Enhancement in Apriori Algorithm to Improve Performance

A Dissertation submitted

By

Harkamal Kaur

(11304803)

To

Department of Computer Science Engineering

In partial fulfillment of the Requirement for the

Award of the Degree of

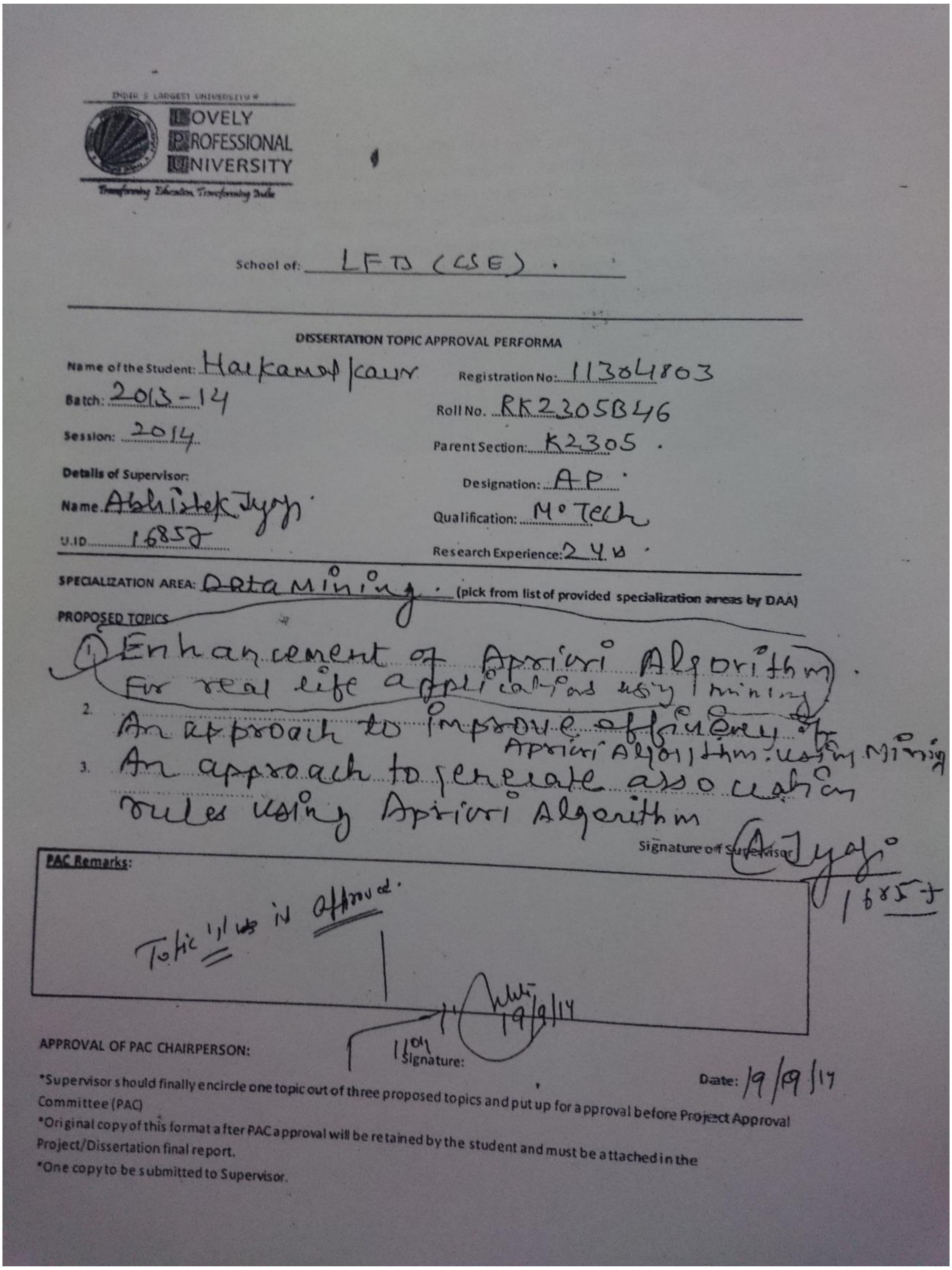
Master in Technology in Computer Science Engineering

Under the guidance of

Mr. Abhishek Tyagi

(April 2015)

PAC FORM



INDIA'S LARGEST UNIVERSITY*

LOVELY PROFESSIONAL UNIVERSITY

Transforming Education, Transforming India

School of: LFTS (CSE).

DISSERTATION TOPIC APPROVAL PERFORMA

Name of the Student: Harkamal Kaur	Registration No: 11304803
Batch: 2013-14	Roll No. RK2305B46
Session: 2014	Parent Section: K2305.
Details of Supervisor:	Designation: A.P.
Name. Abhishek Tyagi	Qualification: M.Tech
U.ID 16857	Research Experience: 2.4 yr.

SPECIALIZATION AREA: Data Mining. (pick from list of provided specialization areas by DAA)

PROPOSED TOPICS

1. Enhancement of Apriori Algorithm For real life applications using mining.
2. An approach to improve efficiency of Apriori Algorithm using Mining
3. An approach to generate association rules using Apriori Algorithm

Signature of Supervisor A. Tyagi 16857

PAC Remarks:

Topic 1,1 is in approved.

19/9/14

APPROVAL OF PAC CHAIRPERSON: Signature: Date: 19/9/14

*Supervisor should finally encircle one topic out of three proposed topics and put up for approval before Project Approval Committee (PAC)

*Original copy of this format after PAC approval will be retained by the student and must be attached in the Project/Dissertation final report.

*One copy to be submitted to Supervisor.

ABSTRACT

Data mining is one of the essentially used and interesting research areas. Mining association rule is one of the important research techniques in data mining field. Many algorithms for mining association rules are proposed on the basis of Apriori algorithm and improving the algorithm strategy but most of these algorithms not concentrate on the structure of database. The performance of the classical apriori algorithm decreases because the algorithm requires multiple scans and when database size is very large, this takes much time. The proposed technique includes transposition of database with further enhancement in this particular transposition technique. This approach reduces the total scans over the database and then time consumed to generate the frequent item sets is also reduced. The proposed idea is implemented in MATLAB which is widely used in all areas of research universities, and also in the industry. Experimental results shows that improved apriori algorithm reduce the time consumed and total transactions in comparison with original apriori algorithm.

CERTIFICATE

This is to certify that she has completed M. Tech dissertation proposal titled 'Enhancement in Apriori Algorithm to Improve Performance' under my guidance and supervision. To the best of my knowledge, the present work is the result of her original investigation and study. No part of the dissertation proposal has ever been submitted for any degree or diploma.

The Report is fit for the submission and the partial fulfillment of the conditions for the award of M. Tech Computer Science & Engineering.

Date: _______________ Signature of Advisor

Name:

UID:

ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of any task would be incomplete without mentioning the name of those people who made it possible, because success results, not only from hard work, but also from stead fast determination, dedication and above all adepts advises. I would like to express my special gratitude to my mentor Mr. Abhishek Tyagi for his guidance and support. I would like to thank him for encouraging my research work and also for his suggestions.

I would also like to appreciate the guidance given by the Department of Computer Science thanks for their valuable advice.

Harkamal kaur

(11304803)

DECLARATION

I hereby declare that dissertation proposal entitled “Enhancement in Apriori Algorithm to Improve Performance” submitted for the M. Tech. degree is entirely my original work and all ideas and references have been duly acknowledged. It does not contain any work for the award of any other degree or diploma.

Date: ………   Investigator

Reg. No. 11304803

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 INTRODUCTION TO DATA MINING

Data Mining is known as the process of analyzing data to extract interesting patterns and knowledge. Data mining is used for analysis purpose to analyze different type of data by using available data mining tools. This information is currently used for wide range of applications like customer retention, education system, production control, healthcare, market basket analysis, manufacturing engineering, scientific discovery and decision making etc [5].

Data mining is studied for different databases like object-relational databases, relational database, data ware houses and multimedia databases etc.

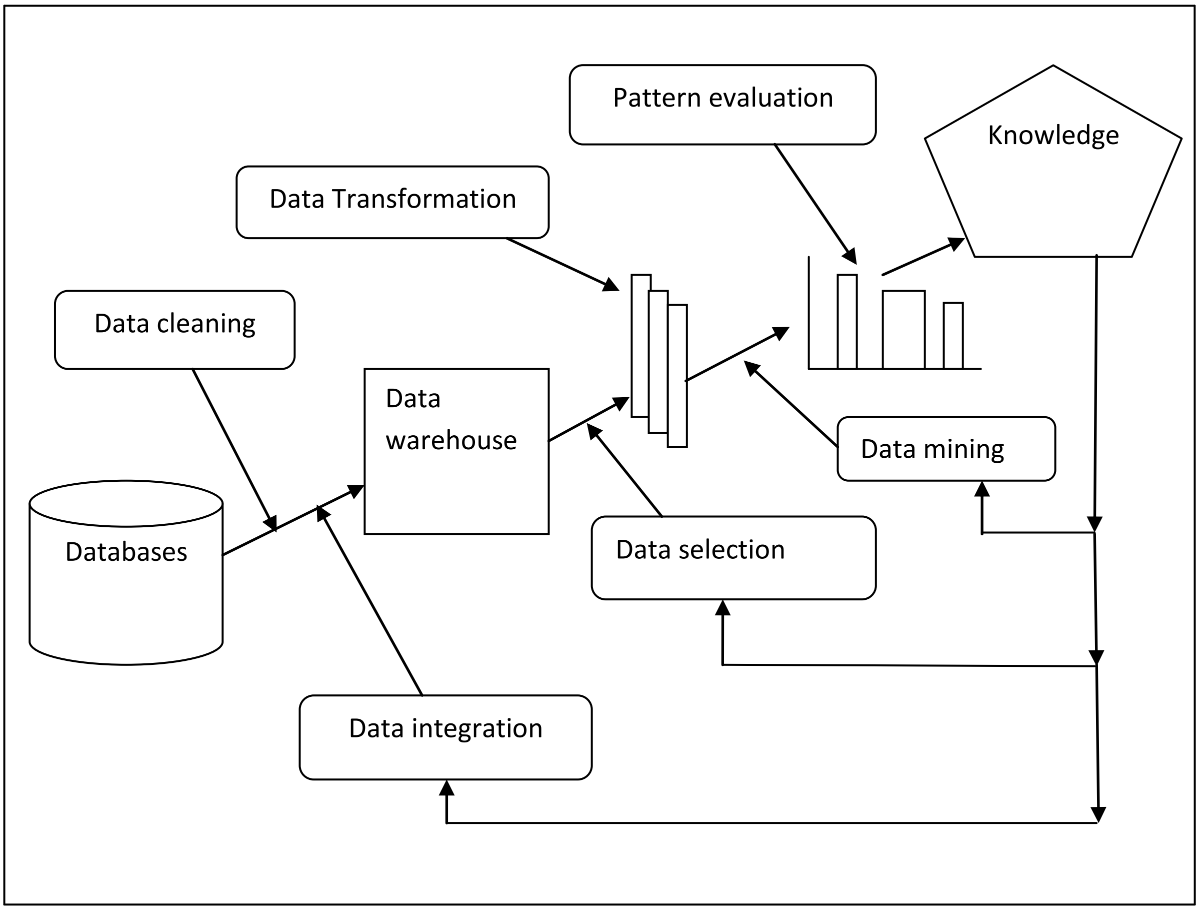


Fig. 1.1 Data Mining Process

Data mining is playing a vital role in many applications like market-basket analysis, etc. Frequent item sets have significant role in data mining which is used to find out the correlations between the fields of database. Association rule is based on discovering frequent item sets and frequently used by retail stores. Mining data in other words, named as Discovery of new knowledge in Databases which further moves to the nontrivial extraction of indirect, new and much required information from data in databases [11].

The role of data mining (KDD) is very important in many of the field such as the analysis of market basket, classification, etc. In data mining, the most important role presented by frequent item set which is used to find out the correlation between the various types of the field that is display in the database. Another name of the data mining is KDD (Knowledge discovery from the database), discovery of frequent item set is done by association rule. The concept of association rules are also used in retail sector for accomplishment of various purposes.

1.2 KDD PROCESS

The Discovery of knowledge in Databases process includes steps to gain unique knowledge. Steps are:

1. Data which is not relevant and contains noise is cleaned. This step contains clearing the data and known as data cleaning.
2. At data integration step, heterogeneous data is combined with the different data sources.
3. In the selection step, applicability of analyzed data is taken into consideration. Data selection is done after the analysis process.
4. Under the transform step, changes in the data are occurs with respect to various mining techniques.
5. Data Mining is used to find the required and unique patterns using many available techniques.
6. Within severely unique patterns, which includes acquaintance are recognized. This step involves, evaluating the required patterns. Thus, pattern evaluation is done under this step, which is used for further extraction of knowledge.
7. In this last step the exposed results, which includes knowledge are represented.

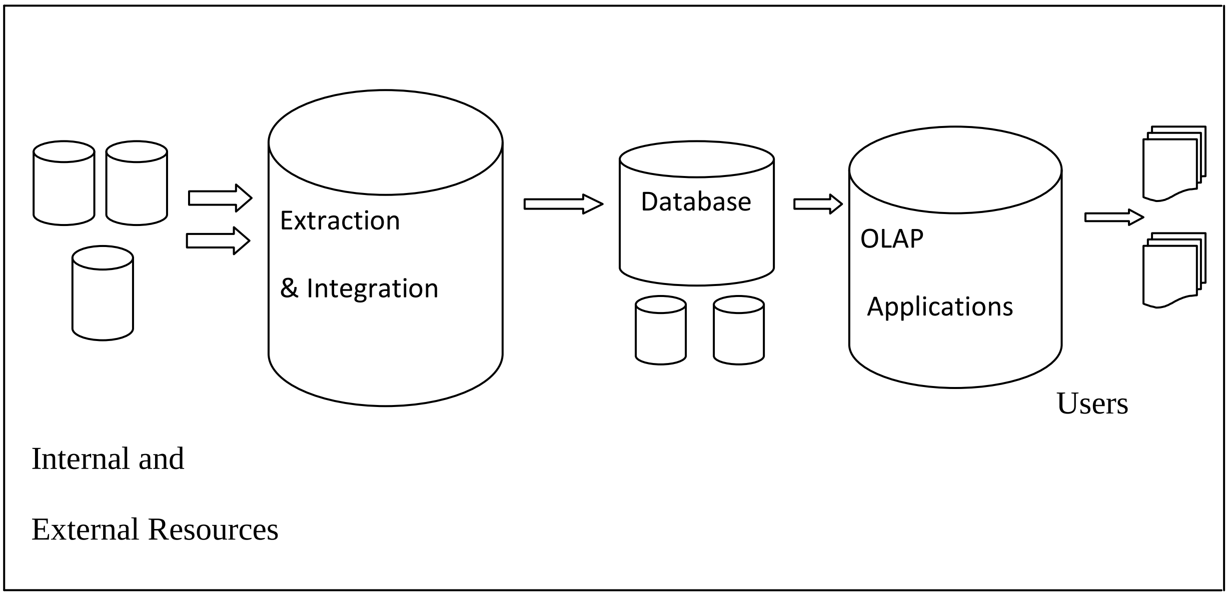


Fig. 1.2 Data Cleaning Process

Components involved in KDD process:

1. **External and Internal Resources:**

   First of all raw data is collected from number of resources either internal resources or external resources.

2. **Extraction and Integration:**

   In this step, first it extracts data from different resources and converts it into original format. Data cleaning is a process in which missing values are filled, it smooth noisy data and remove inconsistencies. In data integration, integration of multiple database, data cubes and files takes place.

3. **Database:**

   After that data is stored in database and data mart.

4. **OLAP Application:**

   OLAP can be used as for discovery in data mining for previously discerned relationship between data items.

1.3 DATA MINING ISSUES

Some of the issues in data mining are explained as:

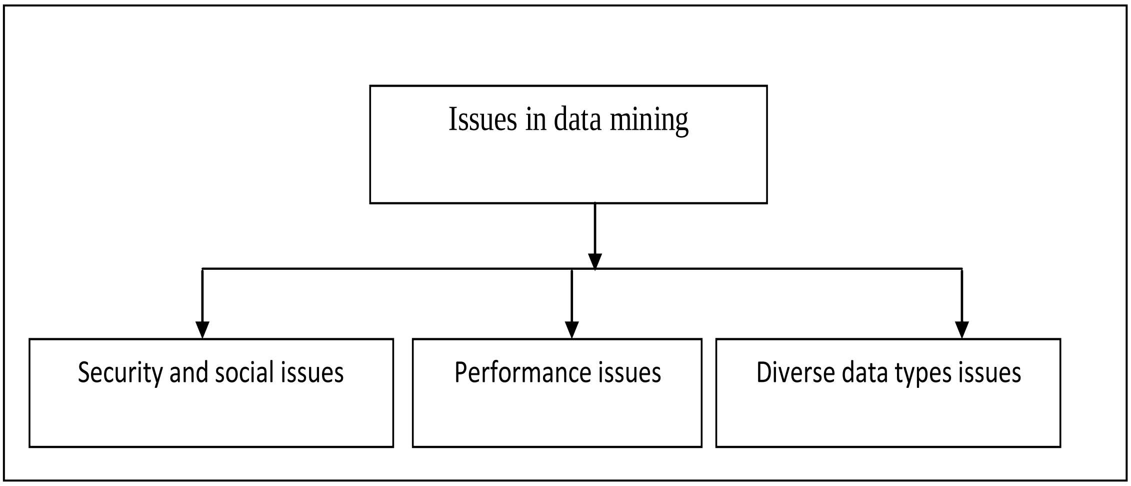


Fig. 1.3 Data Mining Issues

Issues related to security and social: Security provides large amount of responsive and confidential facility to data which is shared among many users.

Issues relevant to performance:

- Efficiently removing the database content which is large in size provides scalability and better performance.
- In many algorithms growth occur because of many reasons such as complexity etc in various methods of data mining. These algorithms processed data by dividing it into partitions which is further parallelized. Then the result from the partitions is combined.

Issues related to different types of data:

- Containing data which includes relations and complexity.
- Databases may contain different sources by involving all information. Therefore mining knowledge from them is a big challenge for data mining.

Nowadays, databases for the storage purpose are required to be increase in size as per need of technology. Data extraction from these databases is needed to be done in a specific manner for better use in future. Databases are of big size and data mining provides help to identify the particular information. In other words, data mining is discovering the latest knowledge from the available sources. Various rules are applicable on market based analysis, medical applications, engineering and science applications, music, data mining, banking etc for decision making. Apriori algorithm which includes the association of items gives relationship among those item sets uses association rule mining.

