ISSN: 2348 - 2117

International Journal of Engineering Technology and Computer Research (IJETCR)
Available Online at www.ijetcr.org
Volume 2; Issue 6; Page No. 113-119

An Efficient Algorithm for Mining Association Rules using Different Parameters for frequent item sets

Priyanka[1], Er. Vinod Kumar Sharma[2]

[1]M.Tech Research Scholar, Department of Computer & Science Engineering, Guru Kashi University, Talwandi Sabo, Punjab, India

Priyankakansal8@gmail.com

[2]Assistant Professor, Department of Computer & Science Engineering, Guru Kashi University, Talwandi Sabo, Punjab, India

Vinod_sharma85@rediffmail.com

ABSTRACT

As the advancement of information technology increases, the amount of data also increases. So, to handle the huge amount of data, Data mining plays an important role. Data mining is a process of extraction of valuable and unknown information from the large databases. There are various techniques and tasks but we will discuss about association rule mining and apriori algorithm. Association rule mining is a descriptive technique that is used to find out the interesting patterns among the data items stored in the database. Apriori algorithm mines the frequent item sets and association rule learning over the transactional databases. In this research work, we have taken different datasets and applied Apriori Algorithm. After that we have analyzed the results by taking different parameters.

Key words: Data mining, Association Rule Mining, Apriori Algorithm, Frequent Item set etc.

INTRODUCTION:

Data mining is also called knowledge discovery in databases (KDD). It is commonly defined as the process of discovering useful patterns or knowledge from data sources, e.g., databases, texts, images, the Web, etc.

1.1 Association Rule Mining:

Association Rule Mining is an important component of data mining .Association rules are an important class of methods of finding patterns in data. Association mining has been used in many applications domains. One of the most known is the business field where discovering of purchase patterns or association between products is very useful for decision making and effective marketing. It aims to extract interesting correlations, frequent patterns, associations among sets of items in the transaction databases. It is to find out association rules that satisfy the predefined minimum support and confidence from a given database. The problem is usually decomposed into two sub problems. One is to find those item sets whose occurrences exceed a predefined threshold in the database; those item sets are called frequent or large item sets. The second problem is to generate association rules from those large item sets with the constraint of minimal confidence. A frequent item set is an item set whose number of occurrences is above a minimum support threshold. An item set of length k is called k-item set and a frequent item set of length k as k-frequent item set. An association rule is considered strong if it satisfies a minimum support threshold and minimum confidence threshold. [4]

Basic Concepts of Association Rules: - Basic objective of finding association rules is to find all co-occurrence relationship called associations. Since it was first introduced in 1993 by R. Agarwal et.al, it has attracted a great deal of attention. Many efficient algorithms, extensions and applications have been reported. The classic application of association rule mining is market basket data analysis, which aims to discover how items purchased by customers in a supermarket (or store) are associated.

1.2 Apriori Algorithm

Apriori is a classic algorithm for learning association rules and mining frequent item sets for Boolean association rules in data mining. It was firstly proposed by R.Agrawal and R.Srikant in 1994 The Apriori algorithm is also called the level-wise algorithm (breadth-first search). To find association rules, the algorithm divided into two steps. The first step is to identify all the frequent item sets, each of these item sets will occur at least as frequently as a predetermined min_sup. Secondly, the strong association rule is generated from frequent item sets. If the confidence of the frequent item sets is not less than the

min_conf given by the user. Apriori algorithm firstly scans the database to find out the number of each item. Apriori algorithm is actually a layer-by-layer iterative searching algorithm, where k-item set is used to explore the (k+ 1) item set. Firstly, the algorithm identifies the collection of frequent 1-Itemset, denoted by L_1. Collection of Frequent 2-Itemset "L_2" is computed from L_1 which is used to find L_2, and so on, until no more frequent k-dimensional item sets can be found. We use the same method to acquire L_k (K≥2) until it cannot findL_{k-1}. Finally, getting the rules from large set of data items

Apriori Algorithm [2]

Find frequent item-sets using an iterative level-wise approach based on candidate generation.

Input: *D*, a database of transactions;

Min- sup, the minimum support count threshold.

Output: *L*, frequent item-sets in *D*.

