 25$, $C = 55$ | 2170 | 1539 | 3439 | 6633 |
| W2 | | 2784 | 1595 | 3539 | 6831 |
| W3 | | 9594 | 3230 | 3722 | 7051 |
| W4 | | 1180 | 2053 | 4527 | 8742 |
| W5 | | 1171 | 2007 | 4442 | 8534 |
| W1 | $\sigma = 45$, $C = 55$ | 140 | 187 | 296 | 531 |
| W2 | | 156 | 203 | 265 | 577 |
| W3 | | 203 | 156 | 297 | 515 |
| W4 | | 281 | 281 | 609 | 850 |
| W5 | | 156 | 219 | 437 | 655 |

Table 3.2 shows the execution time required for APRIORI MR algorithm. Two different thresholds values $\sigma = 45$, $C = 55$ are used.

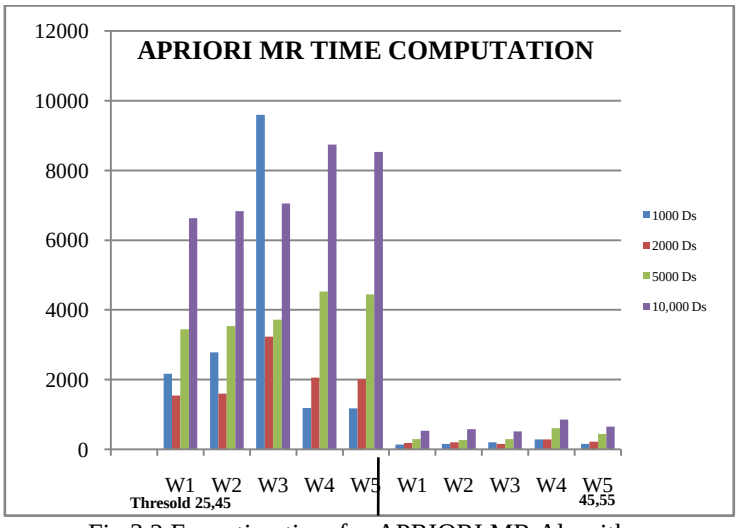



Fig 3.2 Execution time for APRIORI MR Algorithm

Figure 3.2 represented the execution time taken by the APRIORI MR algorithm.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2015

Table 3.3 FP- Tree Growth Algorithm for Rule Generation

| Window Size | Threshold | 1000 Ds | 2000 Ds | 5000 Ds | 10,000 Ds |
|---|---|---|---|---|---|
| | | Rules | | | |
| W1 | $\sigma = 25$, $C = 55$ | 65988 | 68345 | 83678 | 89712 |
| W2 | | 67250 | 67945 | 82444 | 89677 |
| W3 | | 65455 | 69896 | 84990 | 90012 |
| W4 | | 62390 | 65990 | 81266 | 86320 |
| W5 | | 61233 | 66435 | 85484 | 88531 |
| W1 | $\sigma = 45$, $C = 55$ | 41435 | 45412 | 42265 | 49710 |
| W2 | | 44289 | 47892 | 47811 | 50101 |
| W3 | | 49342 | 43672 | 48234 | 49875 |
| W4 | | 43861 | 47239 | 50420 | 51676 |
| W5 | | 50905 | 46021 | 51346 | 49945 |

Table 3.3 shows the number of rule generated using FP-Tree Growth algorithm. Two different thresholds values $\sigma = 45,\ C = 55$ are used.

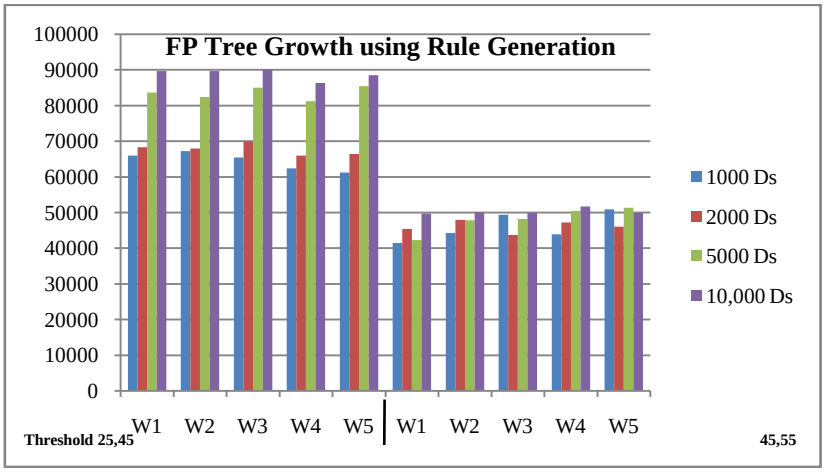



Fig 3.3 Association Rule Generation – FP Tree Growth Algorithm

Figure 3.3 illustrated the number of rule generated using FP-Tree Growth algorithm.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2015

Table 3.4 FP-Tree Growth Algorithm for Time Computation

| Window Size | Threshold | 1000 Ds | 2000 Ds | 5000 Ds | 10,000 Ds |
|---|---|---|---|---|---|
| | | Time (ms) | | | |
| W1 | $\sigma = 25$, $C = 55$ | 1448 | 1210 | 1310 | 1348 |
| W2 | | 1454 | 1221 | 1437 | 1340 |
| W3 | | 1455 | 1221 | 1441 | 1359 |
| W4 | | 1440 | 1446 | 1515 | 1567 |
| W5 | | 1337 | 1474 | 1243 | 1440 |
| W1 | $\sigma = 45$, $C = 55$ | 905 | 917 | 912 | 949 |
| W2 | | 915 | 917 | 918 | 944 |
| W3 | | 917 | 914 | 934 | 955 |
| W4 | | 921 | 919 | 946 | 967 |
| W5 | | 916 | 921 | 938 | 973 |

Table 3.4 depicted the execution time required for FP-Tree Growth algorithm. Two different thresholds values $\sigma = 45$, $C = 55$ are used.