1.4 ASSOCIATION RULE MINING

Association rule mining having two main steps:

- **Creating item sets which are frequent:** the present item sets must be equal to or more than the min support count.
- **Generate strong rules:** the condition for having a rule is strong is that it must satisfy the min support and min confidence. Also introducing the following concepts:

Item set defines the total items present in the set. K item set shows the existence of k items in the set. Example can be taken as, {laptop, Software, pen drive} which is a 3-Itemset. Support count provides the occurrence of items in the given item set. Frequent item set contains the items which satisfy the min support count [15].

1.5 FREQUENT ITEM SETS AND ASSOCIATION RULES

Association rule and frequent item set mining a very popular area for research work. The apriori algorithm includes both frequent item sets and association rules as important terms. Association rules defines items in a given set describes that when item P contain items from database then Q also contains the items of same database.

Using Market Basket Analysis it is analyzed that customer purchase mostly which type of items. Thus frequent item sets are obtained by such analysis using some available techniques.

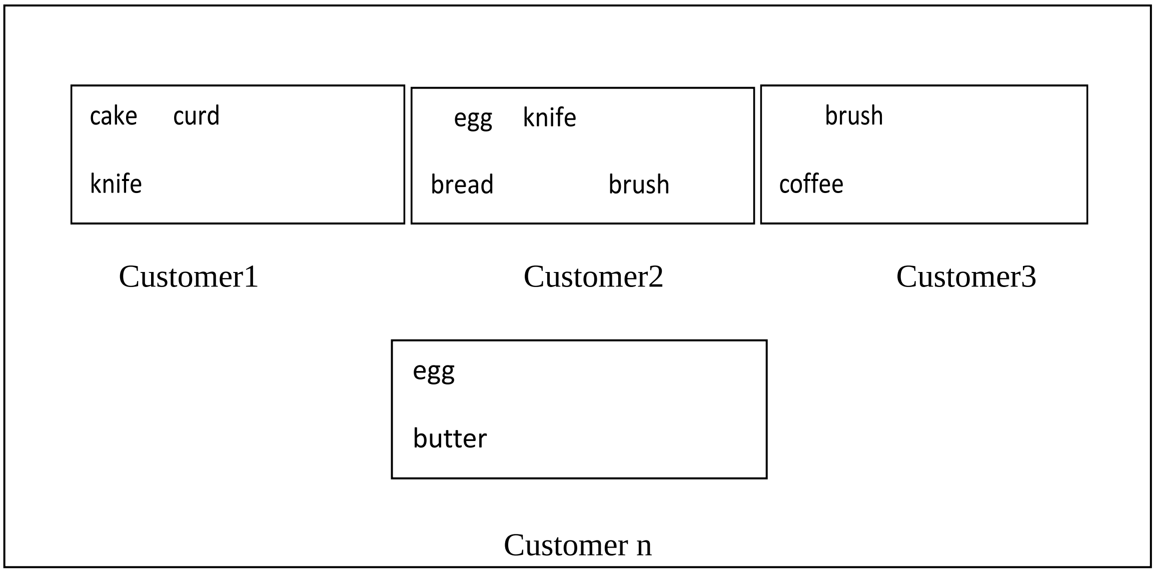


Fig. 1.4 Market Basket Analysis

- Frequent item sets: which are frequently purchased or in other words those items which satisfying minimum support.
- Association Rules: Frequent item sets are used to create rules.

E.g., Bread → Milk [given support = 5%, confidence = 100%],

A collection of items is known as item set and an item set with N number of items is N-item set.

A frequent item set {M, B} =4

M→B= 100% confidence

1.5.1 MEASURES IN ASSOCIATION RULES

Two main components are available in association rules.

Support: The support of an item set is the total transactions that contain all the items of that particular item set. That means containing both A and B: A∪B.

Support= P (A∪B).

Confidence: The confidence of any association rule is the ratio of support and the support of A.

Confidence= P (B | A)

Support and Confidence

Support count: The support count is defined as an item set which contain A item from all transactions. It is represented by A.count.

Support= $((\mathrm{A} \cup \mathrm{B}).\mathrm{count})/(\mathrm{n})$

Confidence= $((\mathrm{A} \cup \mathrm{B}).\mathrm{count})/(\mathrm{A}.\mathrm{count})$

$$\text{Support (XY)} = \frac{\text{Support count of (X}\cup\text{Y)}}{\text{Total number of transactions in database}}$$

Confidence also includes probability which works with condition that if X occurs in an event then Y will also occur.

$$\text{Confidence (XY)} = \frac{\text{Support count of (X}\cup\text{Y)}}{\text{Support(X)}}$$

The association rule must satisfy the requirement to find the problems within the association that both support and confidence have value not less than the given threshold. If the aim is not taken into consideration that that particular item set is neglected from frequent item set [3].

1.6 APRIORI ALGORITHM

Apriori algorithm is an advantageous algorithm presented to mine the items which are frequently purchased and also association rules are generated. The Apriori includes prior knowledge of items which are frequent. This is a step wise search in which by using the previous item sets, the next level item sets are found. Firstly, L1frequent -1 item sets after collecting the occurrence of items by scan the whole data is mined. Then accumulate all the items with minimum support. This result is shown by L1. This frequent-1 item set required to generate the next level item sets until it reaches to null value. It needs one full scanning of dataset to find the result of Lk. Apriori property is available in algorithm to reduce the search [5].

Apriori algorithm having mainly two steps which are:

1) Join step:

   The frequent item sets are joined step wise to find the candidates.

2) Prune step:

   When the items have support less as compared to the given minimum support those items are deleted under this step. Also, the item set having no sub set frequent is deleted.

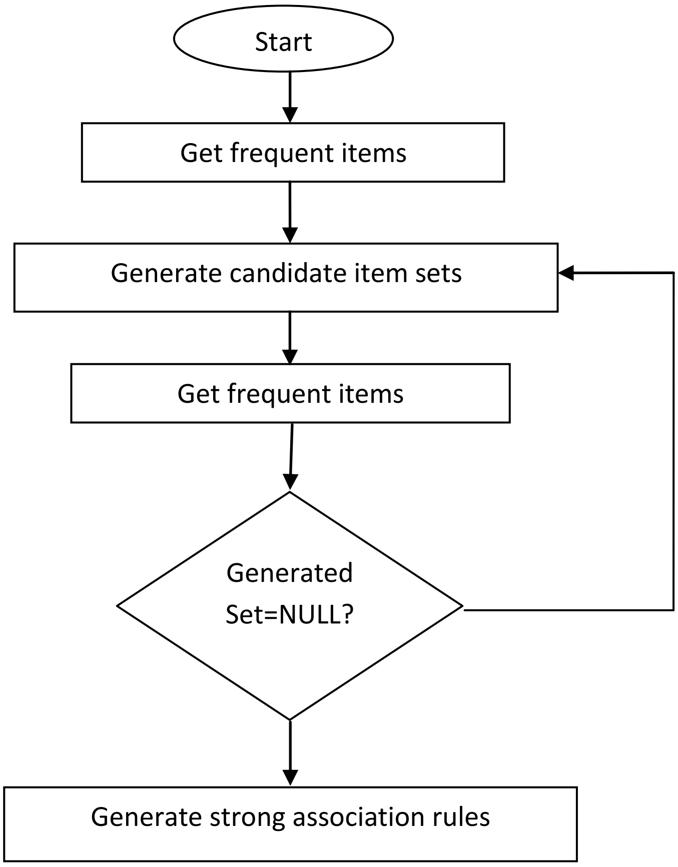

Start
Get frequent items
Generate candidate item sets
Get frequent items
Generated Set=NULL?
Generate strong association rules

Fig. 1.5 Flow Chart of Apriori Algorithm

Step to find the minimum support in Apriori algorithm:
Min support count= number of transaction * sup count.

For example: When the percentage values of support and confidence is given 60% and 60% respectively and number of transactions are 5 then
The result will be 5*60/100 = 3.

The performance of the classical apriori algorithm decreases because the algorithm requires multiple scans and when database size is very large, then it become very time taken process.

Apriori algorithm can be implemented on various applications and discusses how effectively e-commerce application can be used with Apriori algorithm to help in business decisions by knowing the customer buying behavior analysis especially in the retail sector. The role of Apriori algorithm is also explained for finding the item-sets which are frequent and association rules generation is also included. The dataset includes analysis of the set of products purchased by the customer in a period of time is selected. Two main measurements are used in finding the frequent item-sets and strong rules of association.

For all the transactions support is calculated which defines the association of dataset or item set [4].

1.6.1 EXAMPLE OF APRIORI ALGORITHM

Database D having transactions is shown in Table 1.1. Given min sup is 3. The steps in algorithm are: [6].

1. The transactions are scanned to check the number of times each item present and the result is stored in candidate 1- item sets, C1.
2. Given min sup is 3, then the L1, frequent-1 item sets can be find out which contains C1 with min sup.
3. Because the items q5, q6, q7 having support count less than 3, so delete these items.
4. To generate the next level of frequent 2- item sets, that is L2, the Join operation is used on L1 which gives the candidate 2-items C2.
5. The database scanning is done to take the support count of each candidate item set in C2.
6. Now, the next level frequent 2-itemsets, L2, is found, containing items of C2 with min sup.
7. When L2 is created, to find frequent 3-item sets, L3, firstly candidate 3-item sets using Join operation on L2 is generated. Then, C3 contains {q2, q3, q4} andC3 is pruned according to the Apriori property.
8. Using C3 candidate items, by collecting support count of each item on scanning the database the frequent-3 item set, L3 is generated.
9. Because L3 contains only 3 item sets that is the reason of null value in C4. The steps of algorithm will stops when null comes at any level.
10. The candidate item sets will be created till the next level of candidate (CK+1) empty.

D

TID	Items
K1	q1,q3,q7
K2	q2,q3,q7
K3	q1,q2,q3
K4	q2,q3
K5	q2,q3,q4,q5
K6	q2,q3
K7	q1,q2,q3,q4,q6
K8	q2,q3,q4,q6
K9	q1
K10	q1,q3

C1

Items	Sup count
q1	5
q2	7
q3	9
q4	3
q5	1
q6	2
q7	2

L1

Items	Sup count
q1	5
q2	7
q3	9
q4	3

C2

Items	Sup count
q1,q2	2
q1,q3	4
q1,q4	1
q2,q3	7
q2,q4	3
q3,q4	3

L2

Items	Sup count
q1,q3	4
q2,q3	7
q2,q4	3
q3,q4	3

C3

Items	Sup count
q2,q3,q4	3

L3

Items	Sup count
q2,q3,q4	3

Table 1.1 Steps in Apriori algorithm

1.6.2 BENEFITS AND DRAWBACKS OF APRIORI ALGORITHM:

Benefits:

- This algorithm is very simple to understand.
- It can be implemented in an easy way**.**

Drawbacks:

- Repeatedly scan done over the database.
- The frequent item set length is directly proportional to the total database passes.
- For generating the candidate much time, resources are required.
- ARM is not so efficient for large data set.
- ARM treats all items in the database equally.

CHAPTER 2
REVIEW OF LITERATURE

Mohammed AI-Maolegi, BassamArkok (2014) proposed another approach for candidate item-sets generation which overcome wasting time for scanning the whole database as in Apriori algorithm. Before scanning all transaction records, use only before generated frequent item-set(Li) to get the transaction IDs of the minimum support count between X and Y items. Repeat the steps to identify all frequent item-sets. The candidate support count generation in improved algorithm is less time consumed as compared with the Apriori algorithm. Experimental results are shown by using the group of transactions with different available minimum support values, which are applied on both apriori and improved algorithms [8].

Jaishree Singh et al (2013) they have described an Improved Apriori algorithm which removes the unnecessary item sets and thus reducing the time for scanning the database. This proposed algorithm also reduces the generation of similar sub-items in the candidate item sets pruning step. The apriori algorithm performance is also improved in terms of speed for mining association rules. The improved algorithm has some limitations also as this is not efficient algorithm to manage the new database [6].

Chanchal Yadav et al (2013) proposed that various approaches are used to overcome the drawbacks of the Apriori algorithm as to improve its efficiency. The proposed approach is presented which decreases pruning operation of candidate item-set. Data consistency is improved by resolving the problems like bad data or duplicate data and instead of finding the whole dataset, focusing on finding association in the filtered dataset. Other technique is also used to overcome the exceed limit of memory size by both frequent and infrequent item-sets by dividing dataset into horizontal partitions. The proposed idea reduces the size of each transaction and takes less time in comparison to Apriori algorithm to handle the data [1].

Marghny H.Mohamed et al (2013) they proposed new approach to avoid costly generation of candidates and also reduce their test processing by improving count tables for compression of transaction database. Proposed method is far different from apriori algorithm, it includes two algorithms named CountTableFI and BinaryCountTableF which represent all transactions in binary and decimal numbers. These algorithms construct a highly compact count table to discover frequent item sets with intersection in faster manner than traditional algorithms. Proposed method also used to compress the important information related to all item sets, avoid the expenses, and reduce the repeatedly database scans. The results are compared with apriori and FP-Growth which show that this method saves costly database scans in different mining processes. Using proposed method it is also tried to resolve the IMFI (Incremental Maintenance of Frequent Item sets) by applying different theories and solving complexity problems during computations of results [10].

D. Gunaseelan, P. Uma (2012) for mining the frequent items they proposed an improved algorithm in which database transposition is done and reducing the number of scans in database for generating the frequent patterns. This technique reduces the size of database as well as the time consumed. The results are compared with classical apriori algorithm, which shows that the proposed technique is much faster [2].

RuPeng Luan et al (2012) they proposed new algorithm which is more efficient than original apriori algorithm. This new dynamic algorithm is suitable for mining in dynamic database and also applied on web log mining to discover association rules. Algorithm improves the efficiency by greatly reducing the search range for discovering the association rules. Experimental results show that the proposed algorithm's performance has improved as compared to apriori algorithm [13].

Shuo Yang (2012) they described data mining technique which is used with an improved apriori algorithm and this proposed algorithm is applied on shopping areas of e-commerce. For calculation of support count and confidence level the algorithm establishes a new technique. Association rules are required to be obtained from apriori algorithm as results. Proposed algorithm helps in generating association rules more quickly for users and required products in shopping areas are recommended to customers on time. The rules which are obtained are separately stored in a table named as "tb_Association". Every transaction done by customer is stored in the accounts and scanning is done for taking detailed results about what the customers has purchased already [15].

Rahul Mishra, Abha choubey (2012) they presented that web marketing growth has increased because of popularity of internet. Useful patterns are extracted to deal with the weblog records which are used to discover accessing patterns of users from web pages. Databases of weblog provide the useful information about the users and web pages which are accessed. These weblog can identify customers for e-commerce by considering regularities and also improve the quality of services of internet for end users. In data mining, the association rules play a major role to find the interesting patterns in database. Apriori algorithm is used for the mining process to find frequent patterns. Apriori algorithm has some drawbacks such as costly generation of candidate item sets and it takes much time to scan the database. The FP-tree is used for compressed storage of information about frequent patterns and developed a FP-growth, for the efficient mining of frequent patterns. FP-algorithm is more efficient than apriori algorithm, it uses divide and conquer technique and also takes less time by giving better performance [14].

Qiang Yang, YanhongHu (2011). In this paper the application of education training have been used. They found the related rules of particular course to provide the information significance. The improved Apriori algorithm contains whole data from the database of educational information and rules are generated. This algorithm can help in decision making, arrange the course as per requirement, and also improve the teaching skills [12].

Zhuang Chen et al (2011) analyzes the shortcomings of Apriori algorithm and studies the further improvements of Apriori. They proposed improved algorithm named as BE-Apriori which includes pruning optimization and transaction reduction strategies. Pruned optimization strategy have used temporary table to count the frequency of items all in the frequent item-sets and, transaction reduction strategy to compress the size of transaction and reduce the scale of database scanning. The improved algorithm has decreased the number of frequent item-sets generated and also reduced the running time. The experimental results about dataset retail have show the advantages of proposed algorithm of low system overhead, good operating performance and higher efficiency, as compared with pure Apriori [16].

MazaDimitrijevic, ZitaBosnjak (2010). They proposed an analysis on online educational institute for available association rules and these association rules are applied to get important information from online data. Reducing the size of proposed rules and also non-interesting rules are removed. The analysis of association rules confirmed the hypothesis that it is not time consuming to discover interesting and useful association rules in web usage data [9].

Lamine M. Aouad et al (2009) they present a new distributed approach in the proposed approach grid environment is used for mining frequent item-sets which requires pruning strategy locally. It is showed that intermediate communication steps are not locally efficient in classical techniques, which affects their performance. But the present approach improves the performance to acquire more scalability by limiting the overheads related to synchronization and communication. Comparison is done with the local apriori algorithm and results are shown by using workstations of large clusters [7].

CHAPTER 3
PRESENT WORK

3.1 PROBLEM FORMULATION

In association rule mining major problem was to find the different items associations in sales marketing. The items that occur together in any data set are analyzed to find out the association rules. The Apriori algorithm is the most efficient algorithm to generate these association rules. But in apriori algorithm for finding the frequent item sets repeatedly database scan is done which takes much time and degrades the performance of apriori algorithm. The apriori algorithm is mostly applied on many applications to provide the rules generated for that specific application and also give the details about items. In apriori algorithm is applied for large datasets and step wise search is used. Apriori property helps to search the frequent items and rules which are mostly used. Many database scans are necessary in this algorithm when the database size is large. This technique will much time taken and need to be improved. The efficiency of Apriori algorithm will be enhanced by reducing the transitions which further reduce the time need to be consumed to generate the association rules. So, in this work, proposed technique will work on reducing the number of transitions of apriori algorithm.

3.2 SCOPE OF THE STUDY

In mining association rules the Apriori Algorithm is the popular algorithm. Apriori Algorithm which is used to generate all association rules between the items present in the database. This classical approach is inefficient due to multiple database scans. It takes much time for scanning the whole large database. The proposed technique will do an enhancement in Apriori algorithm by reducing the transactions as well as reduce the similar sub-items generation during prune step and also candidate having not frequent sub items are deleted. With the use of proposed technique the apriori algorithm efficiency will be enhanced. This will provide association rules generated with low load on available resources and also in less time.

