Study on Apriori Algorithm and its Application in Grocery Store

Pragya Agarwal
Department of CSE
ASET,Amity University
Sector-125,Noida,UP,India

Madan Lal Yadav
Department of CSE
ASET,Amity University
Sector-125,Noida,UP,India

Nupur Anand
Department of CSE
ASET,Amity University
Sector-125,Noida,UP,India

ABSTRACT

Among the many mining algorithms of associations rules, Apriori Algorithm is a classical algorithm that has caused the most discussions; it can effectively carry out the mining association rules. With large database, the process of mining association rules is time consuming. The efficiency becomes crucial factor. Moreover, Apriori algorithm is improved by reducing the number of scanning data base. The proposed algorithm reduces the storage room, improves the competency of performance with negligible error of the algorithm. Finally, the improved Apriori algorithm can solve the problem of traditional Apriori algorithm. After analyzing the Apriori algorithm, this algorithm is incapable due to it scans the database several times. Based on the planning of getting to database once, a new recoverd algorithm formed on the Apriori is put forward in this paper. Experiments show that it can mostly adds computation competency, i.e. minimize the calculating time and space. This algorithm has been broadly used for Grocery rooms in customer consumer knowledge mining.

1. INTRODUCTION

Data Mining is the core process of KNOWLEDGE DISCOVERY IN DATABASE [28]. It is the process of extraction of useful patterns from the large database[28]. To analyze the large amounts of collected information, the area of Knowledge Discovery in Databases (KDD) provides techniques which extract interesting patterns in a reasonable amount of time. Therefore, KDD employs methods at the cross point of machine knowledge, statistics and database technology. Data mining is the application of competent algorithms to detect the desired patterns contained within the given data.

Data mining has recently attracted considerable attention from database practitioners and researchers because it has been applied to many fields such as market planning, financial forecasts and decision support. Many algorithms have been proposed to obtain useful and invaluable information from huge databases. One of the most important algorithms is mining association rules, which was first introduced. Association rule mining has many important applications in our life. An association rule is of the form X => Y. And each rule has two measurements: support and confidence. The association rule mining difficulty is to get rules that satisfy user-specified minimum support and minimum confidence.

Among the many mining algorithms of associations rules, Apriori Algorithm is a classical algorithm that has caused the most discussions; it can effectively carry out the mining association rules. With large database, the process of mining association rules is time consuming. The efficiency becomes crucial factor. Moreover, Apriori algorithm is improved by reducing the number of scanning data base. The algorithm reduces the storage room, improves the competency of performance and correctness of the algorithm. Data mining is the extraction of hidden descriptive or predictive information from large databases. It has been described as “the task of discovering interesting patterns from large amount of data where the data can be stored in databases, data warehouses or other information store houses.

1.1 Association rule and Apriori algorithm

An “association rule” is a rule like "If a customer buys pizza base and butter, he/she often buys ketchup too." It expresses an association between (sets of) items, which may be products of a market or a mail-order department, special equipment options of a cabe, optional services offered by telecommunication factories etc.

An association rule states that if we pick a customer at random and find out that he/she selected some items (bought some products, chose some options etc.), we can be assured, indicate the quantity by a percentage, that he/she also selected some other items (bought some other products, chose some other options etc.). Of course, we do not want just any association rules, we want "good" rules, rules that are "expressive" and "reliable". The standard measures to assess association rules are the “support” and the “confidence” of a rule, both of which are computed from the help of some item sets.

Apriori algorithm (Agrawal et al. 1993), is the most fundamental and important algorithm for mining frequent things. Apriori is used to find all frequent things in a given database DB. The keynote of Apriori algorithm is to produce multiple passes over the database. It employs an repetitive approach called as a breadth-first search (level-wise search) through the search room, where k-things are used to explore (k+1)-things.

1.2 Need of improved Apriori algorithm

There are some drawbacks or limitations of classical Apriori algorithm.