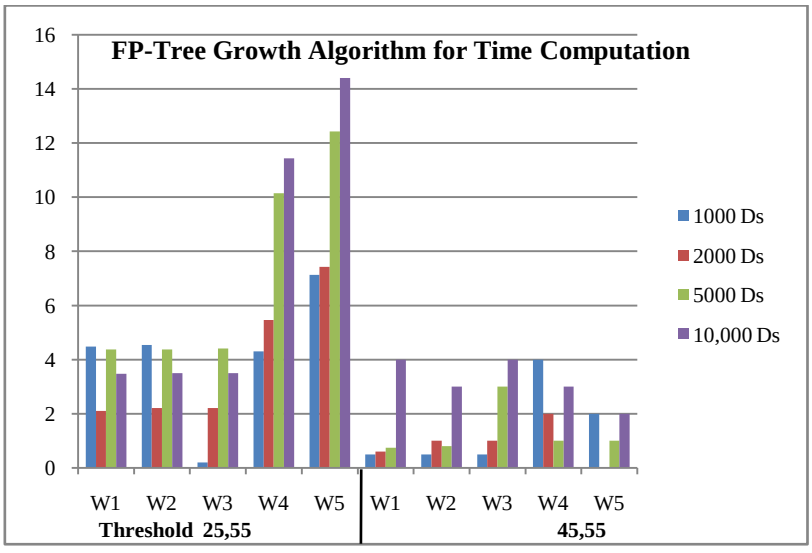



Fig 3.4 Execution Time for FP Tree Growth Algorithm

Figure 3.4 shows the execution time for FP-Tree Growth algorithm.

## V. CONCLUSION AND FUTURE WORK

This research work has compared the traditional association rule mining algorithms for generating association rules in data streams. Traditional association rule mining algorithm was experimented very less number of rules generation and execution time is very high. So, we compared with a new proposed algorithm like as FP Growth and APRIORI MR Algorithm. From the experimental results, it is observed that the performance of FP-Tree Growth algorithm is efficient than APRIORI MR Algorithm. In future work, new algorithms are to be developed in order to reduce the number of rules and execution time.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 3, Issue 9, September 2015**

## REFERENCES

1. Aggarwal, C (2003). A Framework for Diagnosing Changes in Evolving Data Streams. ACM SIGMOD Conference.
2. Agrawal, R. and Srikant, R. Fast Algorithms for Mining Association rules. Proc. 20th VLDB conference, Santiago, Chile, 1994.
3. A. Savasere, E. Omiecinski, and S.B. Navathe, "An efficient algorithm for mining association rules in large databases," Intl. Conf. on Very Large Databases, pp. 432–444, 1995.
4. Charu C. Aggarwal "Data Stream Models and algorithms"-Data stream book 2009, Springer.
5. Christian Hidber. Online Association rule mining. SIGMOD '99 Philadelphia, PA. ACM 1-58113-084-8/99/05, 1999.
6. Charanjeet Kaur, Association Rule Mining using Apriori Algorithm: A Survey ISSN: 2278 – 1323 International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 6, June 2013.
7. "Data mining techniques "by Arun k Pujari.
8. "Data Streams: An Overview and Scientific Applications" Charu C. Aggarwal.
9. "Data Mining: Introductory and Advanced Topics" Margaret H. Dunham.
10. Frequent item set mining data set repository, http:// fimi.cshelsinki.fi/data/ \
11. Han, J., Kamber, M.: "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2006.
12. Jongwook Woo, " Apriori-Map/Reduce Algorithm " International Organization for Scientific Research, 2012.
13. Kuldeep Malik, Neeraj Raheja and Puneet Garg, "Enhanced FP- Growth Algorithm" IJCEM International Journal of Computational Engineering & Management, Vol. 12, April 2011 ISSN (Online): 2230-7893, www.IJCEM.org IJCEM www.ijcem.org.
14. "Mining frequent patterns across multiple data streams" Jing Guo, Peng Zhang, Jianlong Tan and Li Guo, 2011.
15. Nan Jiang and Le Gruenwald, "Research Issues in Data Stream Association Rule Mining"- SIGMOD Record, Vol. 35, No. 1, Mar. 2006.
16. Rakesh Agrawal, Ramakrishnan Srikant; Fast Algorithms for Mining Association Rules; Int'l Conf. On Very Large Databases; September 1994.
17. S. Vijayarani, P. Sathya, " Mining Frequent Item Sets over Data Streams using Éclat Algorithm", International Conference on Research Trends in Computer Technologies (ICRTCT - 2013) Proceedings published in International Journal of Computer Applications® (IJCA) (0975 – 8887) 27.
18. S. Vijayarani, R. Prasannalakshmi, "Association Rule Generation in Data Streams Using Apriori Algorithms", International Journal of Engineering Research And Management (IJERM) ISSN: 2349- 2058, Volume-01, Issue-09, December 2014.

## BIOGRAPHY

**Dr. S. Vijayarani,** MCA., M.Phil, Ph.D is working as Assistant Professor in the School of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of research interest are data mining, privacy and security issues and data streams. She has published papers in the international journals and presented research papers in international and national conferences.

**Mrs. R. Prasannalakshmi,** M.C.A has completed M. Phil in Computer Science in the School of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of interest are data mining and data streams.