3.3 OBJECTIVES OF THE STUDY

1. To study in data mining the algorithms available for generating association rules and the most efficient algorithm is selected.
2. To analyze the shortcomings of Apriori's Algorithm.
3. To propose enhancement in Apriori's Algorithm using different approach, to reduce the number of transitions and processing time to increase efficiency of the algorithm.
4. To identify the transitions and time required by apriori algorithm.
5. To implement the proposed technique and compare it with the existing technique.

3.4 RESEARCH METHODOLOGY

Classical apriori algorithm of association rules is proposed. Apriori algorithm is a step wise search, applying to search for next item set from previous item set. The L1 is the frequent-1 item set which is used to compute next frequent-2 item set that is L2. Thus LK is used until it cannot find LK-1. The proposed method will use the transposition of dataset and then find out the frequent item sets using Apriori algorithm steps. The Apriori property is available which includes that all sub item sets having items, in frequent item sets must also be frequent. Algorithm acquires two steps to find out the frequent items and association rules. In proposed technique, one new value will define, called as the items present in the transition. Improvement will be done in classical algorithm.

To do that, this technique will be further enhanced by modify the minimum support calculation formula in transpose technique. When in any particular transition the number of items below the threshold value then from dataset that transition will be deleted. This approach will reduce the scans over the database, which will further reduce the time to generate the rules. MATLAB tool is used for implementation.

3.4.1 ALGORITHM FOR PROPOSED TECHNIQUE:

- **Objective:** To reduce the complexity (number of transactions) and also reduce the time taken by Apriori Algorithm for generation the frequent item sets proposed technique is used.
- **Parameters:** Set up the following parameters to reach the objective:
  - Dataset is required to implement the new technique.
  - Minimum support calculation formula is used in a different manner.
  - Apriori property is considered to generate the frequent item sets.
- **Transpose of Dataset:** Dataset is transposed, thus implementation is done by applying algorithm on transactions in place of items (As in apriori algorithm items are used to get frequent results).

  L1=(x0, y0, x1, y1)

  Transpose (0, 0, M, M)

  Where, M denotes matrix.
- **Minimum support calculation formula:** To calculate the support for each candidate item set, minimum support calculation formula is used.
- **Finding the frequent item sets:** Use new minimum support calculation by dividing the average number of transactions with total number of transactions. Then, apply the apriori algorithm on the transposed database.

```
LI =find_frequent_I-itemsets( D);

Ck = apriori _gen (Lk-l,minsup), // genration of frequent item sets using Apriori

for each transaction tED {

C1 = subset (Ck, t);

candidates

for each candidate C E C1

c.count+ +;                    //increment the count of all candidates

}

//union of transactions to find exact number of item sets present in each transaction.
```

```
  Return L1;
```

Frequent item sets are generated by using transactions available in dataset. Apriori property is applied on transactions and results are calculated by using matlab tool.

3.4.2 FLOW CHART OF PROPOSED TECHNIQUE:

To generate the frequent item sets in Apriori algorithm technique is proposed, which is transpose technique. In this work, market basket dataset is used on which apriori algorithm and proposed technique applied to get results. Proposed technique involves the transposition of the dataset and then minimum support formula is used to get the minimum support with some enhancement. Frequent item sets are generated by using this transpose technique on Apriori algorithm. This approach can reduce the time and complexity of Apriori algorithm and results analyzed on the basis of these terms.

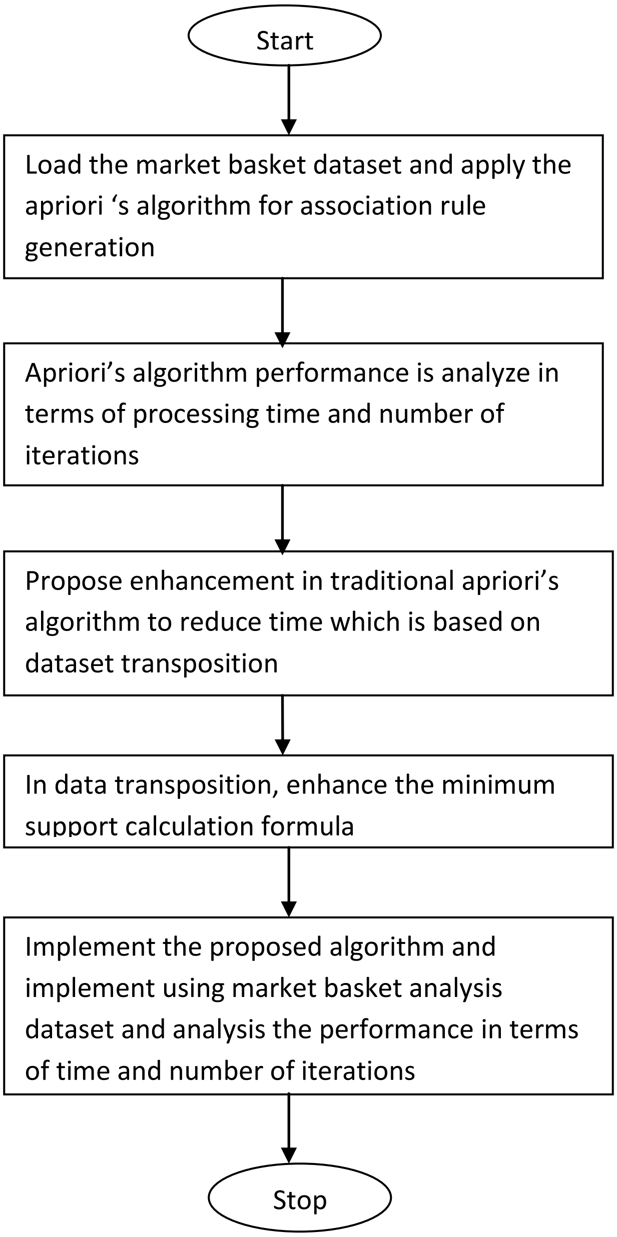

Start
Load the market basket dataset and apply the apriori ‘s algorithm for association rule generation
Apriori’s algorithm performance is analyze in terms of processing time and number of iterations
Propose enhancement in traditional apriori’s algorithm to reduce time which is based on dataset transposition
In data transposition, enhance the minimum support calculation formula
Implement the proposed algorithm and implement using market basket analysis dataset and analysis the performance in terms of time and number of iterations
Stop

Fig. 3.1 Flow Chart of Proposed Technique

CHAPTER 4

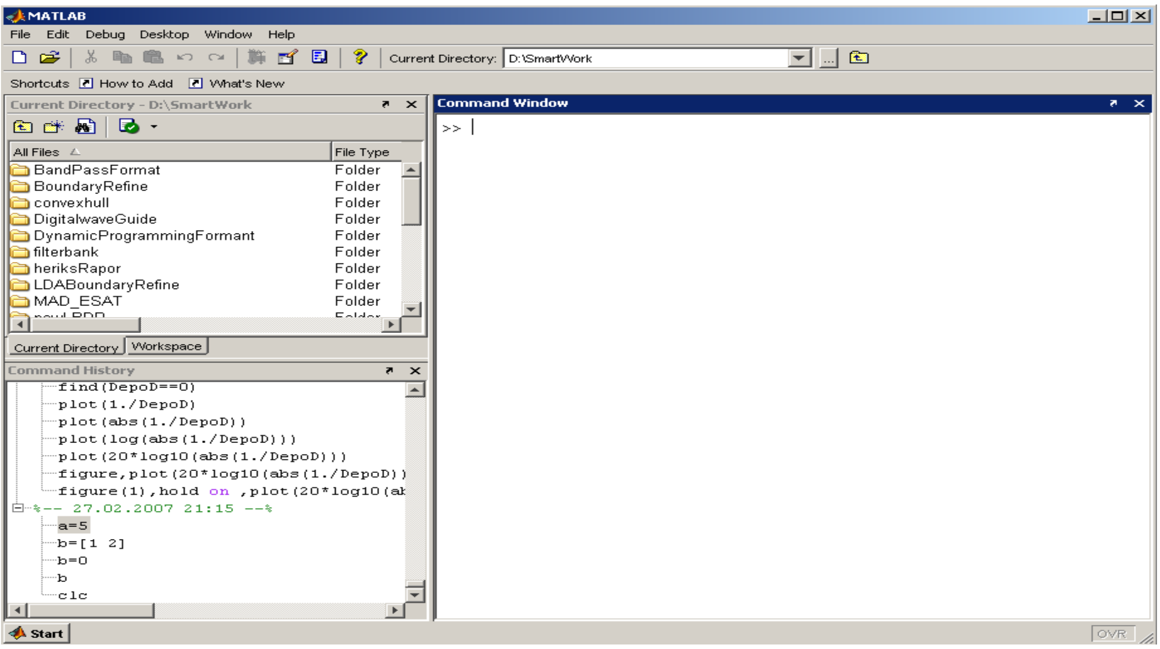
RESULTS AND IMPLEMENTATION

4.1 ABOUT MATLAB TOOL:

MATLAB tool is used for the implementation process. MATLAB (MATrix LABoratory) is a high performance language and is one of the easiest programming languages for writing mathematical programs. MATLAB also has some tool boxes useful for signal processing, image processing, optimization, etc. MATLAB has powerful graphic tools and can produce nice pictures in both 2D and 3D. Matlab also provide an easy-to- use environment and widely used in many research area.

4.1.1 STANDARD WINDOWS IN MATLAB:

- **Command window:** Type commands.
- **Current directory:** View folders and m-file.
- **Workspace:** View program variable and double click on a variable to see it in array editor.
- **Command History:** View past Commands and save a whole session to using dairy.



All Files	File Type
BandPassFormat	Folder
BoundaryRefine	Folder
convexhull	Folder
DigitalwaveGuide	Folder
DynamicProgrammingFormant	Folder
filterbank	Folder
heriksRapor	Folder
LDABoundaryRefine	Folder
MAD_ESAT	Folder

```
find(DepoD==0)
plot(1./DepoD)
plot(abs(1./DepoD))
plot(log(abs(1./DepoD)))
plot(20*log10(abs(1./DepoD)))
figure,plot(20*log10(abs(1./DepoD))
figure(1),hold on ,plot(20*log10(ab
%-- 27.02.2007 21:15 --%
a=5
b=[1 2]
b=0
b
clc
```

```
>> |
```

Fig. 4.1 MATLAB Screen

4.1.2 WORKING FORMAT IN MATLAB:

If the data is store in excel sheet on user's disc, firstly excel sheet is read into matlab using xlsread command. After the data is input into matlab from excel sheet then working on that particular excel sheet can be start. In matlab any data can be stored in excel sheet using xlswrite command.

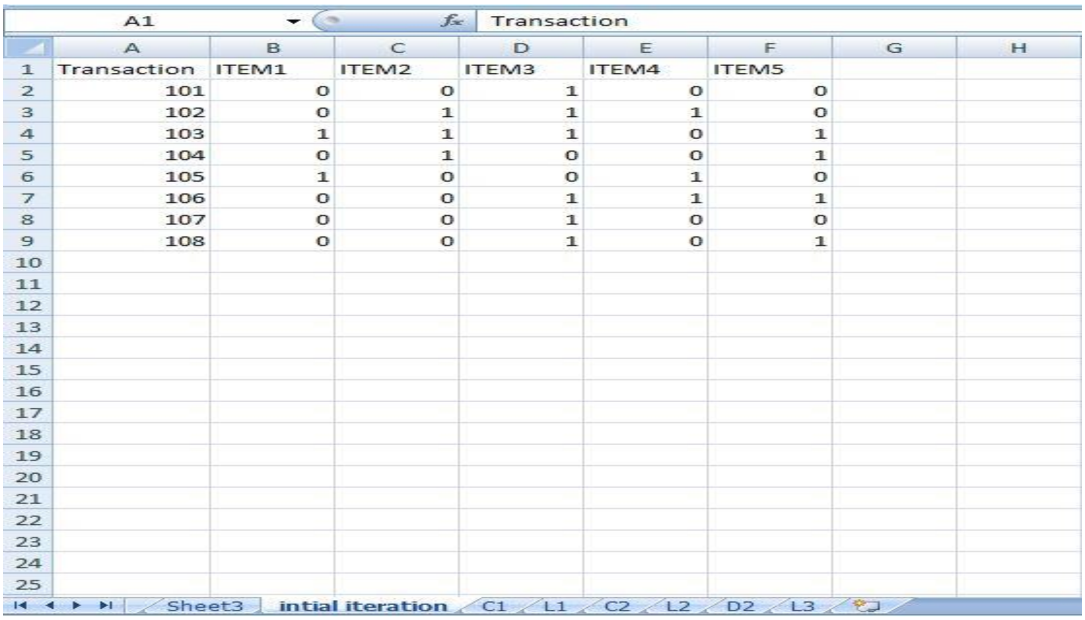
4.1.3 ABOUT INTERFACE:

GUIDE, the MATLAB Graphical User Interface Development Environment, provides a set of tools for creating graphical user interfaces (GUIs) which is used as interface in the implementation.

Guide Tool Properties:

- Allows the user to drag and drop components that is needed in the "layout" area of the GUI.
- All "guide" GUI's start with an opening function.
- Saving automatically generates an .m file and .fig file.
- .fig contains the binary GUI layout and .m contains the code that controls the GUI.

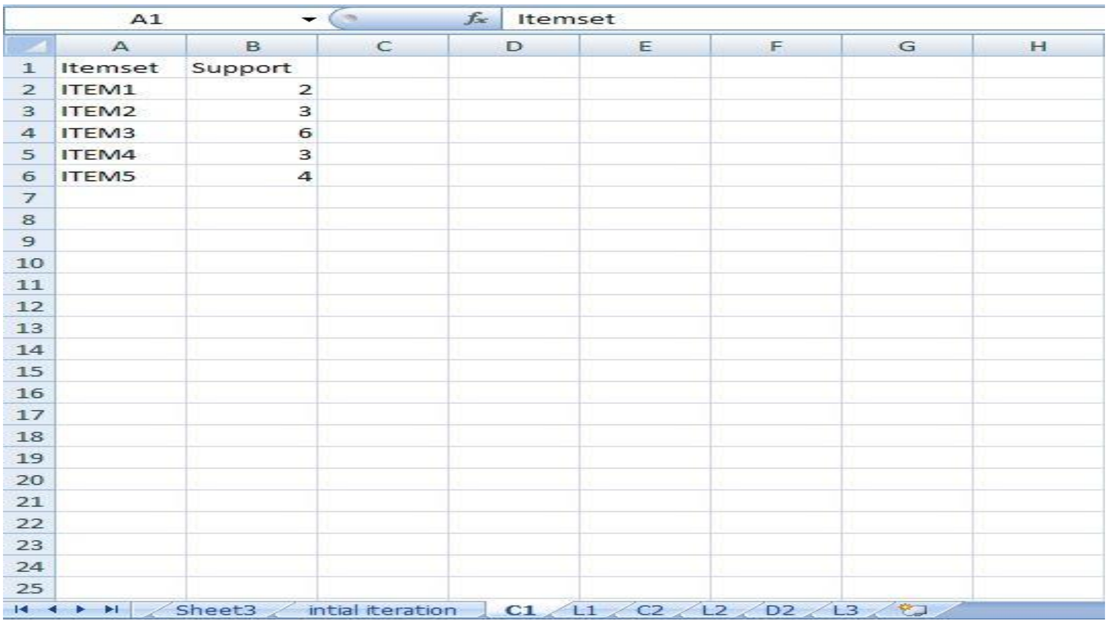
Excel sheet is taken as input dataset which includes transactions and corresponding items are present for every transaction.



Transaction	ITEM1	ITEM2	ITEM3	ITEM4	ITEM5
101	0	0	1	0	0
102	0	1	1	1	0
103	1	1	1	0	1
104	0	1	0	0	1
105	1	0	0	1	0
106	0	0	1	1	1
107	0	0	1	0	0
108	0	0	1	0	1

Fig. 4.2 Dataset Defined

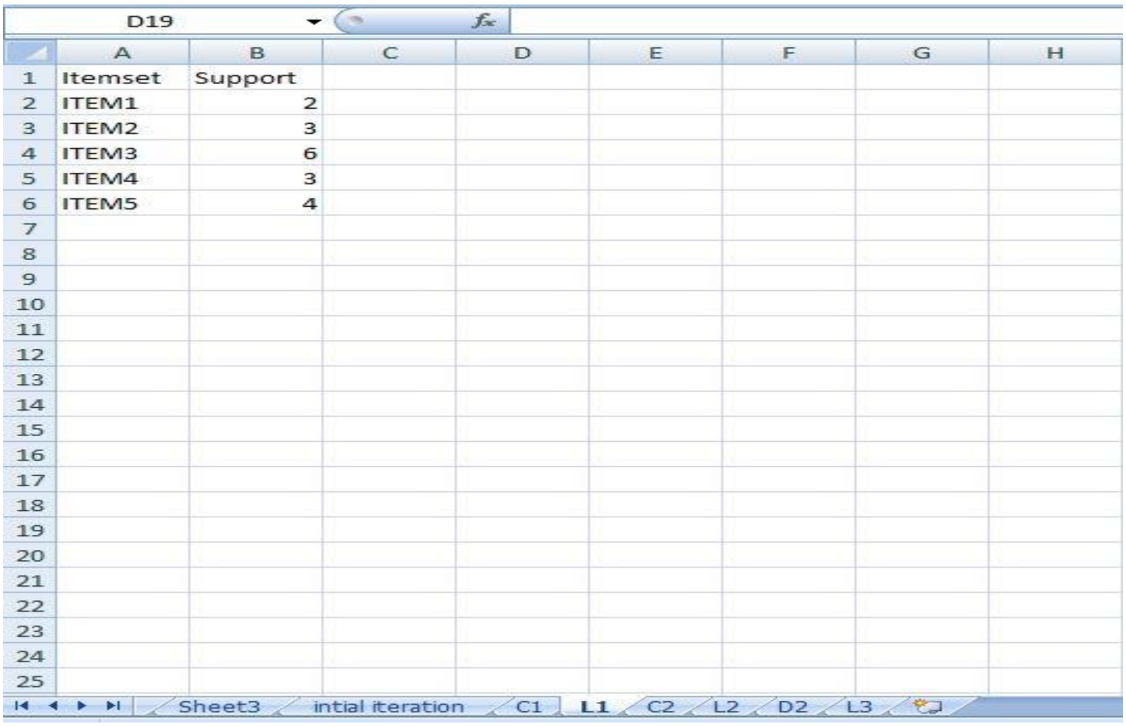
As illustrate in the figure 4.2, the dataset is defined of the super market. On this dataset Apriori algorithm is applied to generate frequent item sets. Minimum support is taken 2.