Limitations of Apriori algorithm-

- Needs many repititions of the data.
- Uses a same minimum support threshold.
- Difficulties to get rarely occuring happenings.
- With large database, the procedure of mining association rules is time taking.
- The competency becomes crucial factor.

2. METHODOLOGY

Data mining is the core process of KNOWLEDGE DISCOVERY IN DATABASE [28]. It is the process of extraction of useful patterns from the large database [28]. To analyze the large amount of collected information, the area of Knowledge Discovery in Database (KDD) provides techniques which extract interesting patterns in a reasonable amount of time [28]. Data mining is the application of efficient algorithms to detect the desired patterns contained within the given data. Data mining is the extraction of hidden descriptive or predictive information from large databases [27].

2.1 Association rule mining

Association rules mining are one of the major techniques of data mining. The purpose of association analysis is to figure out the hidden association and some useful rules of data base, and uses these rules to speculate and judge the unknown matter from the already known information [25]. Association rule mining has many important applications in our life.

2.1.1 Association rule

- An Association rule is one of the form X=>Y. And each rule has two basic needs:
  Support and confidence.
- Things that occur often together can be associated to each other.
- These together occurring things form a **Frequent itemset**.
- Conclusions based on the frequent itemsets make **Association rules.**

2.1.2 Apriori algorithm

Apriori algorithm is a fundamental algorithm mining association rule [19]. It contains two processes:
• Detect all frequent itemsets by scanning DB.
• Form strong association rules in the frequent itemsets.
Process one needs to scan DB several times, which consumes a lot of time and space. As a result, what needs to be improved is the mining competency of frequent group of things in DB [19].

Apriori Algorithim is an significant algorithm for mining frequent itemsets for boolean association rules.

- Apriori algorithm is formed by **AGRAWAL** and **SRIKANT** in 1994.
- It is the most fundamental and important algorithm for mining frequent itemsets.
- Apriori is used to detect all frequent itemsets in a provided database DB.
- The keynote of Apriori algorithm is to form multiple passes over the database.
- It employs an repetitive approach called as a **breadth-first search (level-wise search).**

2.1.3 Key concepts

• **Frequent itemsets:** The itemsets which has minimum help (denoted by Li for ith-itemsets).
• **Apriori Property:** Any subgroup of frequent things must be frequent.
• **Join Operation:** To detect Lk, a group of candidate k-group of things is developed by adding Lk-1with itself.

3. HOW APRIORI WORKS?

Find all frequent itemsets.

- o Get frequent things:
  - Things whose occurrence in database is more than or equal to the minimum help threshold.
- o Get frequent itemsets:
  - Develop candidates from frequent things.
  - Prune the results to detect the frequent itemsets.

Develop strong association rules from frequent itemsets.

- o Rules which satisfy the minimum support and minimum confidence threshold.