Method:

(1) L_1 = find_ frequent_ 1-item sets (D);

(2) For (*k* = 2; L_k-1 ≠φ; *k*++) do begin

(3) *Ck* = apriori _gen (L_k-1);

(4) For each transaction *t* ε D do begin// scan *D* for counts

(5) C_t= subset (C_k, *t*); // get the subsets of *t* that are candidates

(6) For each candidate *c ε Ct do*

(7) c.count++;

(8) End

(9) L_k = {*c* ε C_k|*c. count ≥ min _sup*}

(10) end

(11) return *L* = $U_k L_K$;

1.3 Basic Terms

1.3.1 Item set- A collection of one or more items.

For Example-{Milk, Bread, Diaper}

1.3.2 k- Item set- An item set that contains k items.

1.3.3 Frequent Item set- An item set whose support is greater than or equal to a min sup threshold.

1.3.4 Support count (σ) - Frequency of occurrence of an item set.

For Example- σ ({Milk, Bread, Diaper}) = 2

1.3.5 Support-The support of an item-set is the fraction of the rows of the database that contain all of the items in the item-set. Support indicates the frequencies of the occurring patterns. Sometimes it is called frequency. Support is simply a probability that a randomly chosen transaction t contains both item-sets A and B.

$$\text{Min. Supp.} = \frac{\text{No. of transactions containing both A \& B}}{\text{Total no. of transactions}}$$

1.3.6 Confidence-Confidence denotes the strength of implication in the rule. Sometimes it is called Accuracy. Confidence is simply a probability that an item-set B is purchased in a randomly Chosen transaction t given that the item-set A is purchased.

$$\text{MinConf} = \frac{\text{No. of transactions containing both A \& B}}{\text{Transactions containing only A}}$$

2. Literature Survey

A lot of research has been done in the field of association rule mining and apriori algorithm.

A.H.S et.al (2013) This paper introduces a Apriori algorithm, a classical rule mining algorithm finds its application in areas of data mining, finding association between attributes and in prediction systems. Performance of the redesigned algorithm is evaluated and is compared with the traditional Apriori algorithm. To increase the efficiency of the Apriori algorithm and reduce the time complexity of the proposed algorithm into O (n). [2]

B. Al-Maqeleh et.al (2013) In this paper, A problem with such a process is that the solution of interesting patterns has to be performed only on frequent item sets. An efficient algorithm is proposed to integrate confidence measure during the process of mining frequent item sets, may substantially improve the performance of association rules mining by reducing the search space. The experimental results show the effectiveness of the proposed algorithm in reducing the number of discovered rules comparing with the Apriori algorithm. [3]

D. Ping et.al (2010) This paper proposed the new algorithm to reduce the storage space, improves the efficiency and accuracy of the algorithm based on the drawbacks of the traditional algorithms existed "item sets generation bottleneck" problem, and are very time consuming. An proposed algorithm associating which is based on the user interest item-sets, reduce unnecessary item-sets, is that the support given by date scanning can't be less than the frequent item sets of the minimum support given by the user. In proposed algorithm can save memory and reduce the time of comparing the candidate sets. So the speed of generating frequent item sets will be faster. [5]

M. Patel et.al (2013) in this paper, they have been proposed of many algorithms to mine association rule that uses support and confidence as constraint. We proposed a method based on support value that increase the performance of Apriori algorithm and minimizes the number of candidate generated and removed candidate at checkpoint which is infrequent which interns reduces storage and time required to calculate support of candidate. [11]

Y. Zhou et.al (2010) In this paper, an improved Apriori algorithm is proposed and spending a lot of time to

produce the candidate item-sets, scanning the database. The improved algorithm is consist of three parts to reducing the number of operation, while generating candidate frequent item-sets and association rules the database. [19]

