LOVELY
PROFESSIONAL
UNIVERSITY

Optimization of Apriori Algorithm to reduce the number of transactions for Association Rule in Data Mining.

A Dissertation Proposal

Submitted

By

Abhishek Pathak

TO

Department of Computer Science

In partial fulfilment of the Requirement for the

Award of the Degree of

Master of Technology in Computer Science

Under the guidance of

Nikhil Sharma

Asst. Professor

(April 2014)

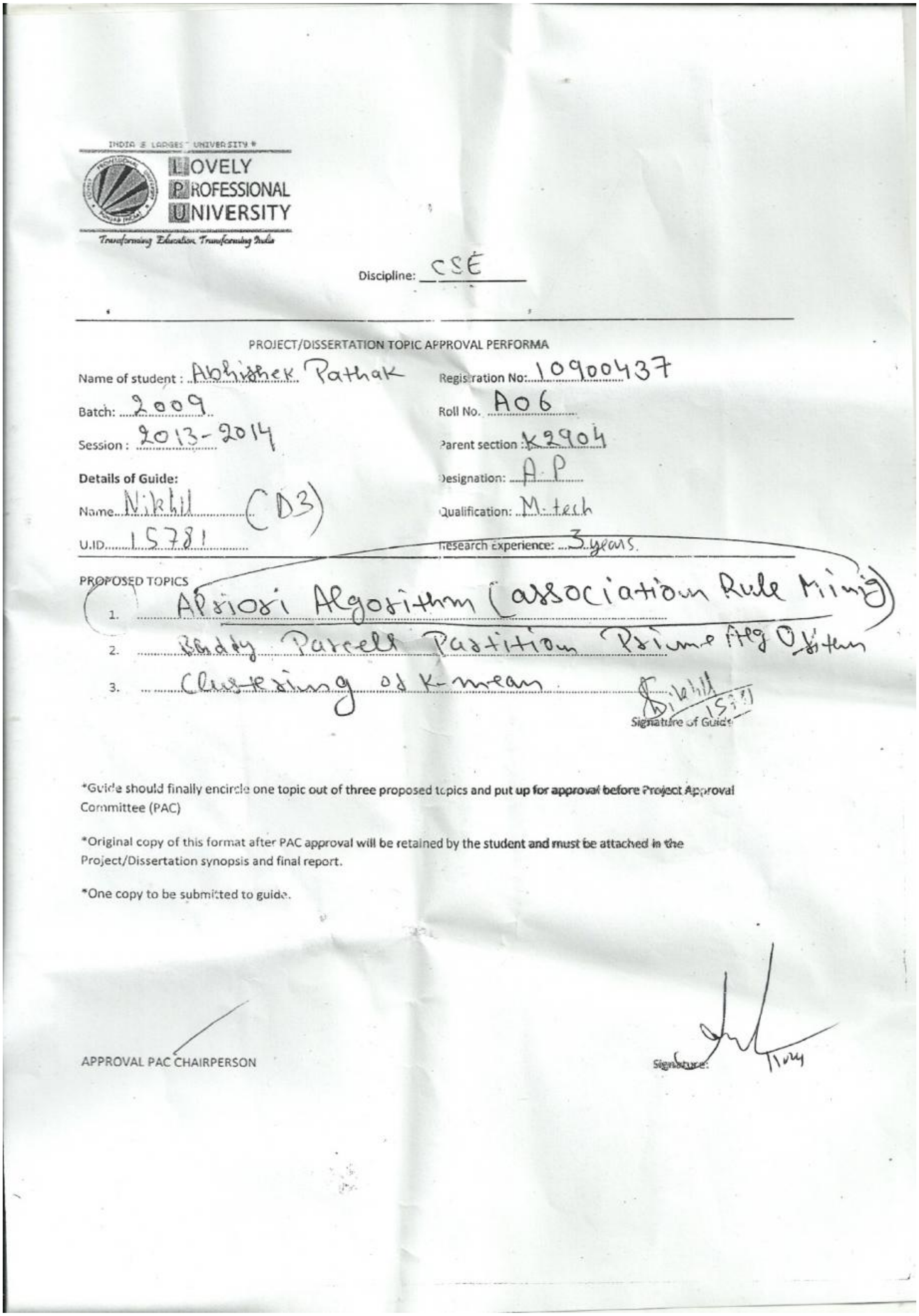
DECLARATION

I hereby declare that the dissertation proposal entitled "**Optimization of Apriori Algorithm to reduce the number of transactions for association rule in data mining"** in data mining "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

Regn. No. 10900437



INDIA'S LARGEST UNIVERSITY

LOVELY PROFESSIONAL UNIVERSITY

Transforming Education, Transforming India

Discipline: CSE

PROJECT/DISSERTATION TOPIC APPROVAL PERFORMA

Name of student : Abhishek Pathak	Registration No: 10900437
Batch: 2009	Roll No. A06
Session : 2013-2014	Parent section : K2904
Details of Guide:	Designation: A.P
Name Nikhil (D3)	Qualification: M.tech
U.ID 15781	Research Experience: 3 years.

PROPOSED TOPICS

1. [✓ Apriori Algorithm (association Rule Mining)]
2. Baddy Parcell Partition Prime Alg Orithm
3. Clustering of k-mean

Nikhil 15781

Signature of Guide

*Guide 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 synopsis and final report.

*One copy to be submitted to guide.

APPROVAL PAC CHAIRPERSON

Signature: 1/14

ABSTRACT

To extract the useful information from the large amount of database data mining Approach is being used.in the large database .association rule have been used to extract the most frequently item set from the huge amount of database .the various type of algorithm are used to extract the most frequently item set from the database. Apriori algorithm is the one in which much more database scans are available and due to this the system sometime halts and giving the most frequently item set after more time. All the input /output overhead that are being generated during repeated scanning the entire database decrease the performance of the CPU, memory and input output overhead. In this paper we proposed new novel approach to reduce the number of transactions in apriori algorithm.

Keywords:-frequent item set, association rule mining, KDD

CERTIFICATE

This is to certified that Abhishek Pathak has completed M.TECH Dissertation proposal title "**Optimization of apriori algorithm to reduce the number of transaction for association rule in data mining**"uder my guidance and supervision.to the best knowledge the present work is the result of his original investigation and study.no part of dissertation has ever been submitted for any other degree or diploma.

The dissertation proposal is fit for the submission and partial fulfilment of the conditions for the award of M.TECH computer science & Engg.

Date

Signature of Adviser
Name: Nikhil Sharma
UID: 15781

AKNOWLEDGEMENT

I would like to place on record my deep sense of gratitude to Assistant Prof. **Nikhil Sharma** department of computer science engineering, LPU, Jalandhar, India for his generous guidance, help and useful suggestion.