High level design

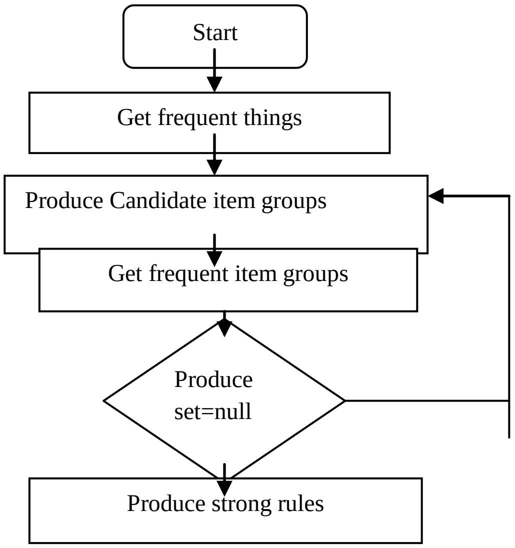


Fig 1: High Level Design of Apriori Algorithm

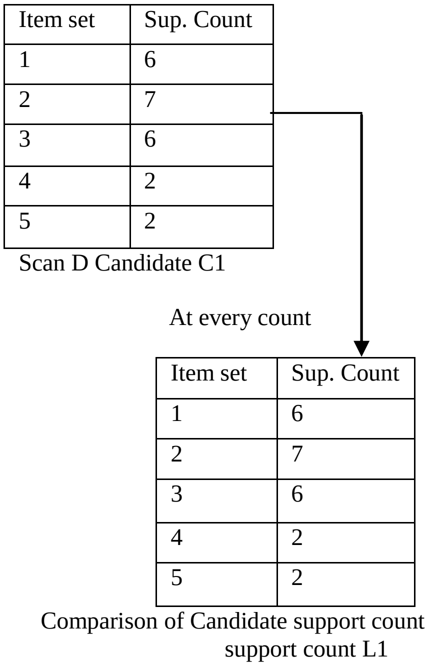
Related products are placed together in such a manner that customers can logically find items he/she might buy which increases the customer satisfaction and the profit. Customers are segmented and association rules are separately generated to satisfy their specific needs in a cost effective manner using some special promotions for the common groups. From the results it is shown that the market basket analysis using K-Apriori algorithm for Grocery Stores improves its overall revenue.

4. TAKE AN EXAMPLE

Table 1. consisting of Transaction Id and List of items

TID	List of the items
T100	1,2,5
T100	2,4
T100	2,3
T100	1,2,4
T100	1,3
T100	2,3
T100	1,3
T100	1,2,3,5
T100	1,2,3

Step 1: Generating 1-itemgroup Frequent Pattern



Item set	Sup. Count
1	6
2	7
3	6
4	2
5	2

Scan D Candidate C1

At every count

Item set	Sup. Count
1	6
2	7
3	6
4	2
5	2

Comparison of Candidate support count with minimum support count L1

Fig 2: Generating 1-itemgroup Frequent Pattern

Consider a database, D , consisting of 9 transactions.
•Suppose min. support count required is 2 (i.e. min_sup = 2/9 = 22 %)
•Assume minimum confidence required is 70%.
•We have to first find out the frequent itemgroup using Apriori algorithm.
•Then, Association rules will be produced using minimum support & minimum confidence.

The set of frequent 1-itemgroups, L1, consists of the candidate 1-itemsets satisfying minimum support.
In the first loop of the algorithm, each item is a member of the set of candidate.

Step 2: Generating 2-itemgroup Frequent Pattern

Scan D for count of each candidate, we get C2

Item set	Sup. Count
[1,2]	4
[1,3]	4
[1,4]	1
[1,5]	2
[2,3]	4
[2,4]	2
[2,5]	2
[3,4]	0
[3,5]	1
[4,5]	0

C2

Item set	Sup. Count
[1,2]	4
[1,3]	4
[1,5]	2
[2,3]	4
[2,4]	2
[2,5]	2

Comparison of Candidate support count with minimum support count L2

Fig 3: Generating 2-itemgroup Frequent Pattern

•To discover the set of frequent 2-itemgroups, L2, the algorithm uses L1 *Join* L1to generate a candidate set of 2-itemsets, C2.
•Next work is,transactions in D are scanned and the support count for each candidate itemset in C2is accumulated (as shown in the middle table).
•The set of frequent 2-itemgroups, L2, is then determined, consisting of those candidate 2-itemsets in C2having minimum support.

•We haven't used Apriori Property till now.

Step 3: Generating 3-itemgroup Frequent Pattern

Item set
[1,2,3]
[1,2,5]

Item set	Sup. Count
[1,2,3]	2
[1,2,5]	2

Scan D for count of each candidate, we get C3

Item set	Sup. Count
[1,2,3]	2
[1,2,5]	2

Comparison of Candidate support count with minimum support count L3

Fig 4: Generating 3-itemgroup Frequent Pattern

•The generation of the set of candidate 3-itemgroups, C3, involves use of the Apriori Property.