3. OBJECTIVE: Objectives are as follows.

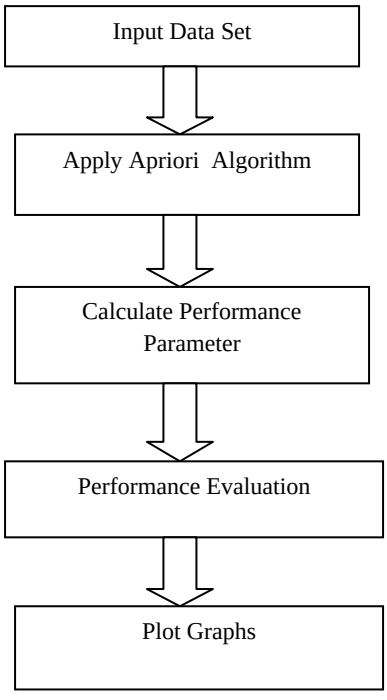
- To optimize the process of finding patterns which should be efficient, scalable, and can detect important patterns which can be used in various ways.
- To different databases for fetching the Rules and Intelligence.
- An Effective Implementation of Rule Based Mining: Apriori Algorithm approach shall be implemented for multiple applications.
- The parameter values used can be analyzed for efficient results.

4. RESEARCH METHODOLOGY

4.1 Introduction of WEKA

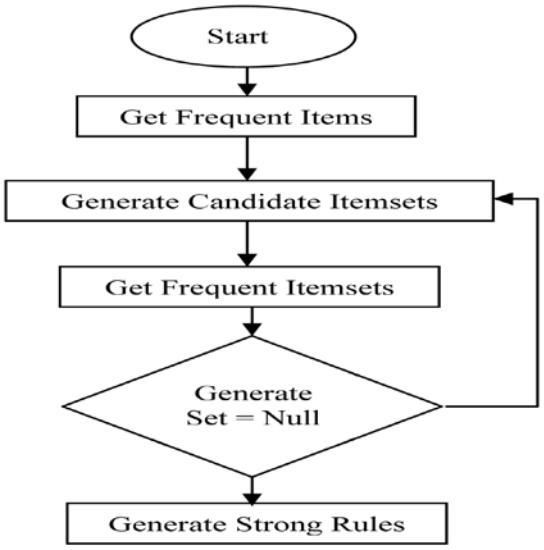
WEKA is open source software under the GNU General Public License. It was developed at the University of Waikato in New Zealand. "WEKA" stands for Waikato Environment for Knowledge Analysis. The system is freely available at http://www.cs.wailato.ac.nz/ml/weka. The system is written using object oriented language java. There are several different levels at which WEKA can be used. WEKA supports many data mining tasks such as data re-processing, classification, clustering, regression and feature selection.

4.2 Steps for Implementing Apriori algorithm



4.2.1 Input data sets: Input data is an integral part of data mining applications. The data used in my experiment is either real world data obtained from UCI data repository and widely accepted dataset available in Weka toolkit, during evaluation multiple data sizes were used, each dataset is described by the data type being used, the types of attributes, the number of instances stored within the dataset, also the table demonstrates that all the selected data sets are used for the association rule mining technique. These datasets were chosen because they have different characteristics and have addressed different areas.

4.2.2 Apply Apriori Algorithm



4.2.3 Performance Parameter

4.2.3.1 Support: supp(*X*) of an item set *X* is defined as the proportion of transactions in the data set which contain the item set.

Supp (X) = no. of transactions which contain the item set X / total no. of transactions

4.2.3.2 Confidence: The rule X → Y holds with confidence conf in T if conf % of transactions that contain X also contain Y. The confidence of a rule is defined:

Conf (X →Y) = P(Y| X) = Supp(X ∪Y) /Supp(X) = P(X and Y)/ P(X)

4.2.3.3 Lift: It is the probability of the observed support to that expected if X and Y were independent. The lift of a rule is defined as:

Lift (X →Y) = Supp(X ∪ Y) / Supp(X) * Supp(Y) = P(X and Y) / P(X) P(Y)

4.2.3.4 Leverage: It measures the difference of X and Y appearing together in the data set and what would be expected if X and Y where statistically dependent

Lev (X →*Y*) = *P*(*X* and *Y*) – (*P*(*X*) *P*(*Y*))

4.2.3.5 Conviction: It was introduced by Brin et al., 1997. Conviction takes the value 1 when A and B has no items

in common and it is undefined when the rule A => B always holds. The conviction of the rule X=>Y can be interpreted as the ratio of the expected frequency that X occurs without Y (that is to say, the frequency that the rule makes an incorrect prediction) if X and Y were independent divided by the observed frequency of incorrect predictions. The conviction of a rule is defined as:

Conv (X→Y) = 1 - Supp(Y) / 1 – conf (X → Y) = P(X) P (not Y) / P(X and not Y)

5. EXPERIMENTAL RESULTS:

Data Set	No. of Instances	No.of Attributes
Weather	5	14
Voting	17	435
Supermarket	217	4627
Contact-Lenses	5	24

Weather Data Set

Metric Type – Confidence

Table 5.1: Default values of Confidence Parameter

Min-Sup	Min-Conf	No. Of Cycles	No. Of Rules
0.15	0.9	17	10

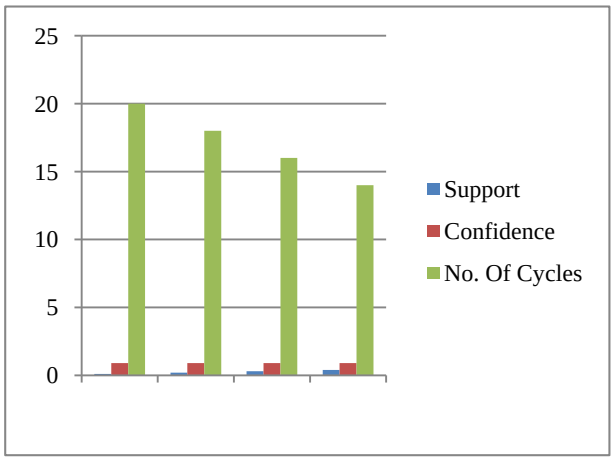
Support values – one by one increase

Confidence values- same

Table 5.2: Calculated Values of Confidence Parameter

Support	Confidence	No. Of Cycles	No. Of Rules
0.1	0.9	18	100
0.2	0.9	16	8
0.3	0.9	14	3
0.4	0.9	12	0

Graph No. 5.1: Comparison of Support, Confidence & No. Of Cycles



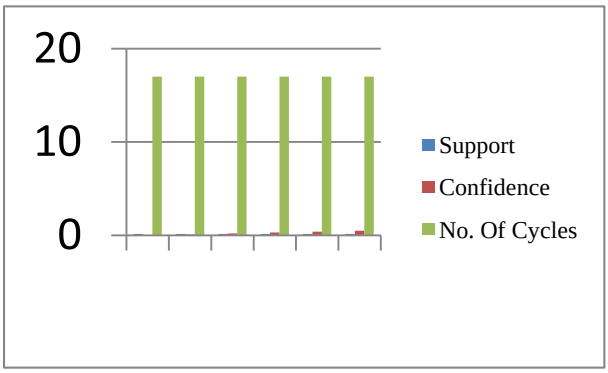
Support values – same

Confidence values- one by one increase

Table 5.3: Calculated Values of Confidence Parameter

Support	Confidence	No. Of Cycles	No. Of Rules
0.15	0	17	100
0.15	0.1	17	100
0.15	0.2	17	100
0.15	0.3	17	100
0.15	0.4	17	100
0.15	0.5	17	100
0.15	0.6	17	100
0.1	0.7	18	100

Graph No. 5.2: Comparison of Support, Confidence & No. of Cycles



1. Metric Type-Lift

Table 5.4: Default values of Lift Parameter

Min-support	Lift	No. of Cycles	No. of Rules
0.3	1.1	14	10

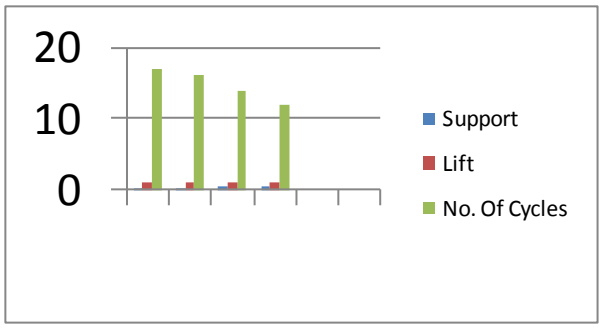
Support Values- one by one increase

Lift vales –same

Table 5.5: Calculated Values of Lift Parameter

Support	Lift	No. of Cycles	No. Of Rules
0.15	1.1	17	100
0.2	1.1	16	54
0.3	1.1	14	18
0.4	1.1	12	4