Itemset	Support
ITEM1	2
ITEM2	3
ITEM3	6
ITEM4	3
ITEM5	4

Fig. 4.3 Items Count to generate candidate-1 item sets (C1)

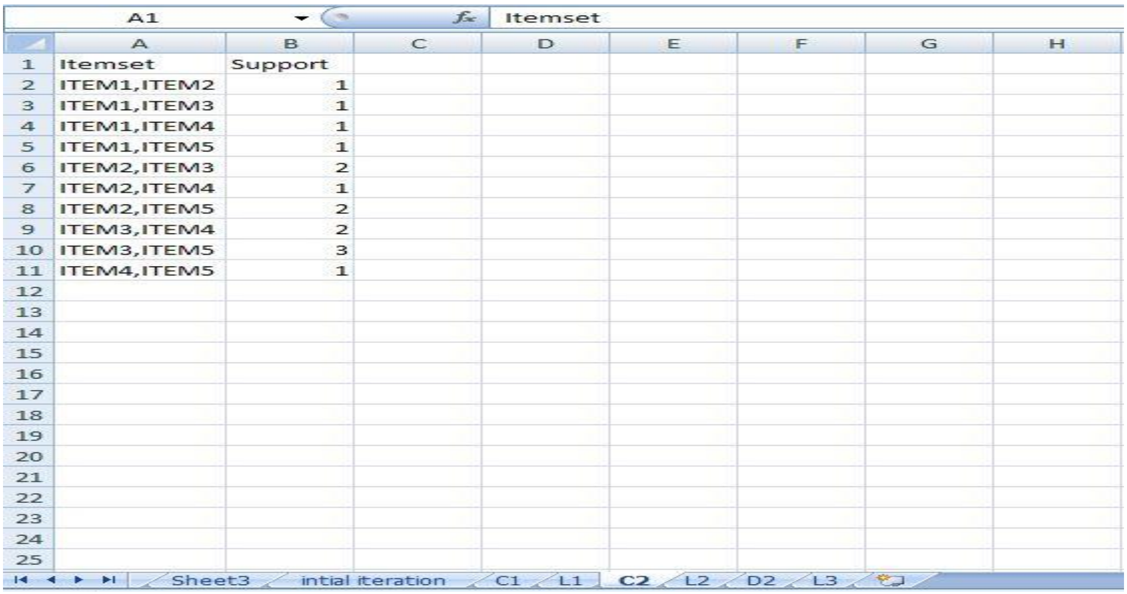
After the Item count, the items which have minimum value than support count will be deleted from the list.



Itemset	Support
ITEM1	2
ITEM2	3
ITEM3	6
ITEM4	3
ITEM5	4

Fig. 4.4 Items count to get the frequent-1 item sets (L1)

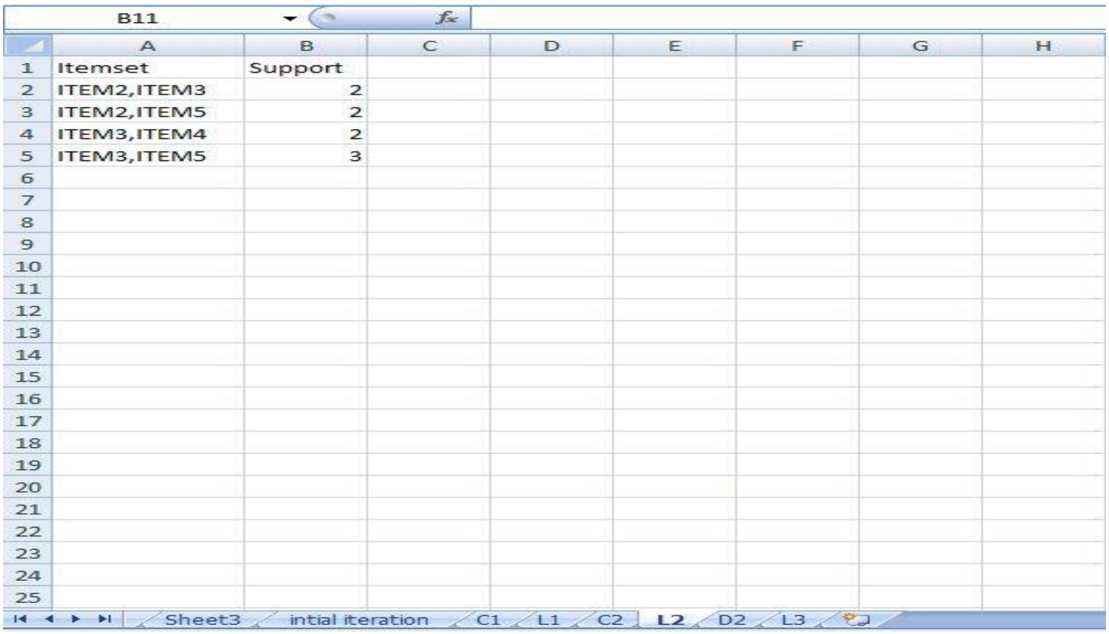
As illustrate in the figure 4.4, the item count is defined which is compared with minimum support count and results are generated for L1.



Itemset	Support
ITEM1,ITEM2	1
ITEM1,ITEM3	1
ITEM1,ITEM4	1
ITEM1,ITEM5	1
ITEM2,ITEM3	2
ITEM2,ITEM4	1
ITEM2,ITEM5	2
ITEM3,ITEM4	2
ITEM3,ITEM5	3
ITEM4,ITEM5	1

Fig. 4.5 Items combined in the pair of two to get candidate-2 item sets (C2).

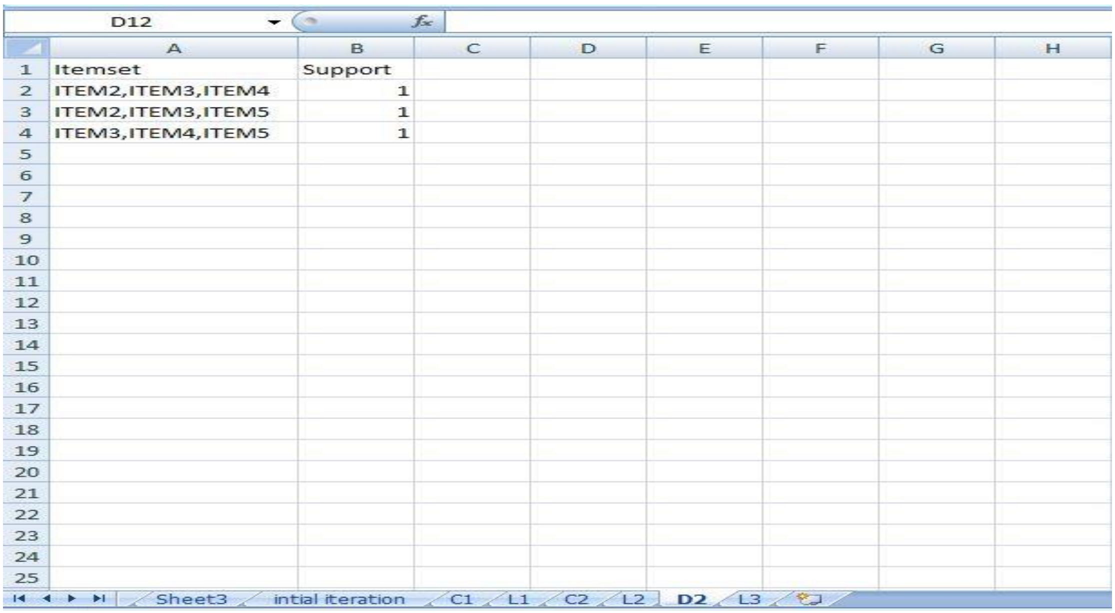
To generate C2, as shown in figure 4.4 two items are combined and support is count for each pair of items from dataset.



Itemset	Support
ITEM2,ITEM3	2
ITEM2,ITEM5	2
ITEM3,ITEM4	2
ITEM3,ITEM5	3

Fig. 4.6 Items count again to get the frequent-2 item sets (L2)

The item which has value less than minimum support count will be deleted from the list to generate L2 item sets.



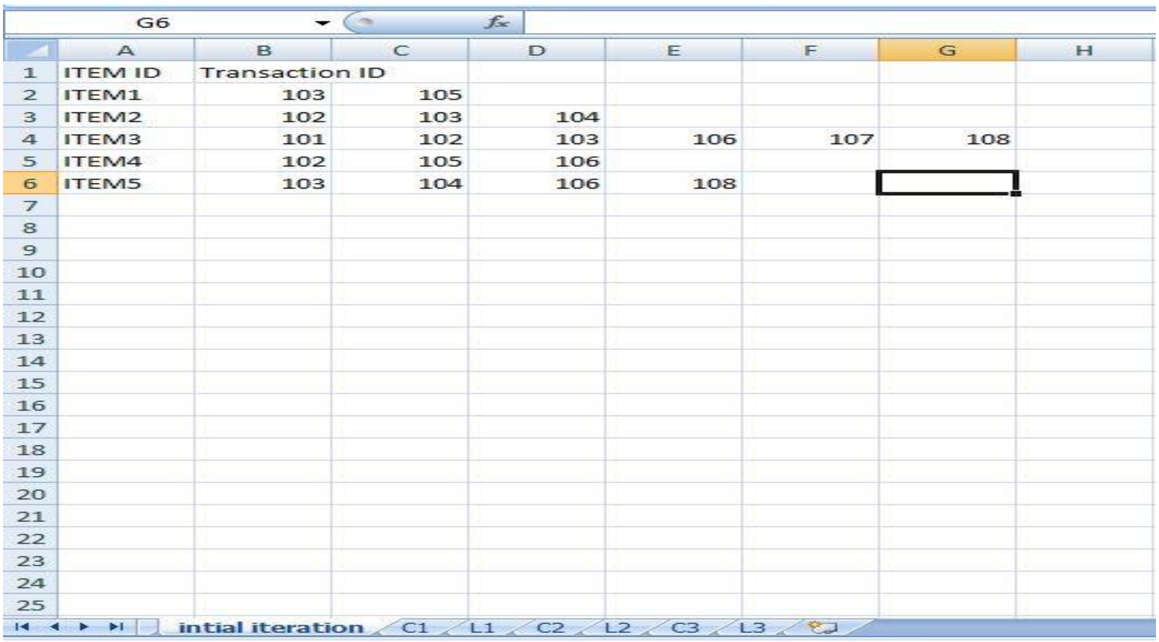
Itemset	Support
ITEM2,ITEM3,ITEM4	1
ITEM2,ITEM3,ITEM5	1
ITEM3,ITEM4,ITEM5	1

.

Fig. 4.7 Items combined in the pair of three to get candidate-3 item sets (C3)

Frequent item sets are generated using Original Apriori algorithm. As no candidate frequent item set satisfies the minimum support criteria, L3 will (frequent item sets 3) contains null.

Results on the basis of proposed work stored in excel sheet is shown as following in which transpose technique is used.

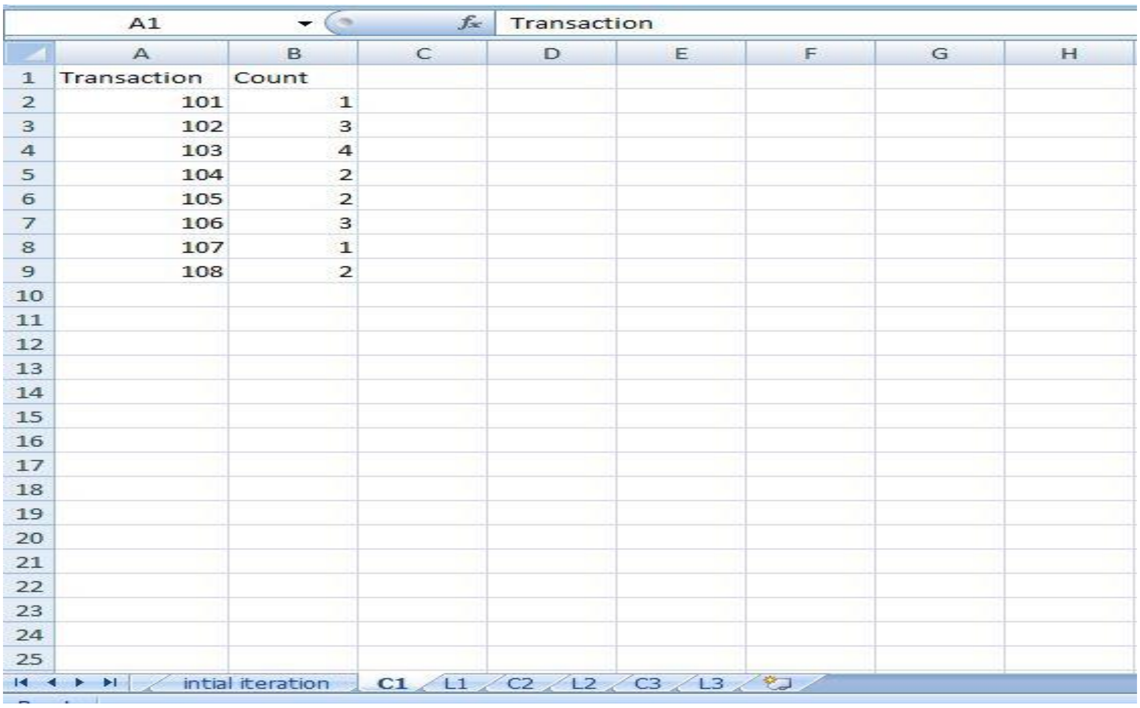


G6

	A	B	C	D	E	F	G	H
1	ITEM ID	Transaction ID						
2	ITEM1	103	105					
3	ITEM2	102	103	104				
4	ITEM3	101	102	103	106	107	108	
5	ITEM4	102	105	106				
6	ITEM5	103	104	106	108			

Fig. 4.8 Transpose of the Dataset

As illustrated in the figure 4.8, to reduce the number of transitions, original data base will be transposed. Minimum support is calculated using support calculation formula by dividing the average of transactions with total number of transactions, which gives the value of 2.25(which is taken as 2). Apriori property is applied to generate the frequent item sets.

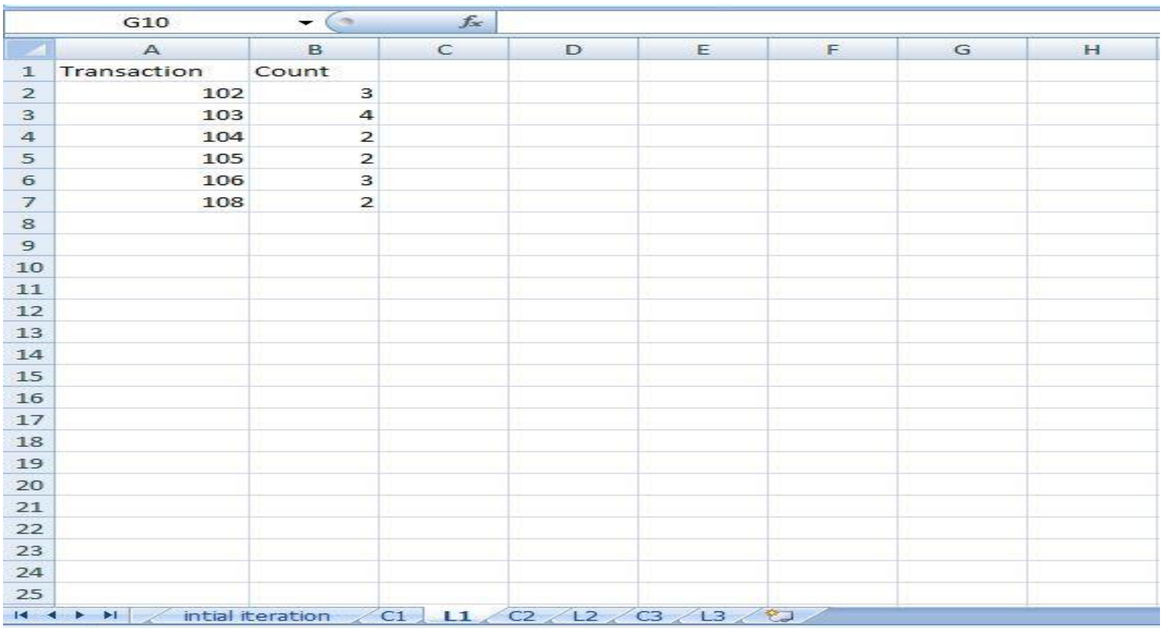


A1 Transaction

	A	B	C	D	E	F	G	H
1	Transaction	Count						
2	101	1						
3	102	3						
4	103	4						
5	104	2						
6	105	2						
7	106	3						
8	107	1						
9	108	2						

Fig. 4.9 Transaction Counts (C1)

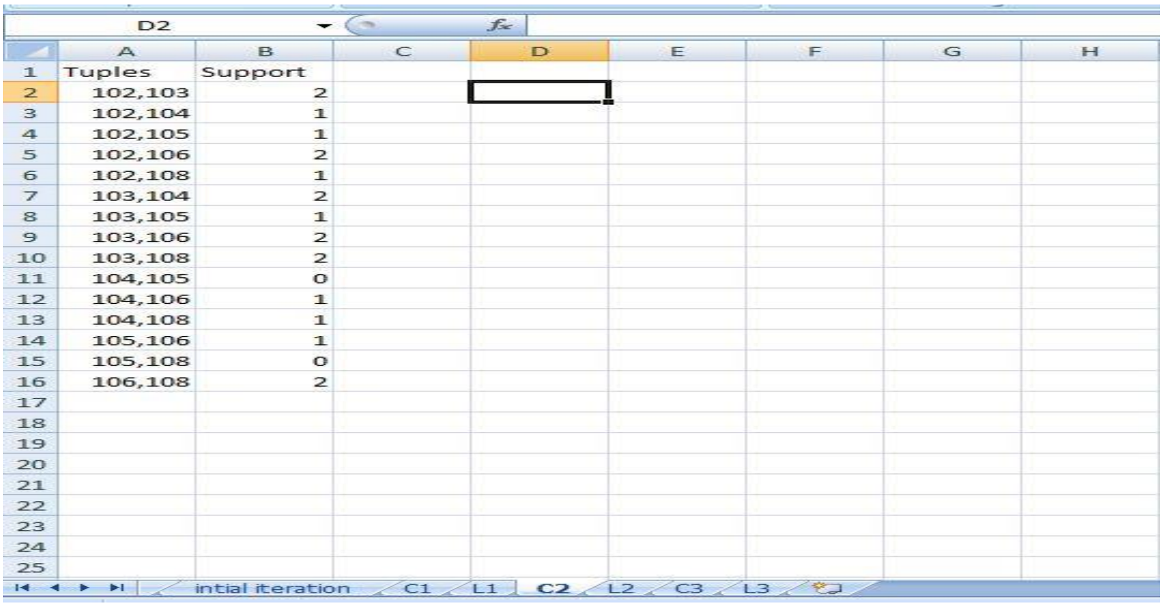
As illustrate in the figure 4.9, the count is defined to generate the candidate-1 item sets on the basis of proposed technique.