•In order to get C3, we compute L2*Join*L2.

•C3= L2 *Join*L2 = [[1, 2, 3], [1, 2, 5], [1, 3, 5], [2, 3, 4], [2, 3, 5], [2, 4, 5]].

•Now, Join step is complete and Prune step will be used to reduce the size of C3. Prune step is used to avoid heavy computation due to large Ck.
Based on the Apriori property that all subsets of a frequent item group must also be frequent so that we can determine that four latter candidates cannot possibly be frequent. How ?

•For example , lets take [1, 2, 3].The 2-item subsets of it are [1, 2], [1, 3] & [2, 3]. Since all 2-item subsets of [1, 2, 3] are members of L2, We will keep [1, 2, 3] in C3.

•Lets take another example of [2, 3, 5]which shows how the pruning is performed. The 2-item subsets are [2, 3], [2, 5] & [3,5].

•BUT, [3, 5] is not a member of L2and hence it is not frequent violating Apriori Property. Thus we will have to remove [2, 3, 5] from C3.

•Therefore, C3= [[1, 2, 3], [1, 2, 5]] after checking for all members of result of Join operation for Pruning.

•Finally the transactions in D are scanned in order to determine L3, consisting of those candidates 3-itemsets in C3 having minimum support.

Step 4: Generating 4-itemgroup Frequent Pattern

•The algorithm uses L3 *Join*L3to generate a candidate set of 4-itemsets, C4. Although the join results in [[1, 2, 3, 5]], this itemset is pruned since its subset [[2, 3, 5]]is not frequent.

•Thus, C4= φ, and algorithm terminates and we have found all of the frequent things. This completes our Apriori Algorithm.

5. RESULTS

We can see the results of these transactions by the following screenshots.
Fig 5: shows the Home page.

Fig 6: Open the file of stored transaction database.
Fig 7: Shows the desired output.