Graph No. 5.3: Comparison of Support, Lift & No. of Cycles



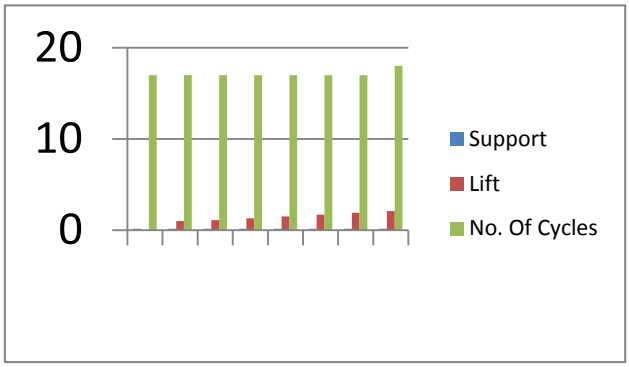
Support values- same

Lift values – one by one increase

Table 5.6: Calculated Values of Lift Parameter

Support	Lift	No. of Cycles	No. Of Rules
0.15	0	17	100
0.15	1	17	100
0.15	1.1	17	100
0.15	1.3	17	100
0.15	1.5	17	100
0.15	1.7	17	100
0.15	1.9	17	100
0.15	2.1	18	100

Graph No. 5.4: Comparison of Support, Lift & No. of Cycles



2. Metric Type- Leverage

Table 5.7: Default values of Leverage Parameter

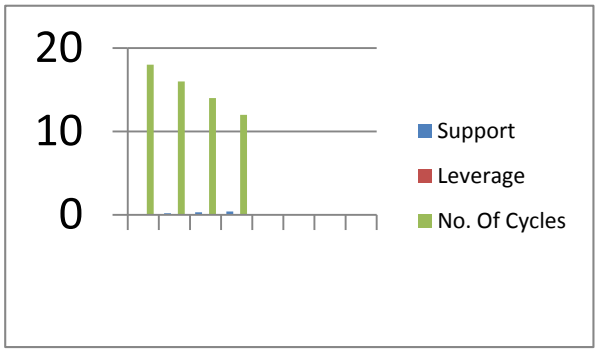
Min-sup	Leverage	No. Of Cycles	No. Of Rules
0.3	0.1	14	10

Support values – one by one increase
Leverage values – same

Table 5.8: Calculated Values of Leverage Parameter

Support	Leverage	No. Of Cycles	No. Of Rules
0.1	0.1	18	34
0.2	0.1	16	18
0.3	0.1	14	10
0.4	0.1	12	2