Transaction	Count
102	3
103	4
104	2
105	2
106	3
108	2

Fig. 4.10 Support count Consideration (L1)

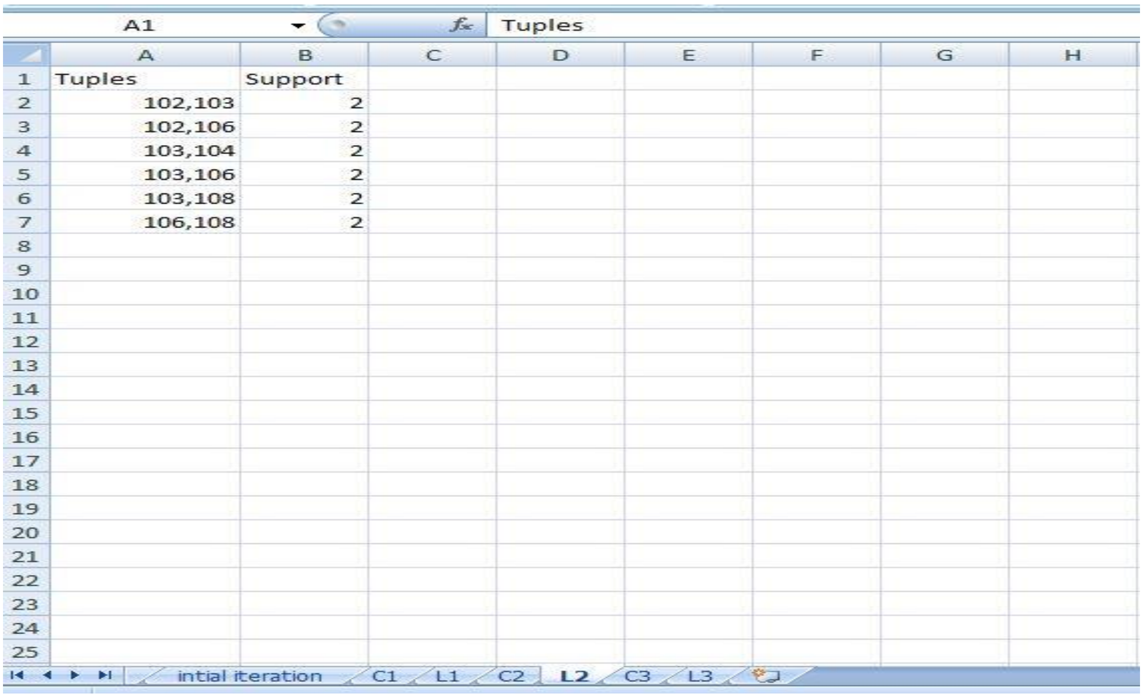
As illustrate in figure 4.10, the values less than minimum support count will be deleted to generate the frequent-1 sets.



Tuples	Support
102,103	2
102,104	1
102,105	1
102,106	2
102,108	1
103,104	2
103,105	1
103,106	2
103,108	2
104,105	0
104,106	1
104,108	1
105,106	1
105,108	0
106,108	2

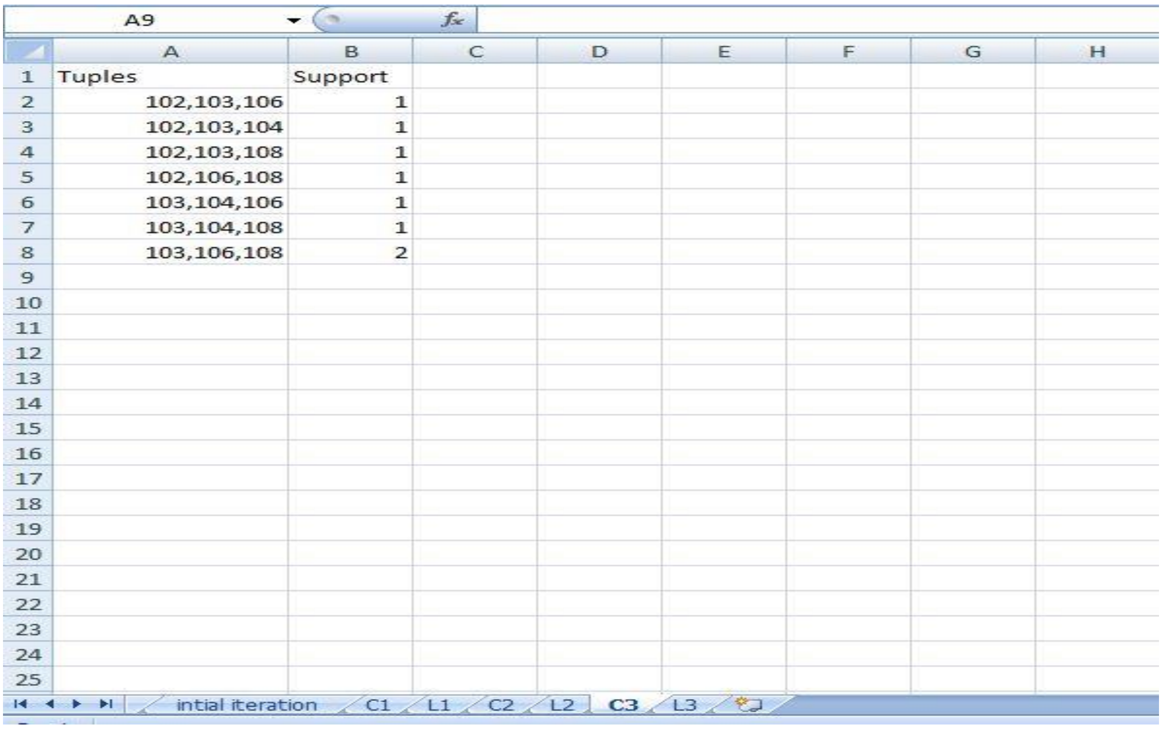
Fig. 4.11 Transactions combined (C2)

Transactions are combined in the pair of two by considering the availability of each pair in database.



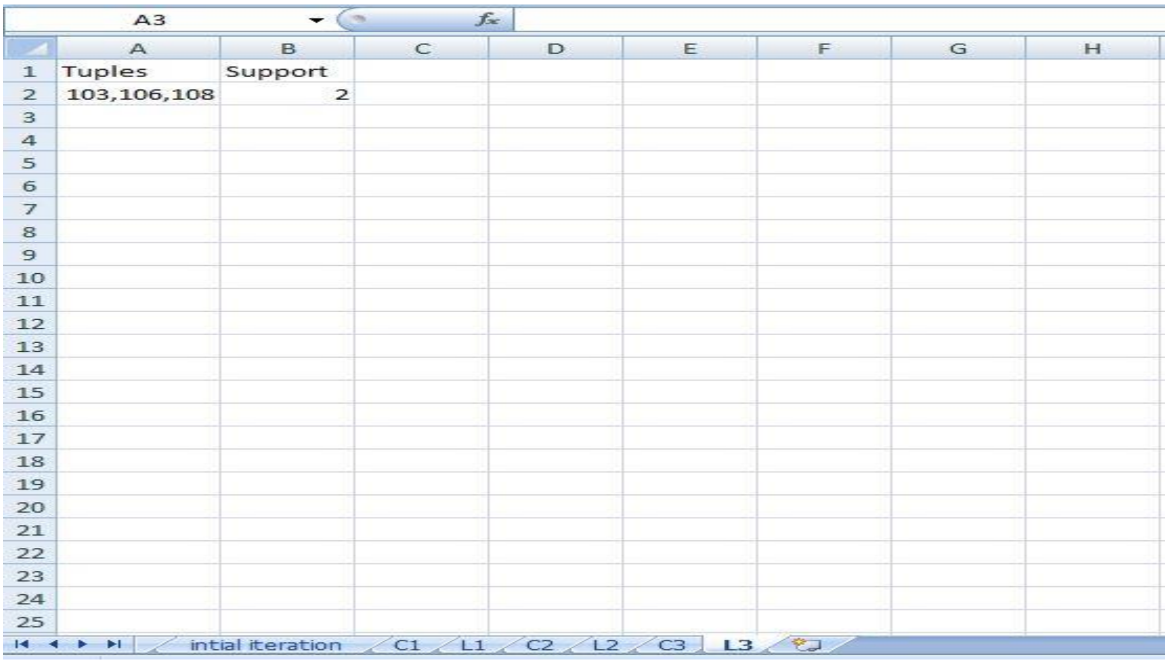
Tuples	Support
102,103	2
102,106	2
103,104	2
103,106	2
103,108	2
106,108	2

Fig. 4.12 Generate frequent-2 item sets (L2)



Tuples	Support
102,103,106	1
102,103,104	1
102,103,108	1
102,106,108	1
103,104,106	1
103,104,108	1
103,106,108	2

Fig. 4.13 Transaction count for candidate-3 item sets (C3)



Tuples	Support
103,106,108	2

Fig. 4.14 Transaction results (L3)

As a resultant, transactions are combined and generate association rules with minimum time and minimum transactions.

Using interface implementation of both existing and proposed technique:

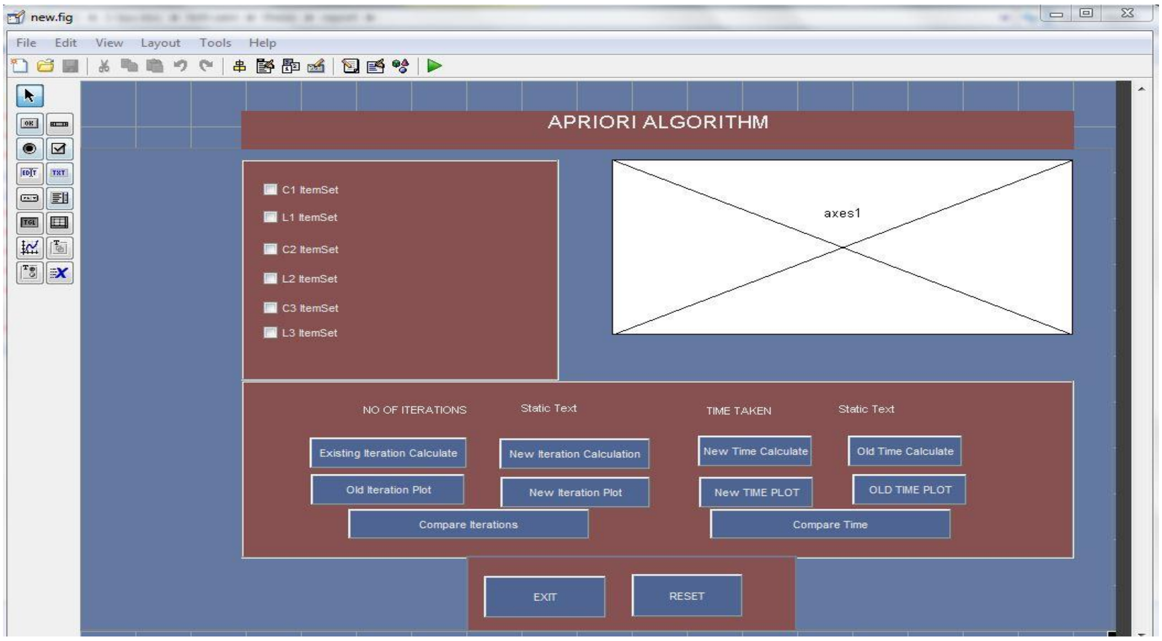


Fig. 4.15 Interface made by using Guide tool

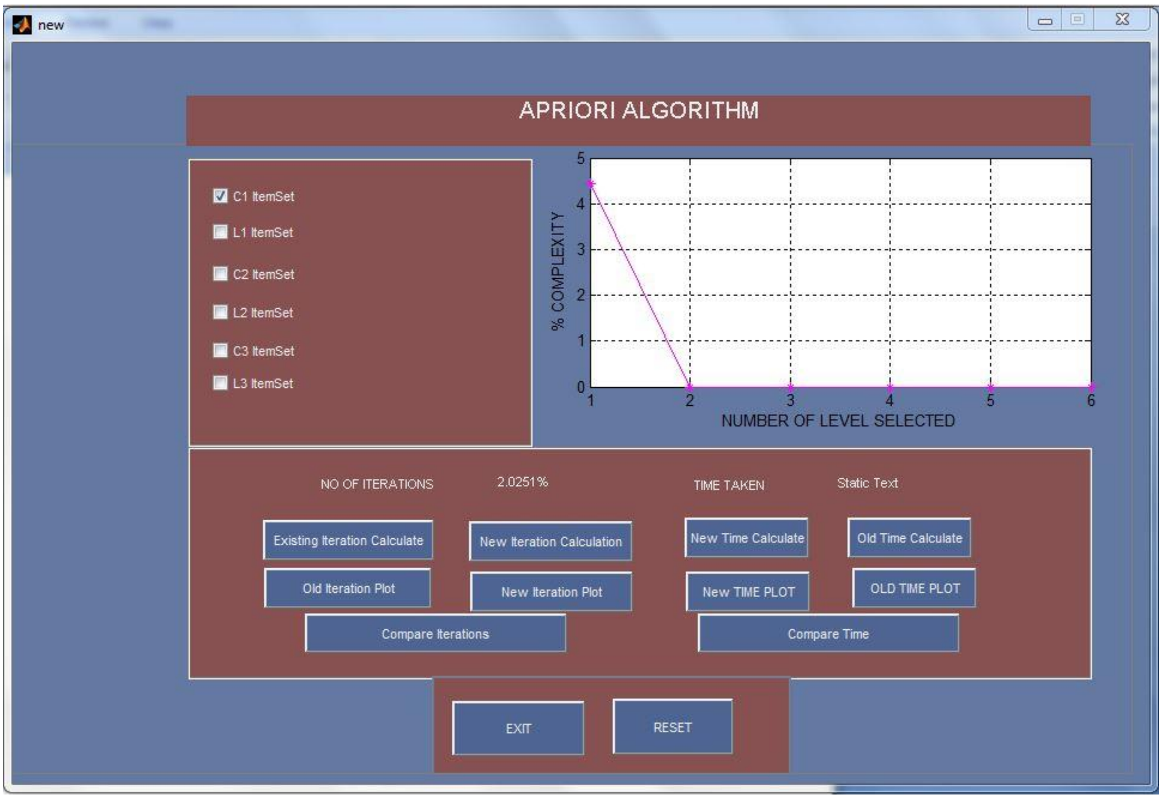


Fig. 4.16 existing iteration calculation only for C1 item set

In figure 4.16, to calculate C1, number of iterations used by apriori algorithm is shown.

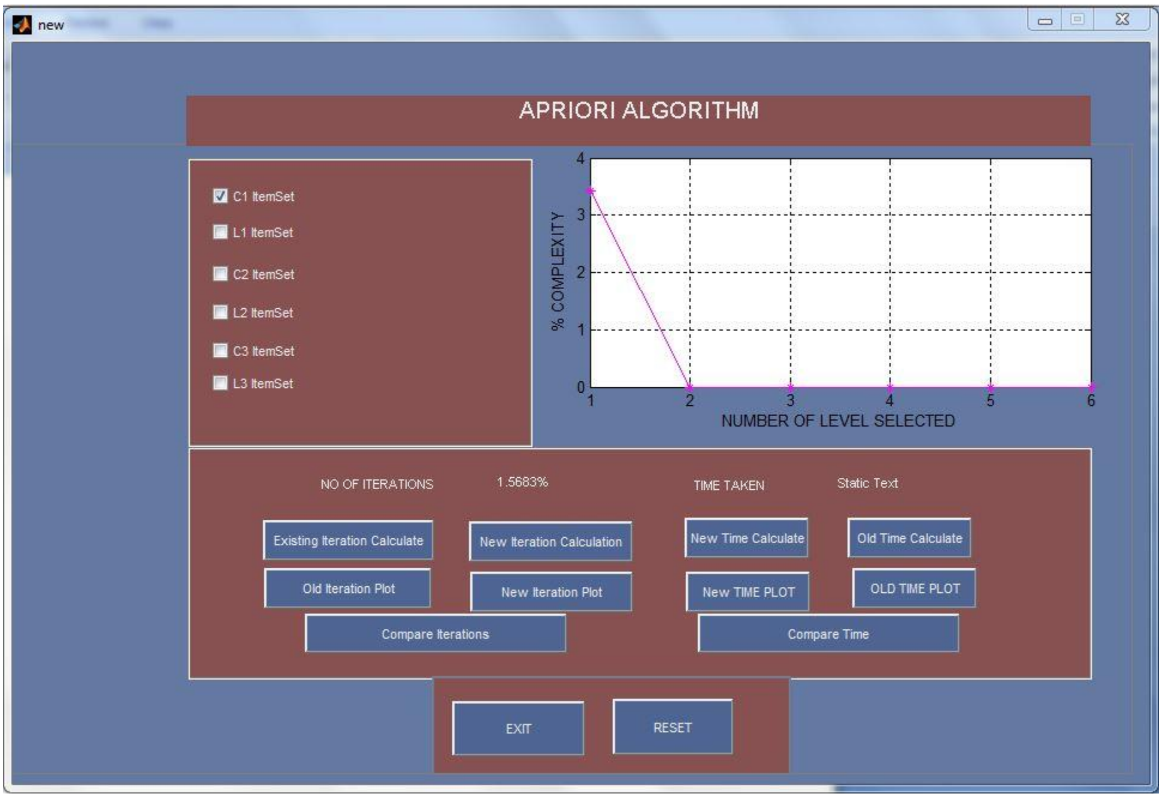


Fig. 4.17 new iteration calculation only for C1 item set

In figure 4.17, number of iterations taken by proposed technique with the complexity line graph for the same is shown to calculate only C1 candidate item set.