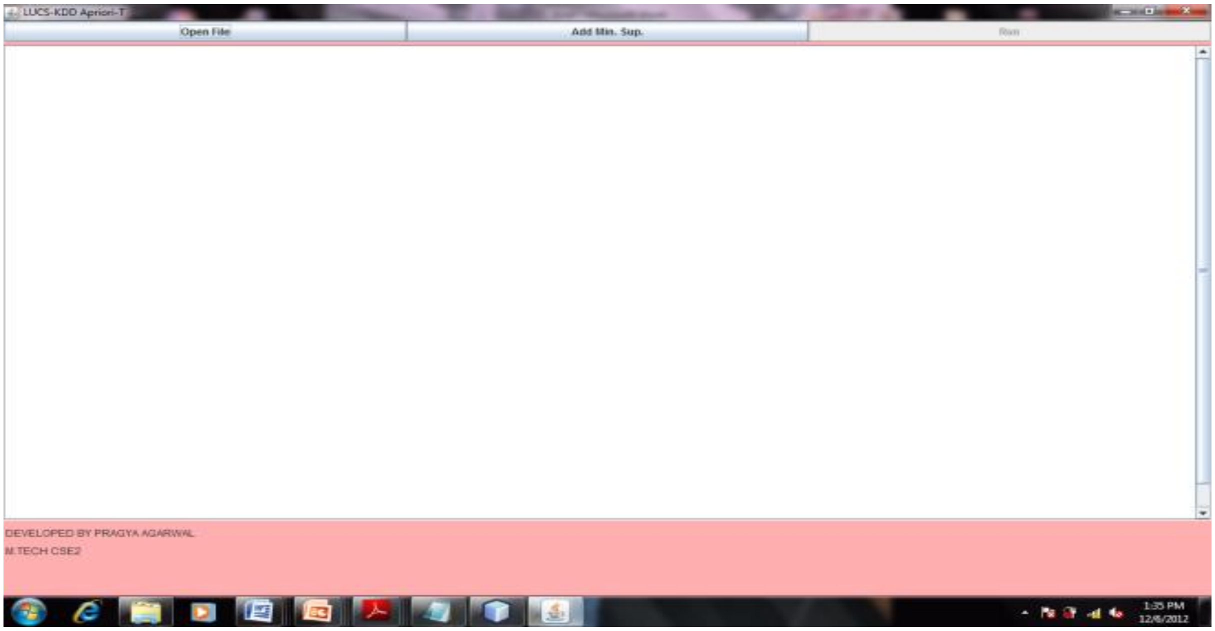
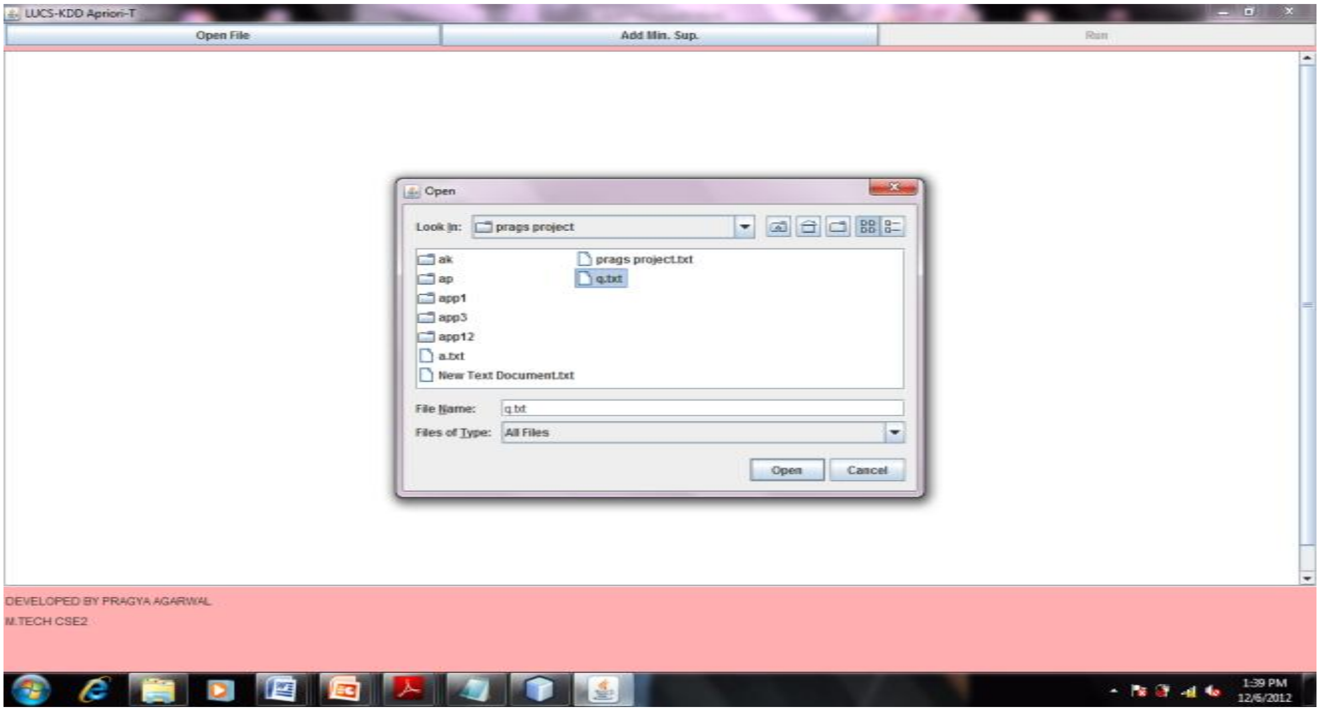


Fig 5: This is the home page, in which we open the file of containing transaction