His assistance is very much beneficial for me to carry out the process of research in this field. Key improvement in the proposed research work would not be possible without the valuable suggestion and feedback of my guide. I would also like to thanks to all my colleagues and classmates for their co-operation and support.

Abhishek Pathak

Table of content

List of Tables

List of Abbreviations

Abbreviations	**Full Name**
KDD	KNOWLEDGE DISCOVERY IN DATABASE
SUP	SUPPORT
CONF	CONFIDENCE
SOT	SIZE_ OF _TRANACTION
ARM	ASSOCIATION RULE MINING
RAAT	REDUCED APRIORI ALGORITHM
C_K	SET OF CANDIDATE K-ITEM
L_K	FREQUENT K-ITEM
MATLAB	MATRIX LABORATORY

List of Figures

Chapter 1
Introduction

1.1 INTRODUCTION

The role of data mining (KDD) is very important in many of the field such as the analysis of market basket, classification, etc. If talk about 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. Retail store also used the concept of association rule for managing marketing, advertising, and errors that are presented in the telecommunication network.

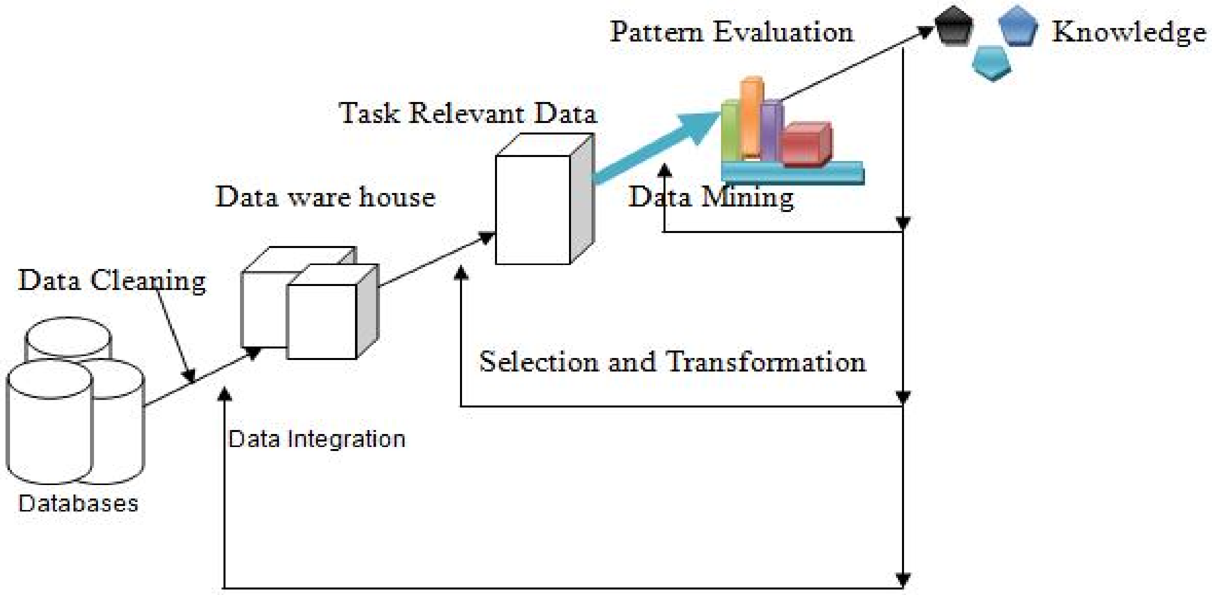


Fig: 1 KDD Steps

As we know Information Technology is growing and databases generated by the companies or organizations are becoming huge. Organization like telecommunications, banking, marketing, transportation, manufacturing etc. To defined valuable data, it is important to explore the databases and efficiently and completely. Data mining which helps to identify information in large amount of databases. KDD is the process designed to generate data that show the well-defined relationship between the variables. KDD has been very interesting topic for the researchers as it leads to automatic discovery of useful patterns from the database. This is also called knowledge discovery from the large amount of database. Many techniques have been developed in data mining amongst which primarily Association rule mining is very important which results in association rules. These rules are applied on market based, banking based etc. for decision making.

The relationship among the items done by association rule. All type of relationship between items is totally based on the co-occurrence of item.

The knowledge discovery in data can be achieved by following steps:

- **Data Cleaning:** In this step, the data that is irrelevant and if noise is present in database than both irrelevant and noise data removed from the database.

- **Data Integration:** In this step the different type of data and multiple data sources joined in a common source.

- **Data Selection:** In this stage, the applicable of data analyse that what data what type of data retrieved from the collection of data.

- **Data Transformation:** In this stage, the selected data is changed into accurate form for the procedure of data mining.

- **Data Mining:** This is the important step in which the technics that is used to extract the pattern is clever.

- **Pattern Evaluation:** In this step severely needed pattern represent acquaintance are based on measures parameters.

- **Knowledge Representation:** This is the final step in which knowledge is visually represented to the user. Knowledge representation use visualization technics to help understanding of user and taking the output of the KDD**.**

Data Mining System: Classification of data mining system is as follow (Manne, 2011):

1. **Classification according to the type of data source mined:**

   In this type, data mining system defined according to the data to be managed or handled such as spatial, Multimedia data etc.

2. **Classification according to the data model:**

   In this type, data mining system defined according to the data to be managed or handled such as relational and object oriented database.

3. **Classification according to the king of knowledge discovered:**

   In this type, the data Ming system is defined according to the functionality of the knowledge established like clustering and classification.

4. **Classification according to the mining techniques to be used**

   In this type, the data mining system is defined according to the different type of approaches such as machine learning, neural network, genetic algorithm etc.

Issues in Data Mining:

Algorithm of data mining represent techniques that have every so often existed for many years but now have recently been practically as scalable and reliable tool that time and break previous classical statistical method moreover before KDD develop into a trusted discipline, many type of issue pending to be addressed. All the issues are not exclusive and are not in sequence (Ford Lumban, 2009).

Security and Social Issues: The important issue of data mining is security with any type of data collection that is synchronize and purposed to be apply for decision making beside this when data is composed for user behaviour understanding, customer profile etc. the huge amount of confidential information is gathered or stored.

Performance Issues

It refers to the various types of issues:

- **Efficiency and scalability of data mining algorithms:** efficiently neglect the information from the large amount of database data mining algorithm must be efficient and scalable**.**

- **Parallel, Distributed, and incremental mining algorithms:** The factor such as large amount of database and KDD complexity method inspires the growth of parallel and distributed data mining parallelism

1.2 Frequent Item sets and Association Rules

These two term i.e. frequent item set as well as association rule mining both have the widely scope in the research area . ARM also used by Apriori algorithm. AR equation P+Q, both P and Q are the item set, it means that In database the transaction that contain item P also contain item Q.