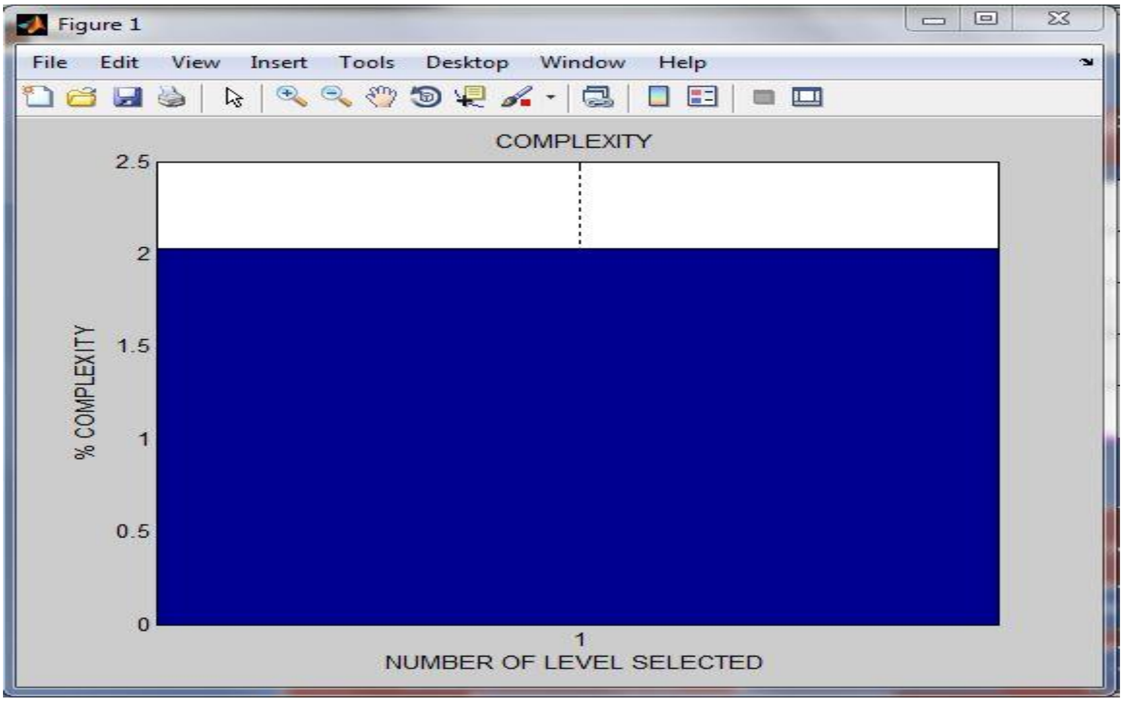


Fig. 4.18 Old iteration graph plot on the basis of complexity only for C1 item set

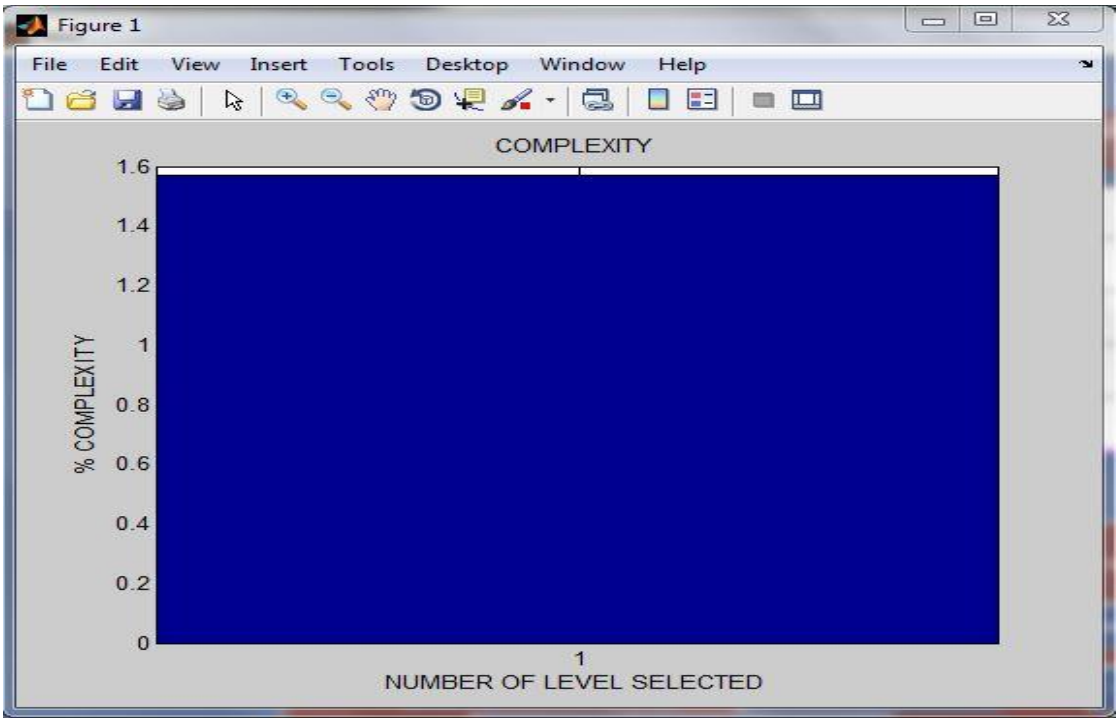


Fig. 4.19 New iteration graph plot on the basis of complexity only for C1 item set

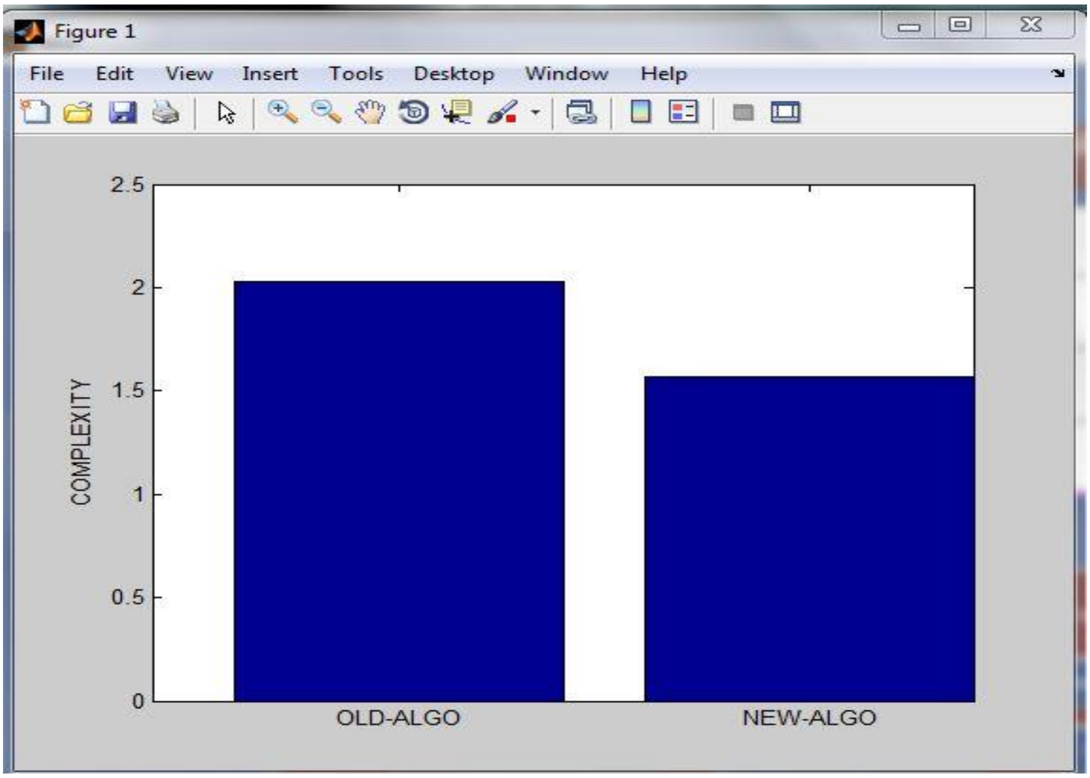
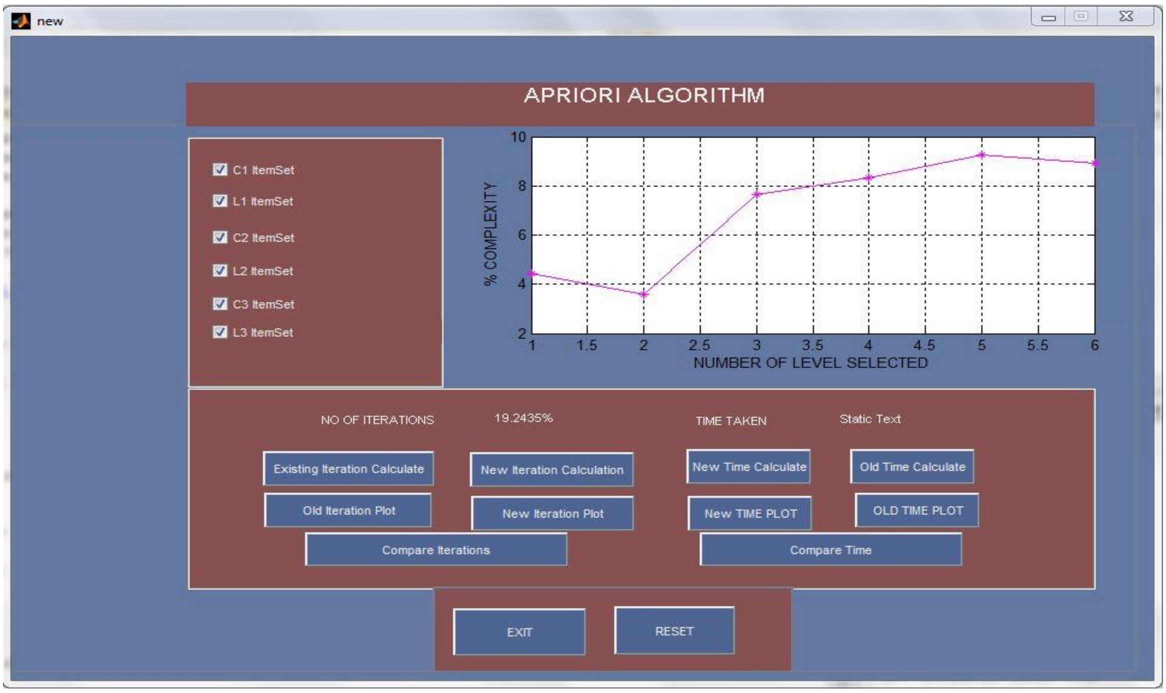


Fig. 4.20 Comparison on the basis of complexity only for C1 item set calculation

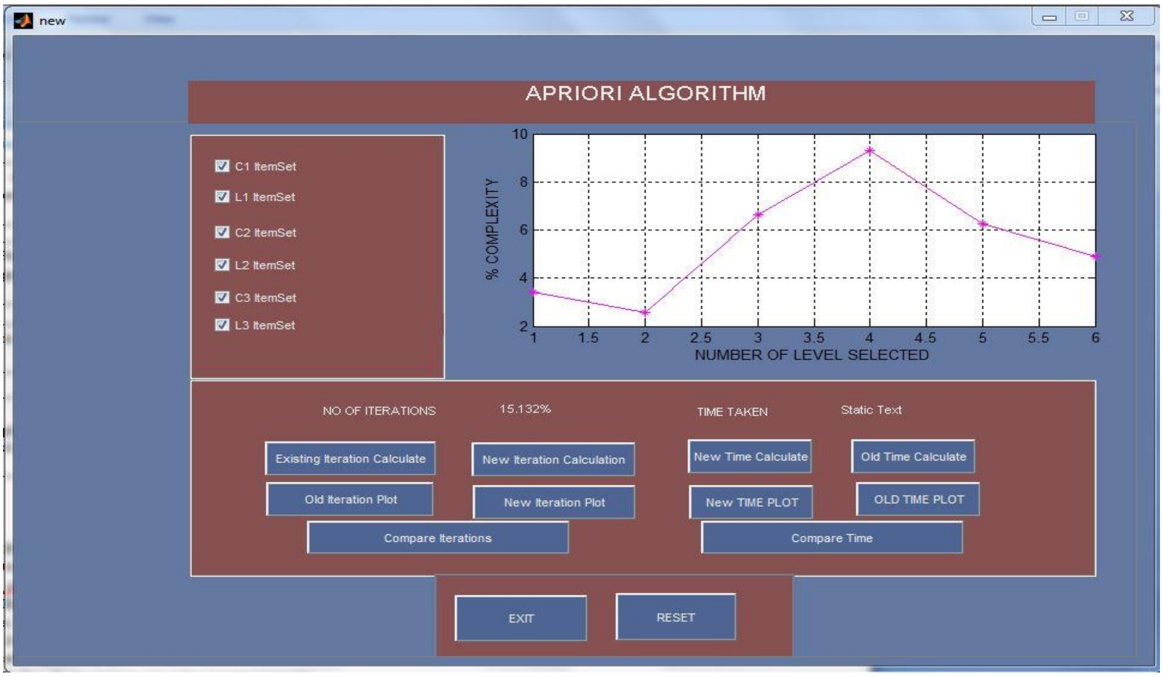


NO OF ITERATIONS 19.2435%

TIME TAKEN Static Text

Fig. 4.21 existing iteration calculation for all frequent item sets

In figure 4.21, number of iterations taken in apriori algorithm and the complexity line graph is shown to calculate all frequent item sets.



NO OF ITERATIONS 15.132%

TIME TAKEN Static Text

Fig. 4.22 new iteration calculation

In figure 4.22, number of iterations taken by proposed technique with the complexity line graph for the same is shown to calculate all frequent item sets.

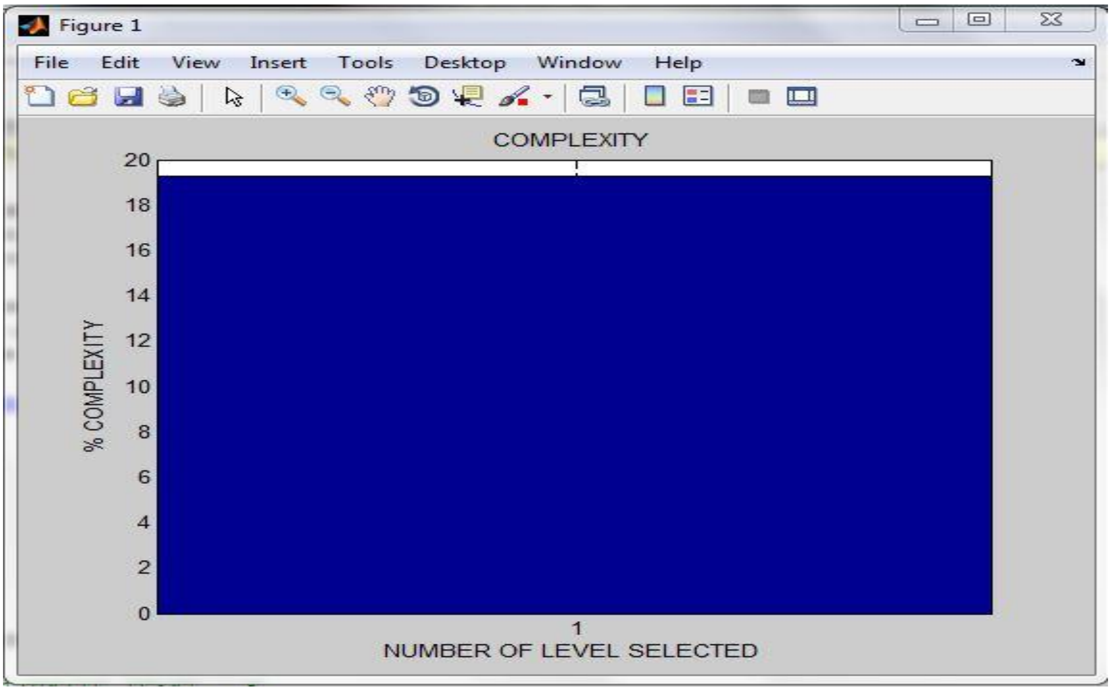


Fig. 4.23 Graph of Apriori algorithm complexity which is equal to the number of iterations taken

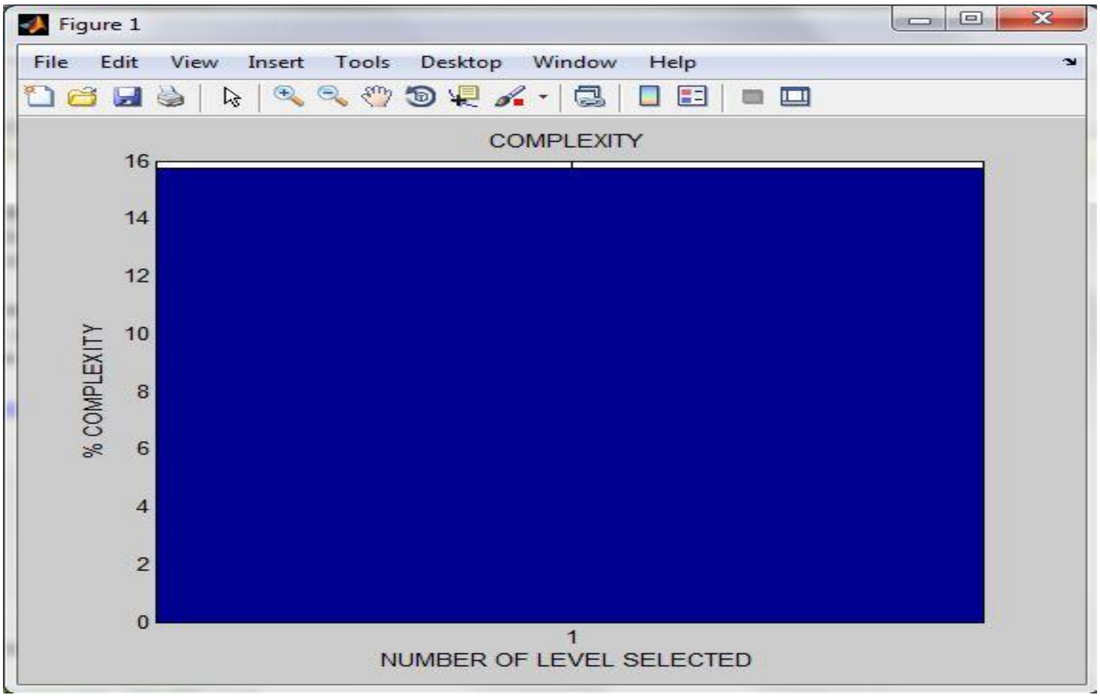


Fig. 4.24 Graph of Proposed technique complexity which is equal to the number of iterations taken