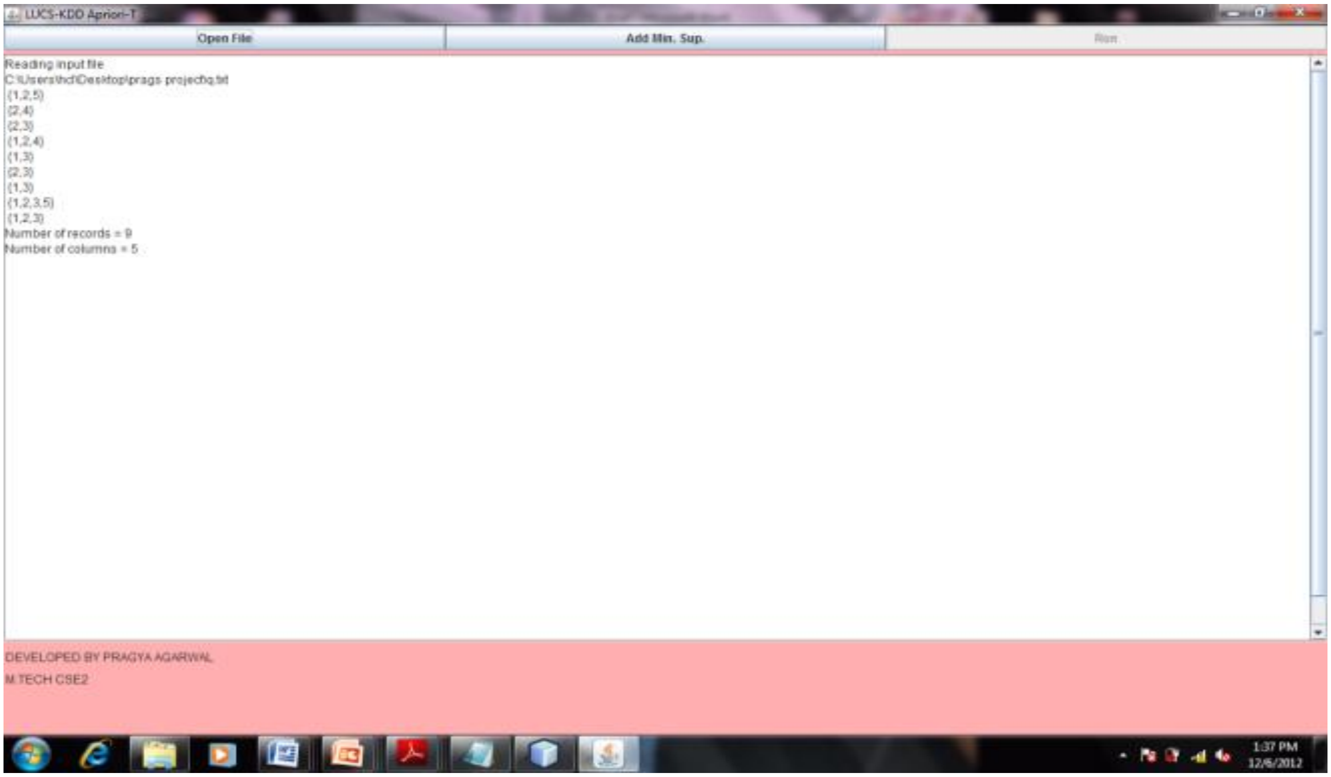
Look In: prags project

File Name: q.txt

Files of Type: All Files

DEVELOPED BY PRAGYA AGARWAL
M.TECH CSE2

Fig 6: Open the file of stored transaction database



```
Reading input file
C:\Users\hcl\Desktop\prags project\q.txt
 {1,2,5}
 {2,4}
 {2,3}
 {1,2,4}
 {1,3}
 {2,3}
 {1,3}
 {1,2,3,5}
 {1,2,3}
Number of records = 9
Number of columns = 5
```

DEVELOPED BY PRAGYA AGARWAL
M.TECH CSE2

Fig 7: It shows the desired output

6. CONCLUSION

The conclusion of the project is to overcome the disadvantages of traditional Apriori algorithm and efficiently mine association rules without generating candidate itemsets. This project has studied the classic association rules mining algorithm and discussed shortcomings of the Apriori algorithm. It takes the grocery store's consumer knowledge mining for example to compare.