The establishment of AR proposed by Agrawal in 1993. AR is mandatory data mining model research extensively by the data mining and database establishers. Initially AR is used in market basket analysis to find out the most frequently item set from the large amount of the database.

$$\text{Copy} \rightarrow \text{Pen} \qquad [\text{sup}=4\% \text{ and conf}=100\%]$$

Let we have data set of text document :{ transaction}

Document treated as “student” keyword:-

d1: play, college, bag,

d2: college, bag

d3: City, college

d4: cricket, volleyball

d5: Spectator, Basketball, Player

d6: volleyball, Team

d7: City, Game, cricket, Team

Transaction containing p that is the set of item in I. the implication form of the association rule: $P \to Q$, where P, $Q \subset I$, *and* $P \cap Q = \emptyset$an item set is a set of items.

E.g., A = {Copy, Pen, Book} is belong to item set.

E.g., {Copy, Pen, and Book} is a 3-itemset.

Rule strength measures in association rules

Support: if finding frequent item set without correlation between the item we use support that contains both the items together i.e. P∪Q both the items purchased together by customer

Support = Prob (P∪Q).

Confidence: confidence used for correlation between the items i.e. Association rule. Confidence includes the concept i.e. if transactions that contain P also contain Q.

Conf = Prob (Q | P)

Support and Confidence:-

Support count support count of the item is denoted by A.count, in a data set T is the number of transactions in T that contain A. Assume T has n transactions.

Then**,**

$$support = \frac{(A \cup B).count}{n}$$

$$confidence = \frac{(\mathrm{A} \cup B).count}{A.count}$$

Relationship between unrelated data is done by association rule. Two component support and confidence used in this whole algorithm properly. Association rule holds in the number of transaction. With the help of %age of T (transaction) has demonstrate the support concept. considered the support of an item 50% the meaning of this statement is that only 50% items are purchased together most frequently by customer

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

The conditional probability included in confidence, Y will also be present.

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

Association rule is to generate all association problems having support and confidence with respect to specified minimum support and minimum confidence

Two steps included in ARM:-

1) **Determined the item set that is used frequently:** the frequency of item sets counted from each transaction in databases should not be less than min_ support.

2) **With the help of frequents item set generate strong Association Rule:** the rules must meet the min_support and the min_confidence criteria and with the help of that item set that is frequently used generating the strong association rule

1.3 The Apriori Algorithm

To find out the most frequently used item set from the large amount of database the used algorithm is apriori algorithm that is establish by R.Agrawal and R.srikant in 1994 to identify the frequent item set for association rule mining. The apriori algorithm based on the technics that used the prior knowledge of the frequent item set. The algorithm also known as the level wise search. For exploring (k+1) item set, firstly we use k-item set. In the 1st scan of the database the 1-item set is generated with the help of matching the criteria of min_support and min_confidence. The generated item set named as L1. With the help of L1 find out the 2-item set and so on. For finding L_k, L_k require full scan on the data set. Bu using apriori property the efficiency of the level wise generation of the frequent item is overcome. Describing the example and characteristics firstly (Jiawei Han and Micheline Kamber).

Two main steps included in the Apriori Algorithm:

- Determine all set of item that have satisfying minimum support criteria
- For creating rule used frequent item set

E.g. frequently used item set {shirt, shoes, paint} [support $=\frac{3}{7}$] and one rule from the frequent item set shoes → paint, shirt [sup $= \frac{3}{7}$, conf $=\frac{3}{3}$]

1) Join:-join the item one by one in level wise sequence after joining creating items are candidate items

2) Prune: Discard items set if support is less than minimum threshold value and discard the item set if its subset is not frequent.

In some areas the performance of these algorithm decrease like genome data in which much more scan are available to finding the frequently item set, due to this type of problem system some time halts and giving the response after a long time.

Algorithm example

In table 1 transactional database presented. Considered that the support count is 3(Jaishree Singh, 2013)

Step 1: initial step covert database in to the desired database i.e. size of transaction .number of transaction and items are giving in the first step.in the first step minimum support criteria neglected, min_support criteria comes when goes to the next step i.e. in the 2nd step.

Step 2 .After initial step write down the items in level wise way and write the frequency of that item for calculating the candidate generation rule. That is write down the transaction id, items as well as SOT (size of the transaction).with the help of candidate item set we can generate the 1-itemset called L1.

Tid	**Items**
K1	p1,p3,p7
K2	p2,p3,p7
K3	p1,p2,p3
K4	p2,p3
K5	p2,p3,p4,p5
K6	p2,p3
K7	p1,p2,p3,p4,p6
K8	p2,p3,p4,p6
K9	p1
K10	p1,p3

X

Tid	**Items**	**SOT**
K1	p1,p3,p7	3
K2	p2,p3,p7	3
K3	p1,p2,p3	3
K4	p2,p3	2
K5	p2,p3,p4,p5	4
K6	p2,p3	2
K7	p1,p2,p3,p4,p6	5
K8	p2,p3,p4,p6	4
K9	p1	1
K10	p1,p3	2

C1

Itemset	**Sup_count**
p1	5
p2	7
p3	9
p4	3
p5	1
p6	2
p7	2

L1

Items	**Sup_count**
p1	5
p2	7
p3	9
p4	3

Tid	**Items**	**SOT**
K1	p1,p3	2
K2	p2,p3	2
k3	p1,p2,p3	3
k4	p2,p3	2
k5	p2,p3,p4	4
k6	p2,p3	2

k7	p1,p2,p3,p4	4
k8	p2,p3,4	3
k10	p1,p3	2

Itemset	**Sup_count**
p1,p2	2
p1,p3	4
p1,p4	1
p2,p3	7
p2,p4	3
p3,p4	3

X1

C2

X2

Tid	**Items**	**SOT**
K3	p1,p2,p3	3
K5	p2,p3,p4	3
K7	p1,p2,p3,p4	4
K8	p2,p3,p4	3

Item set	**Sup_count**
p1,p3	4
p2,p3	7
p2,p4	3
p3,p4	3

L2

Itemset	**Sup_count**
p2,p3,p4	3

C3

Itemset	**Sup_count**
p2,p3,p4	3

Table1: **Example Apriori Algorithm**

Step 3: the number of item generated by transaction in third step called size of transaction.

Step 4: the min_support is 3 determine the 1-frequent item set

Step 5: support frequency of p5, p6, p7 < 3, and the entire not appear in L1. Neglect these data from initial database i.e. In addition, when L1 is established, currently the value of k is 2, neglect those records of transaction having SOT=1 in X. And there won't exist any elements of C2 in the records we find there is only one

Step 6: to generate a candidate item set join L1 ∞L1.