Graph No. 5.5: Comparison of Support, Leverage & No. of Cycles

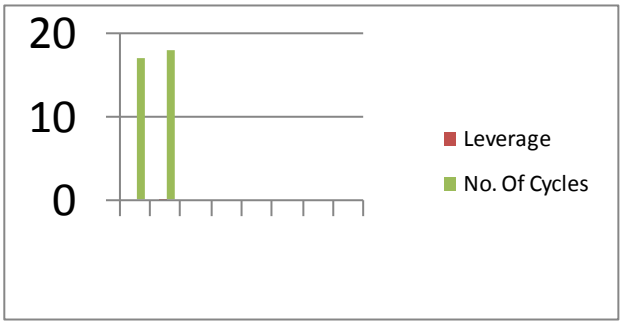


Support Values – Same
Leverage Values – one by one increase

Table 5.9: Calculated Values of Leverage Parameter

Support	Leverage	No. Of Cycles	No. Of Rules
0.15	0	17	100
0.1	0.1	18	34

Graph No. 5.6: Comparison of Support, Leverage & No. of Cycles



3. Metric Type- Conviction

Table 5.10: Default values of Conviction Parameter

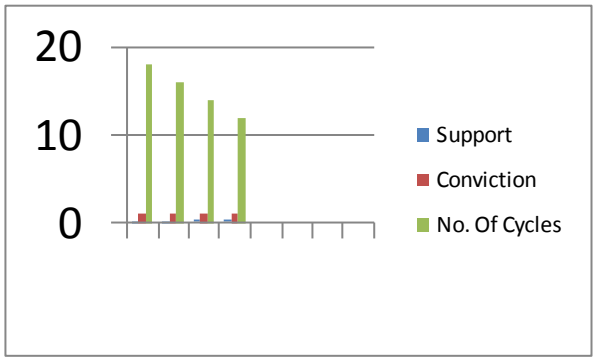
Min- sup	tion	No. Of cycles	No. Of Rules
0.25	1.1	10	15

Support Values- one by one increase
Conviction Values- Same

Table 5.11: Calculated Values of Conviction Parameter

Support	Conviction	No. Of Cycles	No. Of Rules
0.1	1.1	18	51
0.2	1.1	16	18
0.3	1.1	14	8
0.4	1.1	12	2

Graph No. 5.7: Comparison of Support, Conviction & No. of Cycles

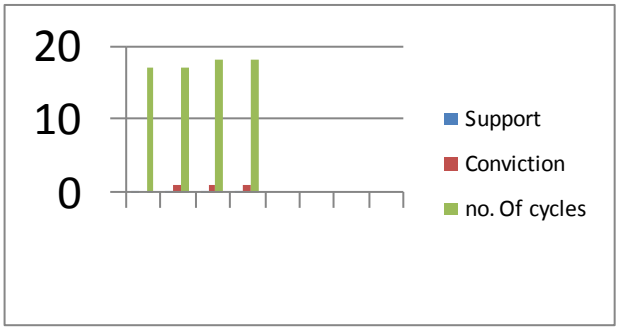


Support Values- Same
Conviction Values- one by one increase

Table 5.12: Calculated Values of Conviction Parameter

Support	Conviction	No. of Cycles	No. Of Rules
0.15	0	17	100
0.15	1	17	100
0.1	1.1	18	51
0.1	1.2	18	34

Graph No. 5.8: Comparison of Support, Conviction & No. Of Cycles



6. CONCLUSION & FUTURE WORK

An efficient way for discovering the frequent item set can be very useful in various data mining problems, such as discovery of association rules. The most widely used algorithm is the Apriori Algorithm. In this research work, the frequent item sets are analyzed the different data sets. To enhance the business good service can be easily provide to the customer by knowing customer and selling purchases. In the future work, the work presented in this thesis can be applied to various other fields and further improvement can be done in case of sensitive data bases.

ACKNOWLEDGMENT

I am really grateful to my guide and my friends for this paper. It would not be possible to complete my paper without their help and valuable support. Last but not the least, I thank to my parents and family members.