In this project we presented the use of an ARM (Association rule mining) driven application is to manage retail businesses that provide retailers with reports regarding prediction of product sales trends and customer behavior. Our goal of project is to find a new scheme for finding the rules out of the transactional dataset which outperforms in terms of running time, number of database scan, memory consumption and the interestingness of the rules over the classical APRIORI Algorithm.
We have proposed mining algorithm for incremental generation of large itemsets. Frequent itemsets discovered depends on value of parameters like support and number of transactions read at a time. Thus execution time of the algorithm depends on transactional datagroups, minimum support value.

7. FUTURE WORK

To overcome these limitations of traditional Apriori algorithm, a frequent pattern-tree (FP-Growth) structure is proposed. Association Rules form an very applied data mining approach. Association Rules are obtained from frequent itemsets. The Apriori algorithm is an competent algorithm for finding all frequent itemsets. A great advantage of FP-tree is that overlapping itemsets share the same prefix path. So the information of the data set is mostly compressed. It only needs to scan the data set twice and no candidate itemsets are required.

7.1 Obtaining frequent patterns without candidate generation

Shorten a large database into a compact, Frequent-Pattern tree (FP-tree)structure.
–Highly compressed, but complete for frequent pattern mining.
–avoid costly database scans.
Develop a competent, FP-tree-based frequent pattern mining method.
–A divide-and-conquer methodology: distill mining tasks into smaller ones.
–Evade candidate generation: sub-database test only!

7.2 Advantages of FP-growth algorithm

The main advantages of FP-Growth algorithm is,
• Uses compact data structure

• Eliminates repeated database scan

FP-growth is faster than other association mining algorithms. The algorithm reduces the total number of candidate item sets by producing a compressed version of the database in terms of an FP-tree. The FP-tree stores relevant information and allows for the efficient discovery of frequent item sets.

8. REFERENCES

[1] Piatetsky Shapiro and G Discovery analysis and presentation of strong rules ; in G Piatetsky-Shapiro & W J Frawley; eds; Knowledge Discovery in Databases; AAAI/MIT Press; Cambridge; MA; 1991.

[2] Agrawal R and Imielinski T and Swami A N Mining association rules between sets of items in large databases; In Proceedings ACM SIGMOD International Conference on Management of Data Vol 22; No 2; of SIGMOD Record; Washington; pp 207–216; 1993.

[3] R.Agrawal; T.Imielinski; and A Swami; Mining association rules between sets of items in large databases; In Proc of the ACM SIGMOD Conference on Management of data; Washington; D C; pp 207-216; May 1993.

[4] Agrawal R ; Srikant R ; Fast algorithms for mining association rules [A]; In; Proceedings of the 20th Internationan Conference on Very Large Databases[C]; Santiago ; Chile ; 1994 487-499.

[5] Agrawal R; Srikant R ; Fast algorithms for mining association rules ; In Proceedings 20th International Conference on Very Large Data Bases (VLDB' 94); pp 487-499; 1994.

[6] Agrawal R and Srikant R and Mining sequential patterns In Proceeding of the 11th International Conference on Data Engineering ; Taipei ; Taiwan; pp 3 -14; 1995.

[7] Agrawal R et al Fast Discovery of Association Rules [M]; In; Advances in Knowledge Discovery and Data Mining; Menlo Park Calif; AAAI/MIT Press; 1996 307-328.