Step 7: now the transaction that is presented in the X scanned in proper way.

Step 8: the candidate 2-item set is generating after joins L1 and L2

Step 9: when L2 is created properly , we can find the record of 5 transactions (K1, K2,K4, K6, K10)are only two in X1..

Step 10: for discovering 3-item set join L2∞ L2 to generate candidate 3-item set .

Step 11: transaction that are presented in X and the support count of each candidate item set in C3 is accumulated. Use C3 to generate L3.

Step 12: therefore L3 has only 3 item sets C4 =$^{\phi}$**.**The algorithm will stop and give out all the frequent item sets.

Step 13: process will be generated for (C_K) until C_k+1 becomes null.

Advantage

- It is very easy and simple algorithm.
- Its implementation is easy**.**

Disadvantage

- It does multiple scan over the database to generate candidate set.

- The number of database passes is equal to the max length of frequent item set.

- For candidate generation process it takes more memory, space and time (Ms. Rina Raval, 2013).

1.4 Limitations of ARM

1. The Result of the ARM does not return in the reasonable time

2. ARM belongs to the absence and presence of the item in the database (Tang, P., Turkia, M., *2005).*

3. If talking about larger data set this ARM is not so efficient

4. ARM treat all items in the database equally (Hegland, M, 2003).

5. ARM used in much application such as market basket analysis, website navigation. The most frequently item set are only those items that specify the minimum support criteria.

1.4.1An Improved Association Rules Algorithm

ARM used to find out the interesting correlation between the items. In 19993 agrawal proposed the association rule. After ARM many algorithm are generated that is apriori algorithm and improvement in apriori algorithm. Han and Fu change the minimum support threshold for association rule; the algorithm that is F-P algorithm there is no need of generating the candidate item in this algorithm. Some of this algorithm very slows to show the result in reasonable time (Yaqiong Jiang, 2011).

Chapter 2

Review Literature

WanjunYu, XiaochunWang et al. (2008) they proposed the algorithm called reduced apriori algorithm with tag .that is used to reduce the redundant pruning candidate items .If the number of frequent 1-itemsets is n then the number of connected candidate two item sets is C_{n2}, while pruning operationsC_{n2} .The algorithm decreases pruning operations of candidate two-item sets and increasing efficiency and saving time. For the bottleneck: poor efficiency of counting support RAAT optimizes subset operation through the transaction tag to speed up support calculations. Experiments of results shows that RAAT outperforms original one efficiency and saving time.

Qiang Yang,YanhongHu et al. (2011)they used apriori algorithm in educational training .They found the correlation rules of course which provided the directive significance information for the curriculum .The improved Apriori algorithm digs data from educational information database, the rules generated by the digging process are helpful for managing courses, educational model and Quality Education and etc. So decision makers can provide help for making reasonable, though they arrange semester course order of the courses, and improve teaching effect of the following courses by strengthen teaching time and teachers of preliminary course(Qiang Yang,YanhongHu , 2011).

Shuo Yang (2012) shuo yang apply a data mining technics in e-commerce shopping areas, with the proper improved version of apriori algorithm. They establish the mechanism of recommendation of commodities and uplift the technique of calculating the support count and confidence level. From the application of the case, users could quickly build association rules by the improved Apriori algorithm and recommend appropriate products to customers in a timely approach. Each time when customers settle accounts, system will scan what the customer has already purchased and query this table to recommend other products to them (Shuo Yang, 2012).

K.Vanitha and R.Santhi (2011) they proposed hash-based technic for candidate generation. The number of candidate 2 -item sets generated by the proposed algorithm is in orders of magnitude, smaller than that by previous techniques thus resolving the performance bottleneck. In this approach scans the database once utilizing an improved version of apriori algorithm. They analysed that the generation of smaller candidate sets enables them to effectively trim the transaction database size at a much earlier stage of the iterations thereby reducing the computational cost for later iterations significantly. As we know frequent objects is one of the most important fields in data mining. That algorithm can achieve a smaller memory usage than the Apriori algorithm. It is well known that the way candidates are defined has great effect on running time and memory need. They have presented experimental results which shows that the proposed algorithm always outperform Apriori. (AprioriK.Vanitha and R.Santhi, 2011).

D. Gunaseelan, P. Uma (2012) for frequent pattern they proposed enhances technics for algorithm using transposition of the database with minor modification of the Apriori-like algorithm. The main advantage of the proposed technique is the database stores in transposed form and in each iteration database is filtered and reduced by generating the transaction id for each pattern. The proposed technique reduces the huge computing time and also reduces the database size.. Hence the proposed technique is very beneficial for the discovering frequent patterns from large datasets. It has been also compared the classical Apriori algorithm with an improved algorithm. It has been presented the experimental results using synthetic data, showing that the proposed algorithm always outperform Apriori algorithm (D. Gunaseelan, P. Uma, 2012).

Ms. Rina Raval[1], Prof. Indr Jeet Rajput[2,] Prof. Vinitkumar Gupta[3] (^{2013}) survey of few better apriori algorithm are shown in this paper. From these surveys they find the good ideas. Survey included that many improvements are needed basically on pruning in Apriori to improve efficiency of algorithm. After doing survey of algorithms conclusion can be given that mostly in improved Apriori algorithms the main aim was to generate less candidate sets and yet get all frequent items. In the technique of Intersection and intersection Record filter is used with the record filter approach where to calculate the support and count the common transaction that contains in each elements of candidate set. In this technique only those transactions are considered that contain at least k items. In other approach set size and set frequency are considered. Set sizes which are not greater than or equal to minimum set size support are eliminated. Improvement by reducing candidate set and memory utilization only needs to compare the count of each element of $L_{k-1}$with the count of each element (X) of C_k. If the count of the element X equals to k, then only keep X. Also the item that not appears in L_{k-1} will no longer appear in Lk so it is deleted. Trade list approach uses undirected item set graph. From this graph by considering minimum support it finds the frequent item set and by considering the minimum confidence it generates the association rule. Second last approach considers frequency and profit of items and generates association rules. Last technique suggests utilization of attributes such as profit, weight to associate with frequent item set for better information gain for user and business standpoint (Ms. Rina Raval, 2013).

Anjali Singla (2013) new parallel partition prime algorithm is proposed in this paper. Partition prime multiple algorithm addresses the shortcoming of previously proposed parallel buddy prime algorithm. For load balance new technics used in this proposed algorithm. The proposed approach for parallel frequent item set mining and load balancing reduces data complexity and the time and divide transactional database efficiently for good load balancing between the processor. The main motive behind most of the approach is to divide transaction equally to each processor.