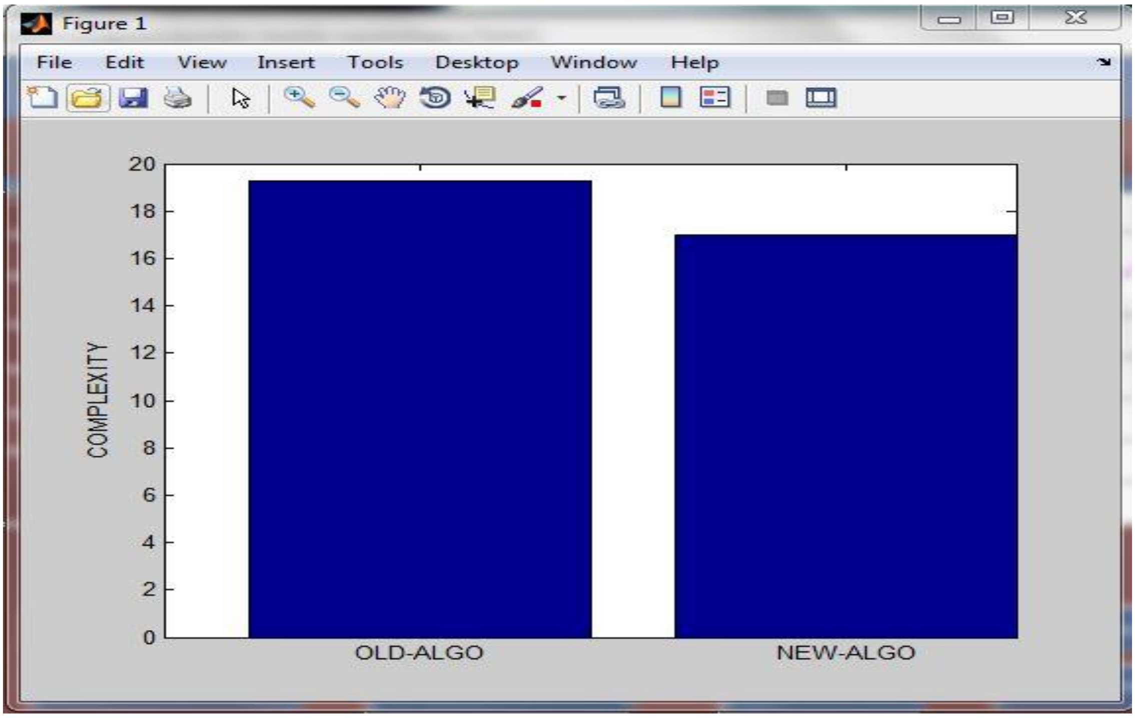
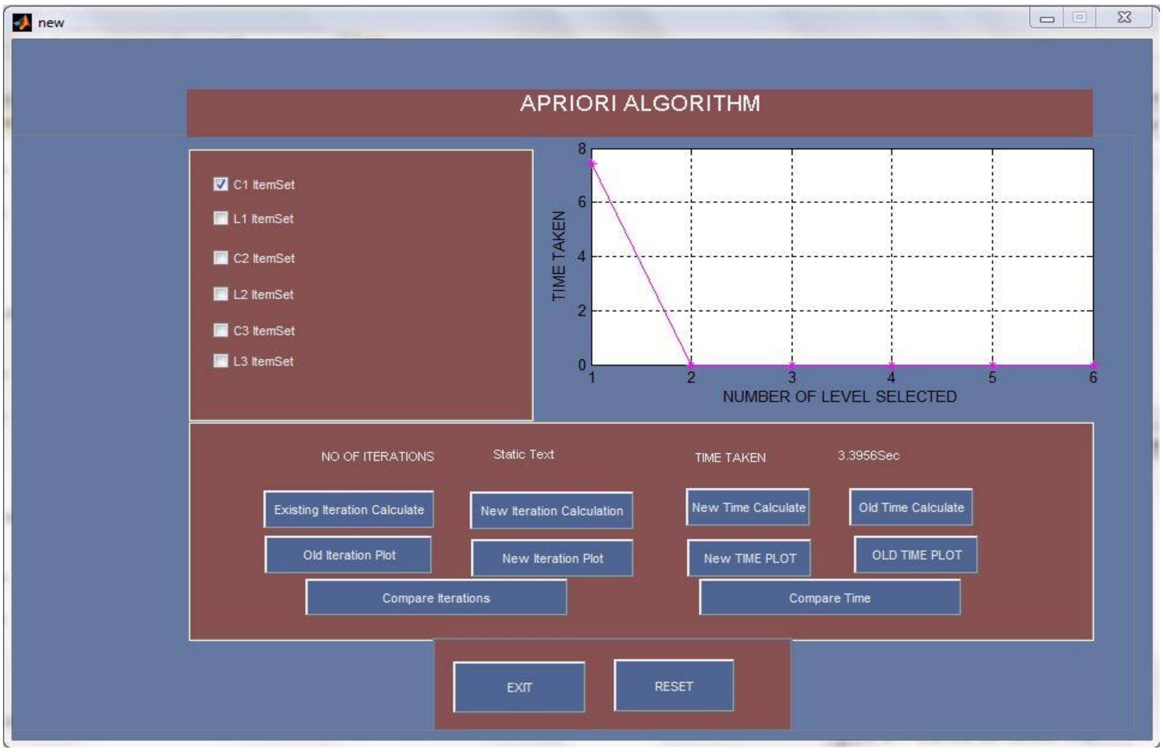


Fig. 4.25 Graph comparison of both algorithms on the basis of complexity which is equal to the number of iterations taken



☑ C1 ItemSet
□ L1 ItemSet
□ C2 ItemSet
□ L2 ItemSet
□ C3 ItemSet
□ L3 ItemSet

NO OF ITERATIONS Static Text

TIME TAKEN 3.3956Sec

Fig. 4.26 Old time calculation only for C1 item set in Apriori algorithm

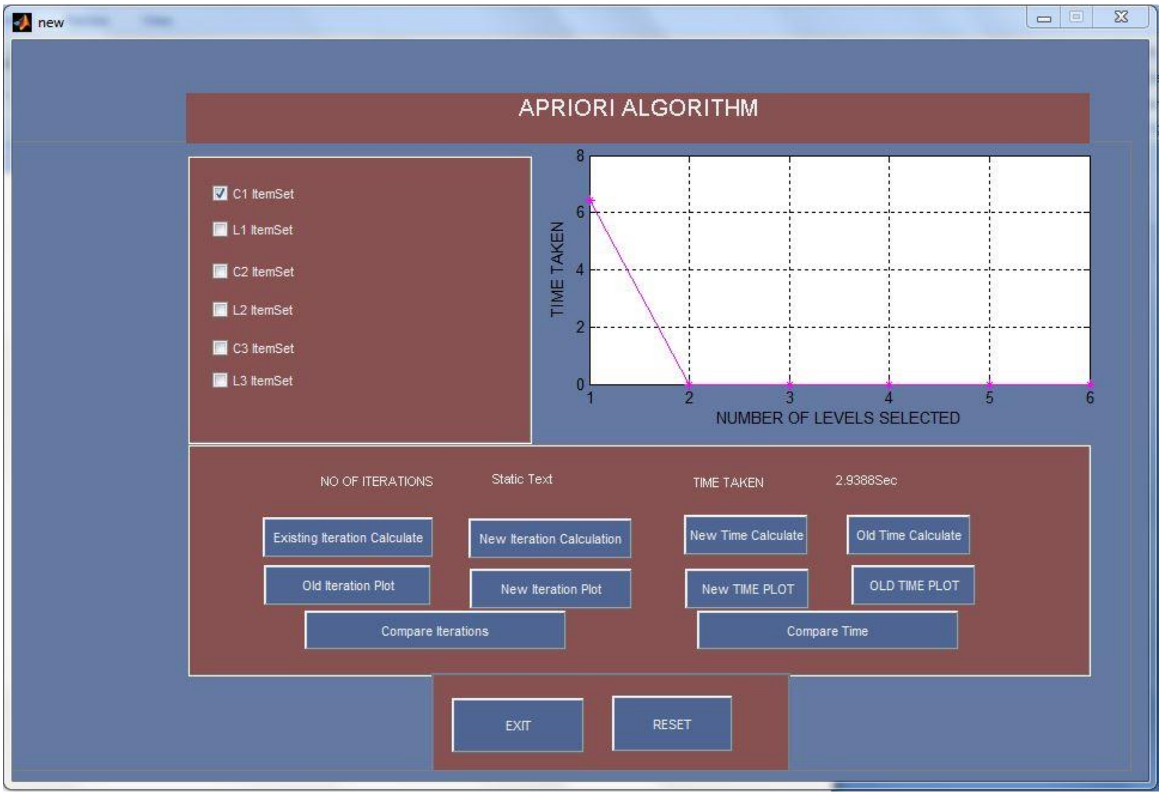


Fig. 4.27 New time calculation only for C1 item set in proposed technique

Time taken by proposed technique to calculate C1 item set is shown in figure 4.27 with the line graph for time.

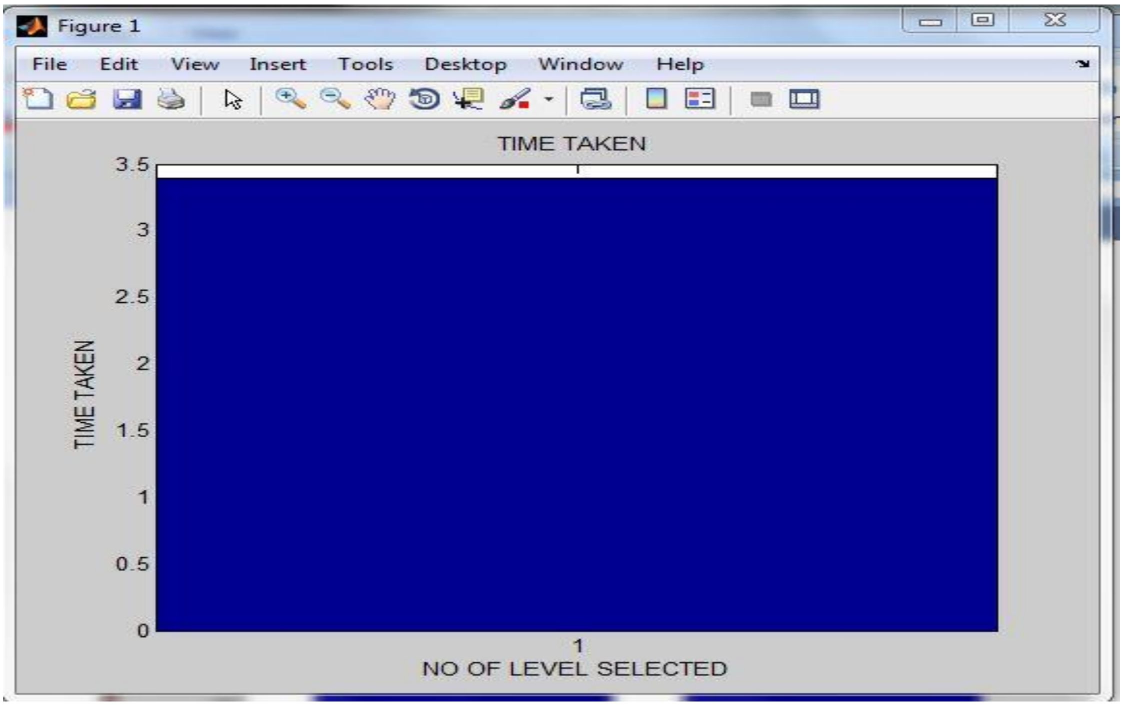


Fig. 4.28 Old time calculation graph plot for only C1 item set

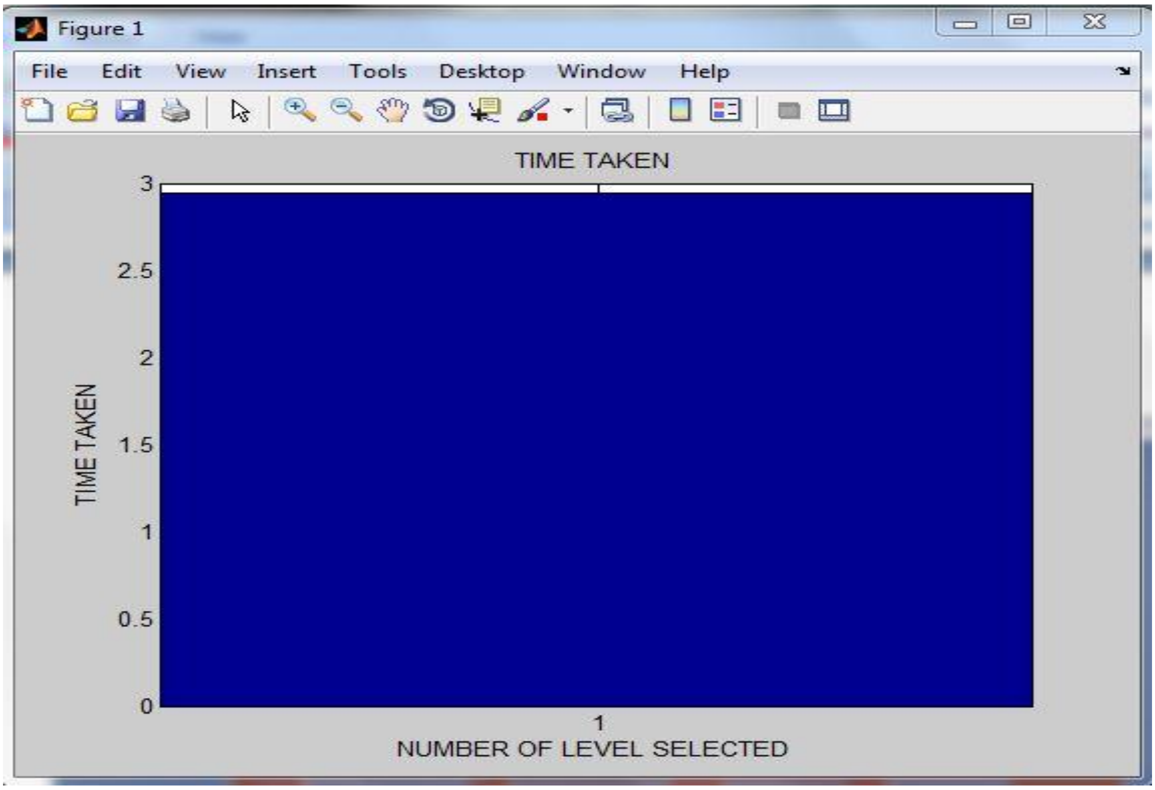


Fig. 4.29 New time calculation graph plot only for C1 item set

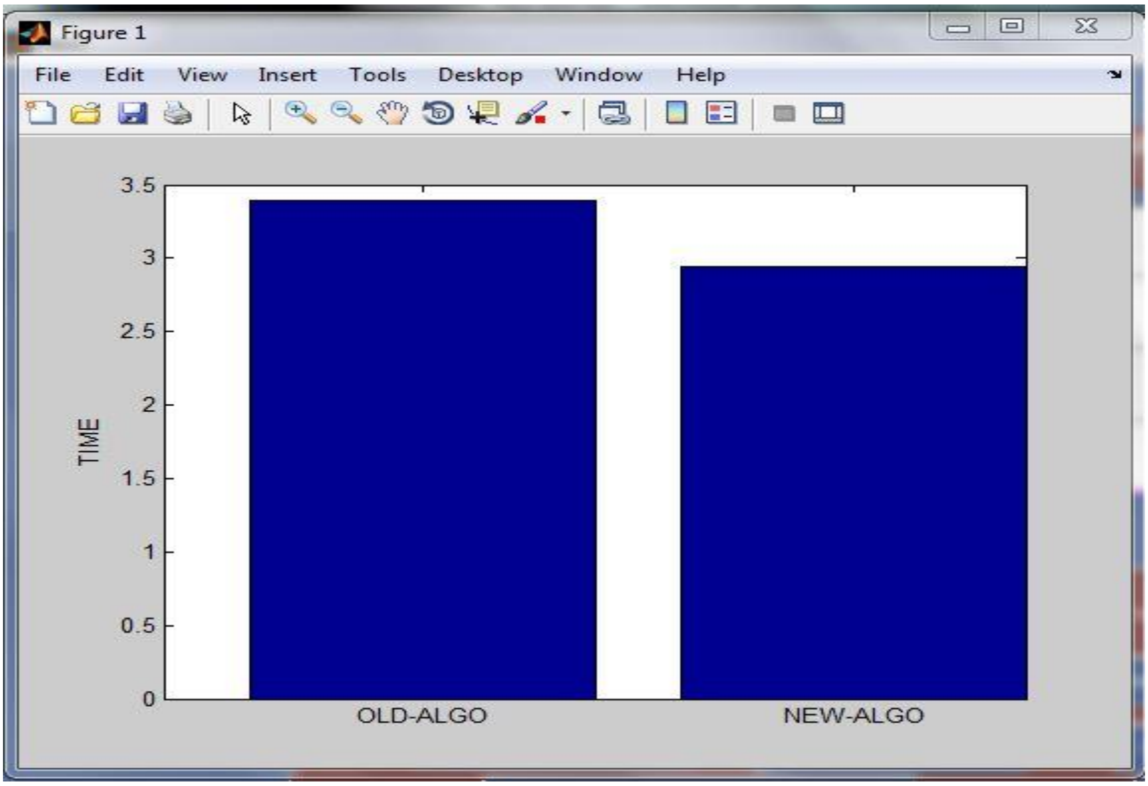
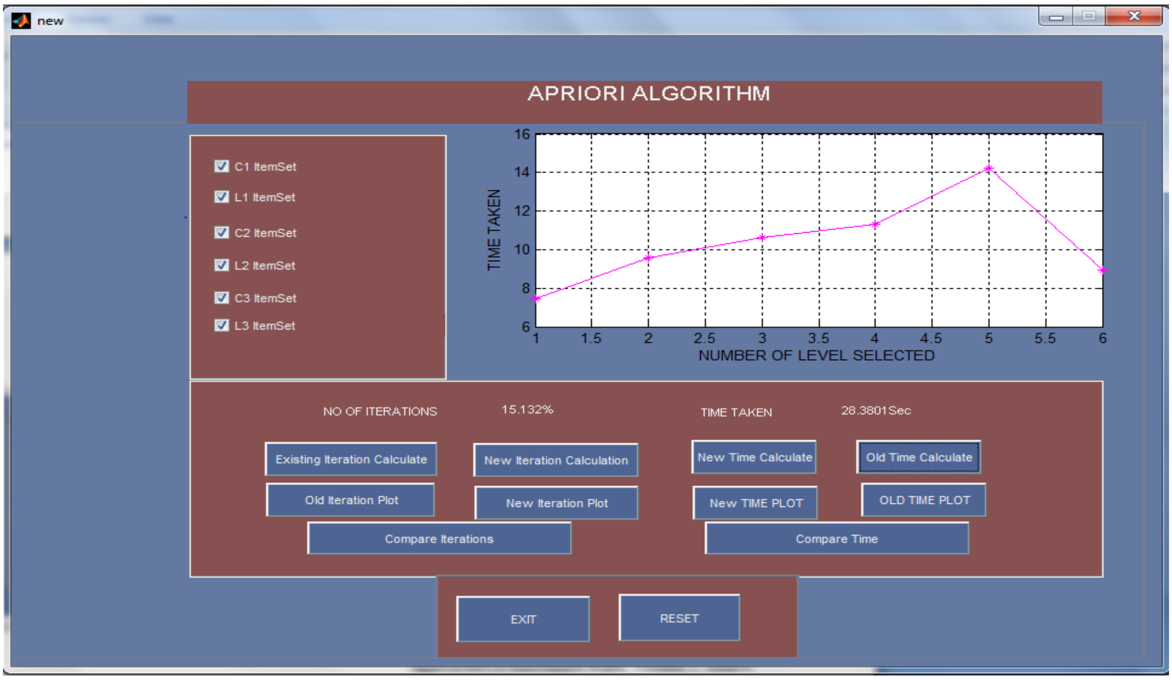