[8] Srikant R ; Agrawal R ; Mining quantitative association rules in large relational tables ; In Proceeding of Association for Computing Machinery- Special Interest Group on Management of Data (ACM SIGMOD) ;pp 1-12;1996.

[9] Fayyad U. M; Piatetsky-Shapiro G; Piatetsky-Shapiro P X ; From data mining to knowledge discovery; an Overview Advances in Knowledge Discovery and Data Mining AAAI Press/MIT Press pp 1-36 1996.

[10] Fayyad U. M; Piatetsky Shapiro G; Smyth; P; From data mining to knowledge discovery in databases AI Magazine Vol 17; No 3; pp 37 54 ; 1996.

[11] Zaki; M. J; Parthasarathy; S; Ogihara; M; Li; W; New algorithms for fast discovery of association rules ; In 3rd Intl; Conf; on Knowledge Discovery and Data Mining I997.

[12] Brin S; Motwani R and Silverstein C and Beyond market baskets and Generalizing association rules to correlations; Data Mining and Knowledge Discovery Journal; Vol 2; pp 39-68; 1998.

[13] Herbert A Edelstein; Introduction to Data Mining and Knowledge Discovery; 3rd Edition; pp 22-26; Oct 1999.

[14] Han J ; Pei J ; Yin Y; Mining frequent patterns without candidate generations; In Proceeding of the ACM SIGMOD; pp 1–12; 2000.

[15] Han J; Jian P; Yiwen Y Mining frequent pattens without candidate generation[C]; In; Proceedings of the 2000 ACM SIGMOD International Conference Management of Data 2000;

[16] Zaki; M.J; SPADE: An Efficient Algorithm for Mining Frequent Sequences; In Machine Learning; Kluwer Academic Publishers. Manufactured in The Netherlands, 42, pp. 31–60, 2001.

[17] Jiawei Han ; Micheline Kamber.Data mining; concepts and techniques[M; 2001.

[18] Pujari A. K.; Data mining techniques; Universities Press (India) Private Limited; 2001.

[19] Lin Lu; Pei-qi Liu **"**Study On An Improved Apriori Algorithm And Its Application In Supermarket" Xi'an 710055; China 2001**.**

[20] Tan P.-N; Steinbach M; Kumar V; Introduction to data mining ; Addison Wesley; 2006.

[21] Han J ; Kamber M; Data mining concepts and techniques ; Elsevier Inc; Second Edition; San Francisco; 2006.

[22] Cheng J; Ke Y; Ng W; Effective elimination of redundant association rules; Data Mining and Knowledge Discovery Journal; Vol 16; pp 221–249; 2008.

[23] Tian Lan, Runtong Zhang and Hong Dai; A New Frame of Knowledge Discovery; in Proc 1st International Workshop on Knowledge Discovery and Data Mining; WKDD 2008; pp 607 – 611; Jan 2008.

[24] Bo Wu; Defu Zhang; Qihua Lan; Jiemin Zheng **"**An Efficient Frequent Patterns Mining Algorithm based on Apriori Algorithm and the FP-tree Structure" Xiamen 361005, China 2008**.**

[25] Liu Jing; Lu Yongquan; Wang Jintao; Gao Pengdong; Qiu Chu; Ji Haipeng; Li Nan; Yu Wenhua "An Improved Apriori Algorithm for Early Warning of Equipment Failue" BeiJing, China 2009 IEEE.

[26] Kumar B S; Rukmani K.V; Implementation of Web Usage Mining Using APRIORI and FP Growth Algorithms; in Int. J of Advanced Networking and Applications Vol 01; Issue; 06; pp 400-404; 2010.

[27] Shilpa; Sunita Parashar "Performance Analysis of Apriori Algorithm with Progressive Approach for Mining Data" October 2011**.**

[28] Sanjeev Rao and Priyanka Gupta and Implementing Improved Algorithm Over APRIORI Data Mining Association Rule Algorithm" 2012**.**