Jaishree Singh[1], Hari Ram[2] Dr. J.S. Sodhi[3] (2013) they have described an Improved Apriori algorithm which minimize the scanning time by cutting down unusable transaction records as well as minimize the redundant generation of sub-items during pruning the candidate item sets, which can form directly the set of frequent item sets and eliminate candidate having a subset that is not frequent. The improved algorithm in this paper not only optimizes the algorithm of reducing the size of the candidate set of k-item sets, C_k but also minimize the I / O spending by cutting down transaction records in the database. The performance of Apriori algorithm is improved so that we can mine association information from massive data faster and better. Although this improved algorithm has optimized and efficient but it has overhead to manage the new database after every generation of L_k. So, there should be some techniques which have very less number of scans of database (Jaishree Singh, 2013).

Ajay Kumar (2007) in this paper author proposed a new Parallel Partition Prime Multiple Algorithm for association rules mining. Proposed algorithm addresses the shortcoming of previously proposed parallel buddy prima algorithm. Therefore it minimizes the time and data complexity. The parallel Partition Prime Multiple algorithm can be improved to provide dynamic load balancing. With the increasing data it is important to create more efficient algorithms to extract knowledge from the data. Unless the volume of data size is rising much faster than CPU execution speeds which have a strong effect on the performance of software algorithms. The performance of Parallel Computing however cannot increase linearly as the number of the parallel nodes grows (Ajay Kumar, 2007).

Swe Swe Nyein (2011) an algorithm is proposed that extract the main content from the web documents. The algorithm based on Content Structure Tree (CST). Firstly the proposed system use HTML Parser to build DOM (Document Object Model) tree from which create Content Structure Tree (CST) which can easily extract the main content blocks from the other blocks. The proposed model then introduced cosine similarity measure to evaluate which parts of the CST tree represent the less important and which parts showed the more important of the page. The proposed system can define the ranking of the documents using similarity values and also extracts the top ranked documents as more relevant to the query. Web page typically contained many information blocks. Apart from the main content blocks, it usually has such blocks as navigation panels, copyright and privacy notices, and advertisements which are called noisy blocks. These noisy blocks can seriously harm Web data mining (Swe Swe, 2011).

K.Rajkumar (2011) this paper introduced a new method of segmentation is introduced (DWS) which segments web pages based on either reappearance based technique by analysing reappearance tag patterns from the DOM tree structure of a web page. Based on the analysis of tag patterns it gave implicit nodes to segment the nested block correctly nor it will segment pages based on web layout data like <TABLE>, <DIV> and <FRAME> tags based on key pattern in the web page. If it consist of reappearance tag in tag pattern means it will segment based on reappearance based segmentation. Else it will segment based on web layout data. From that segmented block hyperlink is displayed on the mobile first and after that user select hyperlinks based on his area of interest. The interested data information alone is displayed to the user. Based on the detection of tag patterns it build implicit nodes to segment the nested block correctly. From that segmented block hyperlink is displayed on the mobile device first and then user select hyperlinks based on his area of interest (K.Rajkumar, 2011)

Shuang Lin, Jie Chen, Zhendong Niu (2012) this paper proposed a content extraction approach that joins a segmentation-like scheme and a density-based scheme. In their approach they designed structures called BLE&IE blocks to gather related contents or noises. Next they used this density-based technique and redundancy removal to obtain the final content. Based on their scheme a tool called Block Extractor was developed. Experimental results on their data set showed that this novel technique was effective and robust compared to three other density-based method. Density-based method in content extraction whose task was to extract contents from Web pages are commonly used to obtain page contents that are critical to many Web mining applications. But traditional density-based method cannot effectively manage pages that consist of short contents and long noises. To overcome this problem they proposed a content extraction approach (Shuang Lin, 2012)

Manne suneetha (2011) this paper aims at organizing the web search results into clusters facilitating quick browsing options to the browser providing an excellent interface to results precisely. Suffix tree clustering creates comparatively more informative and accurate grouped results. Image Repetition is a basic problem during image searching in any search engine. This can be minimized by using the L-Point Comparison algorithm a specially worked out method in field of Information Retrieval systems which is also discussed with a practical example. Search result clustering and Web mining as a whole have very promising perspectives; constantly improving Information storage and exchange media allow us to record almost anything. From our project, the user can easily access the results in the form of clusters so that he can browse the relevant contented cluster. Avoidance of repetition of images is an excellent we can further study to face the problem with almost every image search engine (suneetha, 2011).

Ford Lumban Gaol (2010) this paper proposed web log sequential pattern mining by using Apriori-all algorithm. They called as Apriori-all Web Log Mining. The experiments have been conducted base on the idea of Apriori-all algorithm which firstly stores the original web access sequence database for storing non-sequential data. The experiment observation have given with analysis on further refinement. From the results of experimental analysis, we can conclude that the greater the number of combinations is produced, the less the number of users who perform a combination of these, while the fewer number of combinations generated then it is likely the number of users who perform a combination of these will be even greater (Ford, 2010).

Vivek Arvind, B Swami Nathan, J Viswanathan. K. R. (2010*)* In this paper they proposed an intelligent recommendation model which that utilises (a) Boosted item based collaborative filtering for the efficient rating of predicted items and (b) Association rule mining technique for making a personalised recommender system for the target user. This give improvement in the overall web recommendation precision. They have suggested a framework of web personalization technique using a recommender system which gets a major contribution from the association rule mining that recommends personalized items to the active users based on the web objects predicted by boosted item based collaborative filtering, suggested to overcome the bottleneck of sparse data collected from the user profiles and proved our suggestion to be highly reliable than the existing collaborative filtering techniques (Vivek,2010).

Rui Xie (2012) In the paper, based on Latent Dirichlet Allocation (LDA) they proposed a model to construct a lexicon for sentiment analysis task which is domain independent. Through experiments they compare our generated lexicon with some widely used lexicons and with trivial lexicon construction algorithm. The experiments represents their technique is competitive and flexible. In this paper they proposed a probabilistic model to model the latent topic/sentiment information of review corpus. The lexicon the model builds is domain-specific and corpus related. It exploits more information than the general lexicon. We use our lexicon to do the task of sentiment classification and have achieved good result. The TSTM model gives distribution of words over different sentiment labels, which gives a flexible way to determine the real sentiment role of words (Rui Xie, 2012).

Miguel Dar (2009) This paper presents a recommendation-based web content mining model used for the navigation data from an university community. This recommendation is based on an offline module that does the grouping of the web documents using through a the Bisecting K Means algorithm and vector space model and an online module that selects the closest cluster and documents of the query document (actual navigation).Tasks for pre-processing the web documents recommendation strategies experiments and a supervised validation are presented. The analysis suggests that the relationship between the query document and recommendations are good for near half (Miguel, 2009).