Fig. 4.30 Comparison on the basis of time only for C1 item set calculation

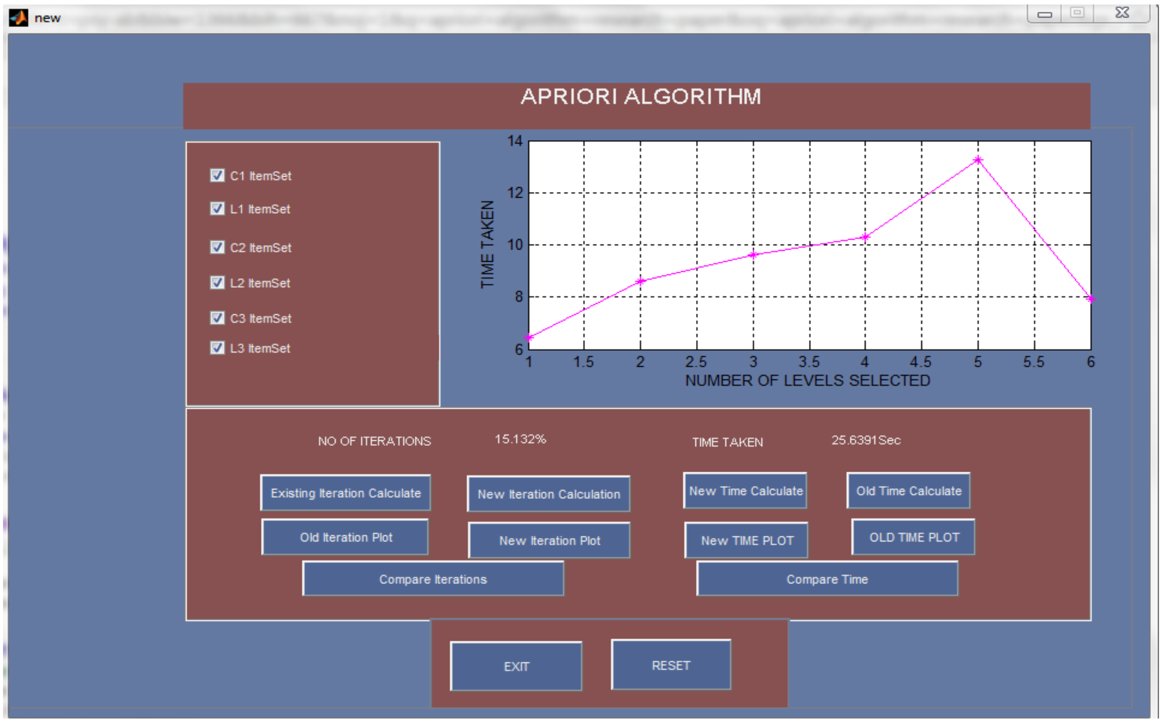


☑ C1 ItemSet
☑ L1 ItemSet
☑ C2 ItemSet
☑ L2 ItemSet
☑ C3 ItemSet
☑ L3 ItemSet

NO OF ITERATIONS 15.132%

TIME TAKEN 28.3801Sec

Fig. 4.31 Old time calculation in Apriori algorithm for calculating all frequent item sets



☑ C1 ItemSet
☑ L1 ItemSet
☑ C2 ItemSet
☑ L2 ItemSet
☑ C3 ItemSet
☑ L3 ItemSet

NO OF ITERATIONS 15.132%

TIME TAKEN 25.6391Sec

Fig. 4.32 new time calculation for calculating all frequent item sets using proposed technique

Time taken by apriori algorithm and proposed technique to calculate all frequent item sets with line graph is shown in figure 4.31 and figure 4.32 respectively.

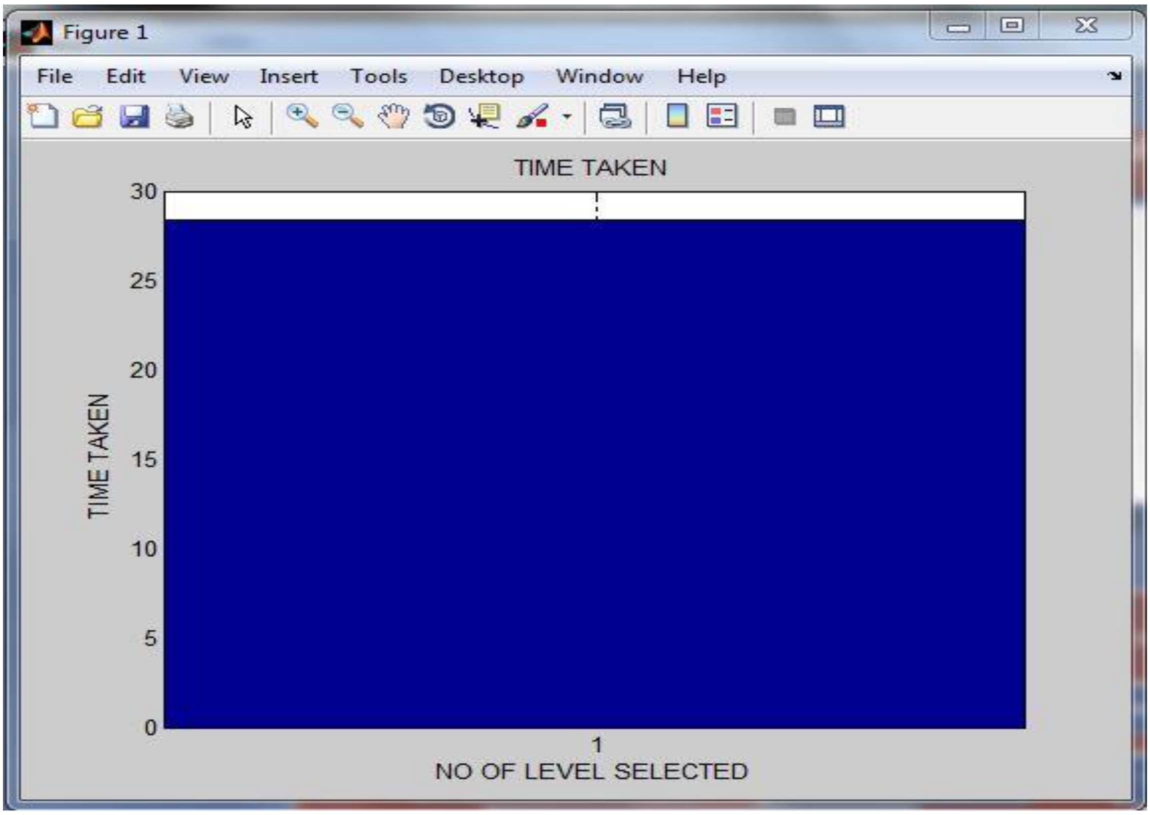


Fig. 4.33 Graph for old time calculation using Apriori algorithm

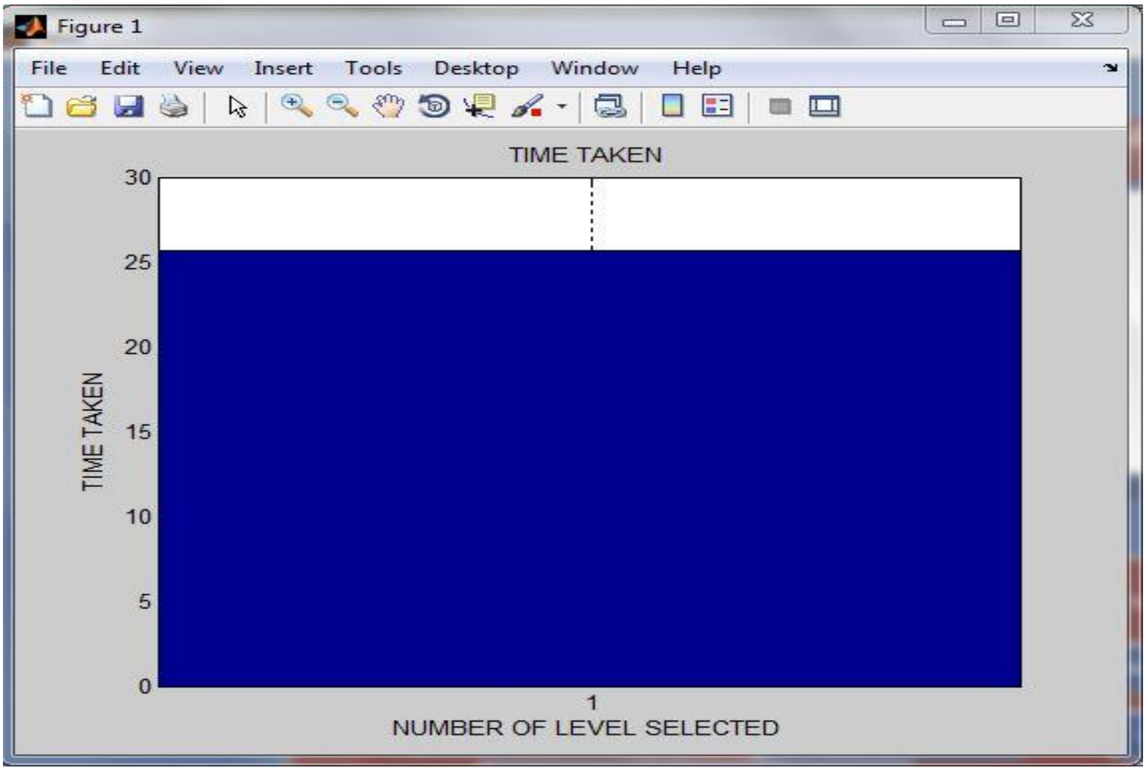


Fig. 4.34 Graph of New time calculation using proposed technique

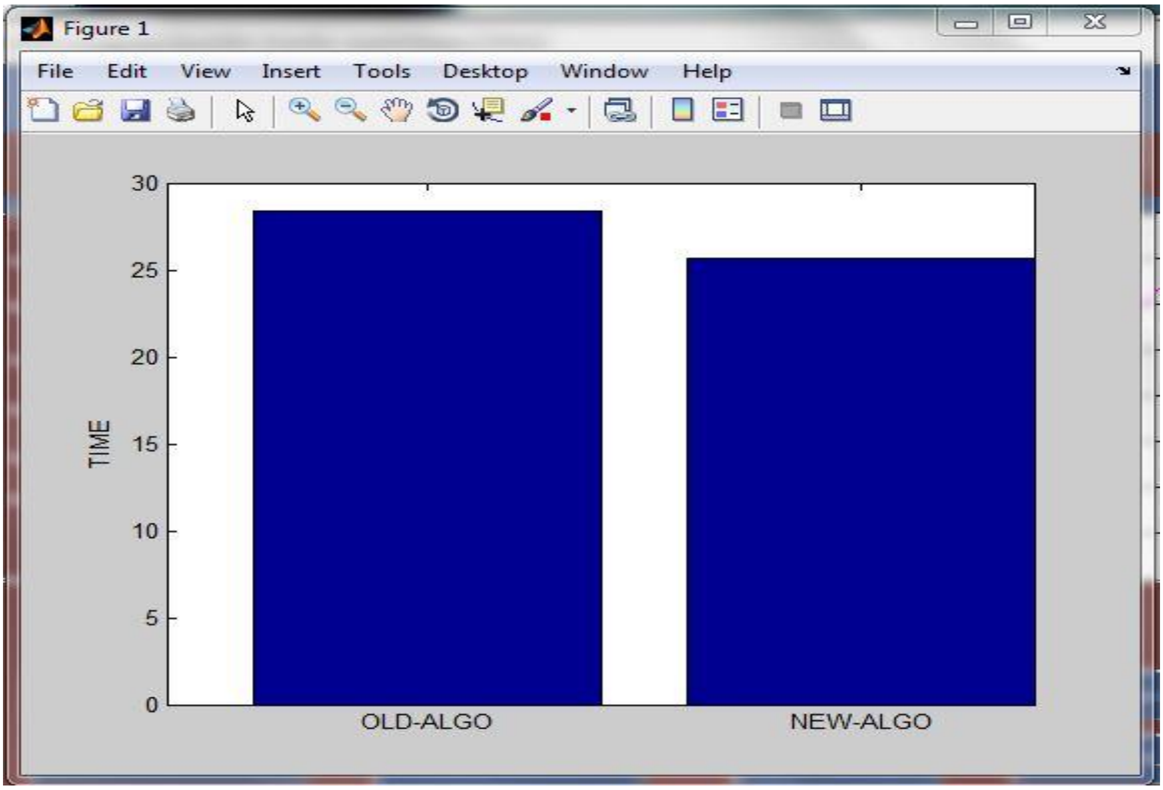


Fig. 4.35 Comparison on the basis of time

CHAPTER 5
CONCLUSION AND FUTURE WORK

CONCLUSION

Association rule mining is one of the important techniques of research area used in the data mining. For generating association rules, the Apriori algorithm is considered as the most efficient one. Apriori algorithm with a different technique is discussed which include enhancement in Apriori algorithm with the use of transposition of database. Further improvement is done in transposition technique using some different calculations of minimum support count. This approach improves the performance as compared to apriori algorithm by reducing the number of transactions and time taken to generate the frequent item sets. Results are taken using MATLAB in the form of graphs and also the detailed steps of both algorithms are stored in excel sheet for further comparison purpose.

FUTURE WORK

As mentioned above generation of frequent item sets using Apriori algorithm involves much complexity and time. These issues are tried to be resolved by proposed transpose technique. In future if any better approach than that we purposed, will help in improving the results in more accurate manner.

REFERENCES

[1] Chanchal Yadav, Shuliang Wang, Manjot Kumar (2011), “An Approach to Improve Apriori Algorithm Based On Association Rule Mining”, IEEE.

[2] D. Gunaseelan, P. Uma (2012), “An improved frequent pattern algorithm for mining association rules”, International Journal of Information and Communication Technology Research Volume 2 No. 5.

[3] Gagandeep Kaur and Shruti Aggarwal, “Performance Analysis of Association Rule Mining Algorithms”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 8,August 2013.

[4] Jugendra Dongre, Gend Lal Prajapati, S.V.Tokekar (2014), “The role of apriori algorithm for finding the association rules in data mining” IEEE.

[5] Jiawei Han, Micheline Kamber, Jian Pei (2012) *Data Mining : Concepts and Techniques*, Morgan Kaufmann Publishers, USA.

[6] Jaishree Singh, H. R. (2013). “Improving Efficiency of Apriori Algorithm Using Transaction Reduction ”.International Journal of Scientific and Research Publications.

[7] Lamine M. Aouad, Nhien-An Le-Khac, Tahar M. Kechadi (2009), “Performance study of distributed apriori-like frequent item-sets mining”, springer, Knowledge Information System.

[8] Mohammed AI-Maolegi, BassamArkok(2014), “An Improved Apriori Algorithm for Association Rules”, International Journal on Natural Language Computing Vol.3, No.1.

[9] M. Dimitrijevic, and Z.Bosnjak,“Discovering interesting association rules in the web log usage data”, Interdisciplinary Journal of Information, Knowledge, and Management, 5, 2010, pp.191-207.

[10] Marghny H. Mohamed • Mohammed M. Darwieesh (2013), “Efficient mining frequent itemsets algorithms” Springer.

[11] N. Padhy, Dr. P. Mishra, and R. Panigrahi (2012), “The Survey of Data Mining Applications and Feature scope”,IJCSEIT (International Journal of Computer Science, Engineering and Information Technology), Vol.2, No.3, June.

[12] Qiang Yang, Y. (2011), “Application of Improved Apriori Algorithm on Educational Information”.Fifth International Conference on Genetic and Evolutionary Computing.

[13] RuPeng Luan*, SuFen Sun, JunFeng Zhang, Feng Yu, Qian Zhang (2012), “A Dynamic Improved Apriori Algorithm and Its Experiments in Web Log Mining”, IEEE.

[14] Rahul Mishra, Abha choubey (2012), “Discovery of Frequent Patterns from Web Log Data by using FP-Growth algorithm for Web Usage Mining”, Volume 2, Issue 9, September, IJARCSSE (International Journal of Advanced Research in Computer Science and Software Engineering).

[15] Shuo Yang, (2012). “Research and Application of Improved Apriori Algorithm to Electronic Commerce”. 11th International Symposium on Distributed Computing and Applications to Business, Engineering & Science .

[16] Zhuang Chen, Shibang Cai, Qiulin Song and Chonglai Zhu (2011), “An Improved Apriori Algorithm Based on Pruning Optimization and Transaction Reduction”, IEEE.

WEBSITES

http://in.mathworks.com/products/matlab/

http://www3.cs.stonybrook.edu/~cse634/lecture_notes/07apriori.pdf

http://en.wikipedia.org/wiki/Apriori_algorithm

APPENDIX

LIST OF ABBREVATIONS

KDD	Knowledge Discovery Database
TID	Transaction Identifier
ARM	Association Rule Mining
FP	Frequent Pattern
MATLAB	MATrix LABoratory
GUIDE	Graphical User Interface Development Environment
GUI	Graphical User Interface