REFERENCES

1. Lekhal, Dr. C V Srikrishna and Dr. Viji Vinod , "Utility Of Association Rule Mining: A Case Study Using Weka Tool", International Conference On Emerging Trends In VLSI, Embedded System, Nano Electronics And Telecommunication System (ICEVENT), pp-1-6, IEEE, 2013.
2. Anand H.S. and Vinodchandra S.S., "Applying Correlation Threshold on Apriori Algorithm", International Conference On Emerging Trends In Computing, Communication And Nanotechnology (ICE-CCN), pp- 432-435, IEEE, 2013.
3. Basheer Mohamad Al-Maqaleh And Saleem Khalid Shaab , " An Efficient Algorithm For Mining Association Rules Using Confident Frequent Item Sets", Third International Conference On Advanced Computing And Communication Technologies (ACCT), 2013 ,pp- 90-94 , IEEE , 2013.
4. Chunzhang , Dezan Xie, Dezan Xie, Honghuili, And Fengliu, "The Improvement Of Apriori Algorithm And Its Application In Fault Analysis Of Crh Emu", International Conference on Service OPErations, Logistics, and Informatics (SOLI), pp-543 – 547, IEEE, 2011.
5. Du Ping, And Gao Yongping, "A New Improvement Of Apriori Algorithm For Mining Association Rules", International Conference on Computer Application and System Modeling (ICCASM), Vol.-2, pp- V2-529 - V2-532, IEEE, 2010.
6. D. N. Goswami, Anshu Chaturvedi and C.S. Raghuvanshi , "Frequent Pattern Mining Using Record Filter Approach", International Journal of Computer Science Issues, Vol. 7, Issue 4, No 7, pp 38-43, 2010.
7. Huiying Wang and Xiangwei Liu, "The Research of Improved Association Rules Mining Apriori Algorithm", Eighth International Conference on Fuzzy Systems and Knowledge Discovery(FSKD), vol-2, pp-961-964, IEEE, 2011.
8. J Han , "Data Mining Concepts and Techniques", Second Edition Morgan Kaufmann Publisher 2006.
9. Jingyao Hu, "The Analysis on Apriori Algorithm Based on Interest Measure", International Conference on Control Engineering and Communication Technology (ICCEC T), pp-1010-1012, 2012.
10. Libing Wu, Kui Gong, Yanking He, Xiaohua Ge, and Jianqun Cui , "A Study of Improving Apriori Algorithm", 2nd International Conference on Intelligent Systems and Applications(ISA), pp- 1-4, IEEE, 2010.
11. Mihir R. Patel, Dipti P. Rana and Rupa G. Mehta, "FApriori: A Modified Apriori Algorithm Based on Checkpoint", International Conference on Information Systems and Computer Networks (ISCON), pp-50-53, IEEE, 2013.
12. Mrs. R. Sumithra, and Dr (Mrs). Sujni Paul, "Using distributed apriori association rule and classical apriori mining algorithms for grid based knowledge discovery", Second International conference on Computing, Communication and Networking Technologies (ICCCNT), pp-1-5, IEEE, 2010.
13. Qiang Yang and Yanhong Hu, "Application of Improved Apriori Algorithm on Educational Information", Fifth International Conference on Genetic and Evolutionary Computing (ICGEC), pp-330-332, IEEE, 2011.
14. Rui Chang and Zhiyi Liu, "An Improved Apriori Algorithm", International Conference on Electronics and Optoelectronics (ICEOE), Vol-1, pp-V1-476-V1-478, IEEE, 2011.
15. Sanjeev Rao, and Priyanka Gupta, "Implementing Improved Algorithm Over APRIORI Data Mining Association Rule Algorithm", International journal of

compter science and technology (IJCST), Vol. 3, Issue 1, pp-489-493, 2012.

16. Shuo Yang , "Research and Application of Improved Apriori Algorithm to Electronic Commerce", 11th International Symposium on Distributed Computing and Applications to Business, Engineering & Science (DCABES), pp-227-231, IEEE, 2012.
17. Sunil Joshi and Dr. R. C. Jain, "A Dynamic Approach for Frequent Pattern Mining Using Transposition of Database", Second International Conference on Communication Software and Networks (ICCSN'10), pp-498-501, IEEE, 2010.
18. WANG Lu-Feng, "Association Rule Mining Algorithm Study and Improvement", 2nd International Conference on Software Technology and Engineering(ICSTE), vol.-2,PP-v2-362-v2- 364, IEEE, 2010.
19. Yubo Jia, Guanghu Xia, Hongdan Fan, Qian Zhang, and Xu Li, " An Improved Apriori Algorithm Based on Association Analysis", Third International Conference on Networking and Distributed Computing (ICNDC), pp-208-211, IEEE, 2012.
20. Yanfei Zhou, Wanggen Wan, Junwei Liu, and Long Cai, "Mining Association Rules Based on an Improved Apriori Algorithm", International conference on Audio Language and Image Processing,(ICALIP), pp-414-418, IEEE, 2010.
21. Zhiyi Liu and Rui Chang, "Study on Efficient Algorithm of Frequent Item-set Mining", International Conference on Electronics and Optoelectronics (ICEOE), vol-1, pp-V1-222-V1-225, IEEE, 2011.
22. Zhuang Chen and Shibang Cai, Qiulin Song and Chonglai Zhu, " An Improved Apriori Algorithm Based on Pruning Optimization and Transaction Reduction", 2nd International Conference on Artifical Intelligence ,Management Science and Electronic Commerce `(AIMSEC), pp-1908-1911, IEEE 2011.