Chapter 3
Present Work

3.1 Problem Formulation

Data mining is one of the important research areas in today's database technology. The important concept to be to discover valuable patterns from a large collection of data for users. In KDD mining association rule is one of the important research technics. The real difficulty shown by association rule to determine the correlation among sales of different items from the analysis of a large set of super market data. For generating association rule apriori algorithm is the best algorithm. In Apriori like algorithm, the step wise process of determined sets of all frequent item sets is in a combined form, frequent item sets. Apriori algorithm identifies all frequent k-item sets, denoted as L_k. C_k is the set of candidate *k*-item sets obtained from L_k-1, which are suspected frequent *k*-item sets. For every transaction, the candidate k-item C_K is generated, support count also increased by $1/|D|$. The support of candidate *k-item set* are > or equal to user defined minimum support, they immediately become frequent *k*-item sets. At the end of level *k*, all frequent item sets of length *k* or less have been discovered. Significantly the performance affected due to the database continually read the candidate. Item set with all transaction records of the database. The Apriori algorithm is generally applied on the large datasets like on the data set of super markets to generate the association rule. The association rule are generated on the super market dataset are on the basis of number of transitions. The items which are sailed who many time and with which item it is sailed. The number of transitions defined the threshold value of confidence. The threshold value of confidence will help to derive the most frequent data items association rules. The large databases scans are required and lot many transitions are required to generate the association rules. This approach will consume time and system resources. The Apriori algorithm's efficiency will be enhanced if the number of transactions will be reduced. The number of transactions, if reduce then the time required to generate the association rule will also be less. So, in this work, we work on to reduce the number of transactions of Apriori algorithms to generate the association rule.

3.2 Objectives of Research

1. To identity the mostly used algorithms for the association rule
2. To select the most efficient association rule algorithm among all the algorithms
3. To identify the number of transactions and time required by the apriori algorithm to generate the association rule
4. To enhance the Apriori Algorithm, so that less number of transactions are required to generate the association rule
5. The proposed enhancement will be implemented in MATLAB by taking super market dataset
6. The results of proposed technique and previous technique will be compared graphically

3.3 Scope of the Study

Association Rule Mining is the essential part of KDD. Apriori algorithm establish association rule in the database. This algorithm is not so efficient due lots of scans on database. If the data set is huge, then more time filter to scan the database. In this work, we will propose an enhancement in apriori algorithm which will reduce the transaction size as well as removing the redundant candidate item set. Use the redundant generation of sub-items during pruning the C_K item sets that can form directly the set of frequent item sets and eliminate C_K having a subset that is not frequent. When the numbers of transaction are reduced to generate the association rule, efficiency of the apriori algorithms will be enhanced. The less time will be required to generate the association's rules from the large datasets. This will leads to fewer loads on the system resource and less time is required to generate the association rule from the large datasets.

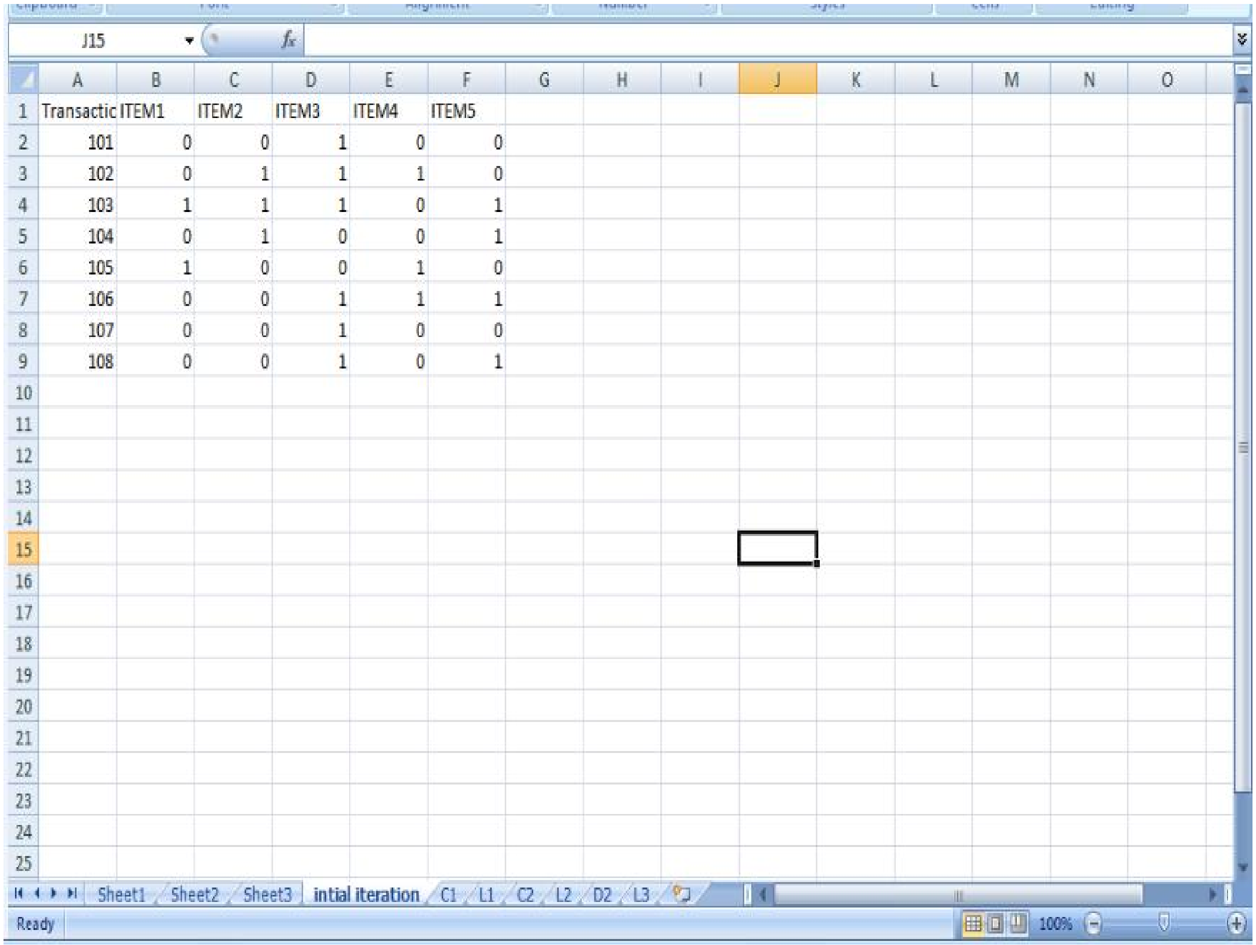
3.4 Research Methodology

Apriori is an iterative method of level wise search, applying k-Item for searching (k+1) item set. First of all algorithm find out the frequent 1-item set, that is defined as L1.L2 is find out by L1. Apriori algorithm introduces "Apriori characteristic" which means that non-empty subset of a collection of frequent item sets must be frequent. Using the Apriori characteristic, non-frequent item sets in the set of candidate item sets “C_k” will be pruned. In this work, we will define one more threshold value which is called “number of items in the transactional database”. If the item that is not satisfy the minimum support criteria than that item neglect from the database. The approach will reduce the number of scans on the dataset. If the number of scans is reduced, less time is consumed to generate the association rules. As, number of transactions are directly proportional to time consumption.

The proposed idea will be implemented in MATLAB which is widely used in all areas of research universities, and also in the industry. MATLAB is beneficial for mathematics equations(linear algebra) moreover numerical integration equations are also solved by MATLAB It is also a programming language, and is one of the simplest programming languages for writing mathematical programs.it has various type of tool boxes that are very beneficial for optimization and so on.

Chapter-4

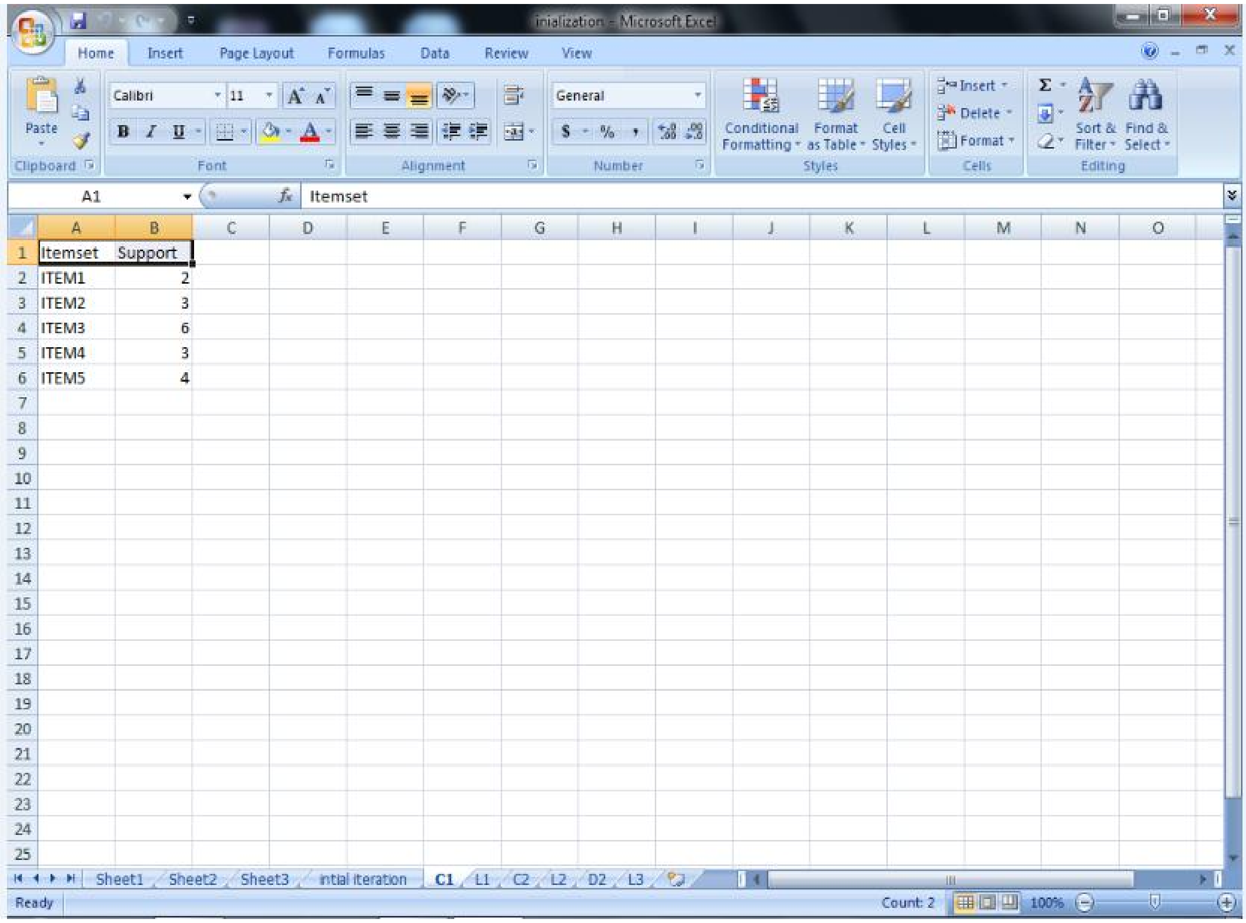
Results and Implementation



Transactic	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

Fig2: Dataset Defined

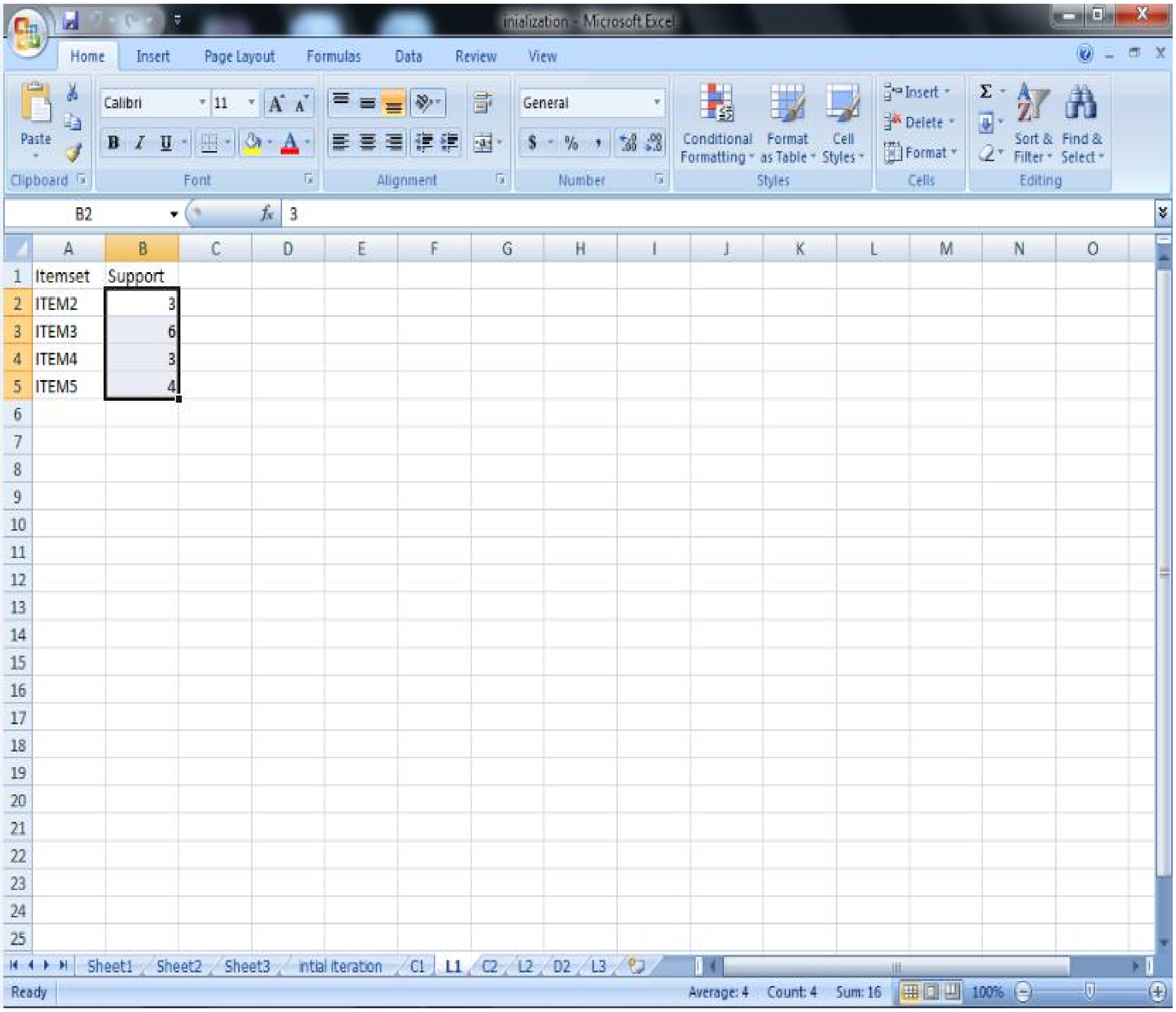
As illustrate in the figure 1, the dataset is defined of the super market. On this dataset Apriori algorithm is applied to generate association rules.



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

Fig3: Item Counts

As illustrate in the figure 2, the item count is defined on the basis of Apriori algorithm



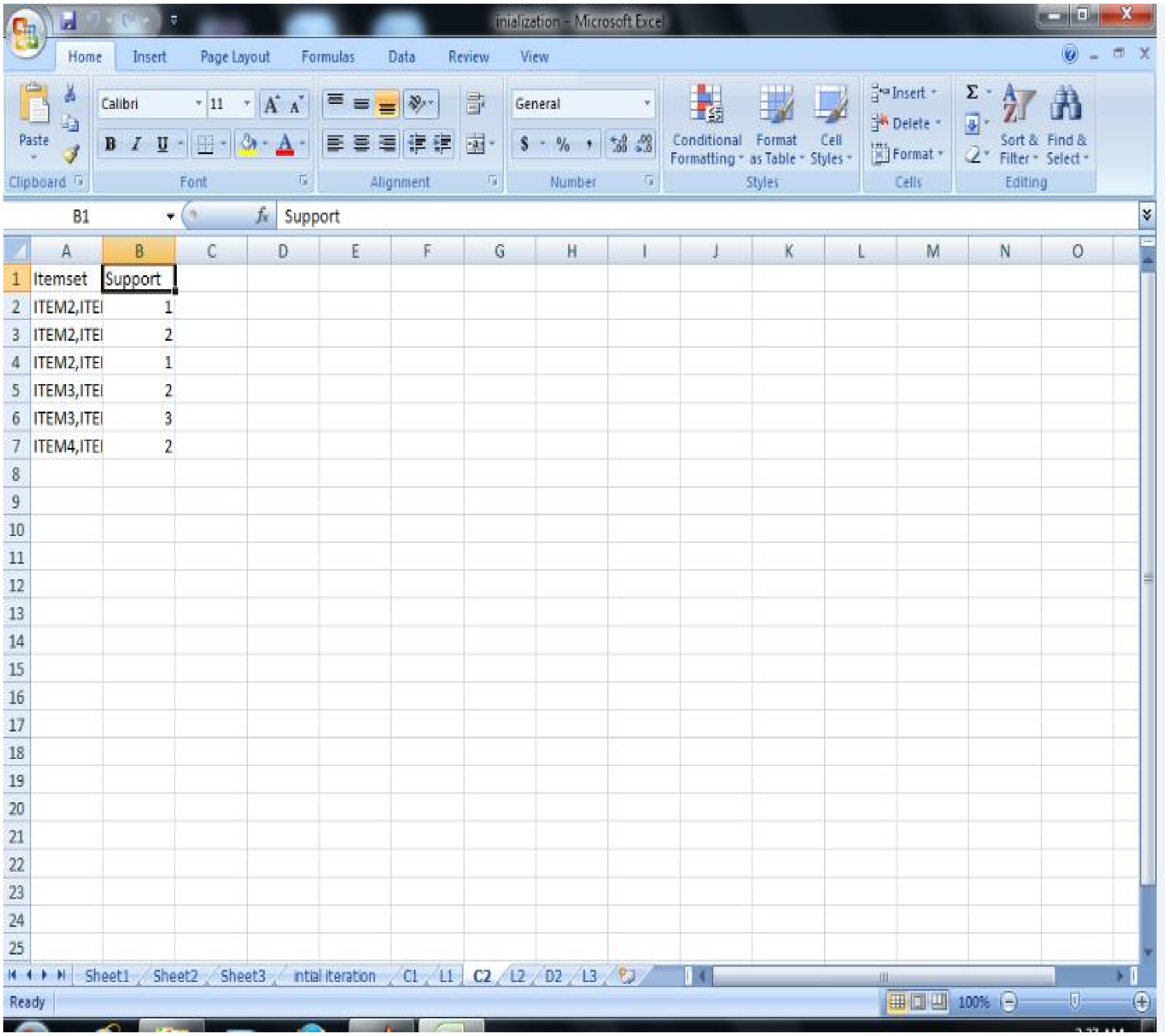
B2 | f_x 3

Itemset	Support
ITEM2	3
ITEM3	6
ITEM4	3
ITEM5	4

Average: 4 Count: 4 Sum: 16

Fig 4: Count support value

After the Item count support value is fixed. The items which have minimum value than support count will be deleted from the list as shown in above figure.

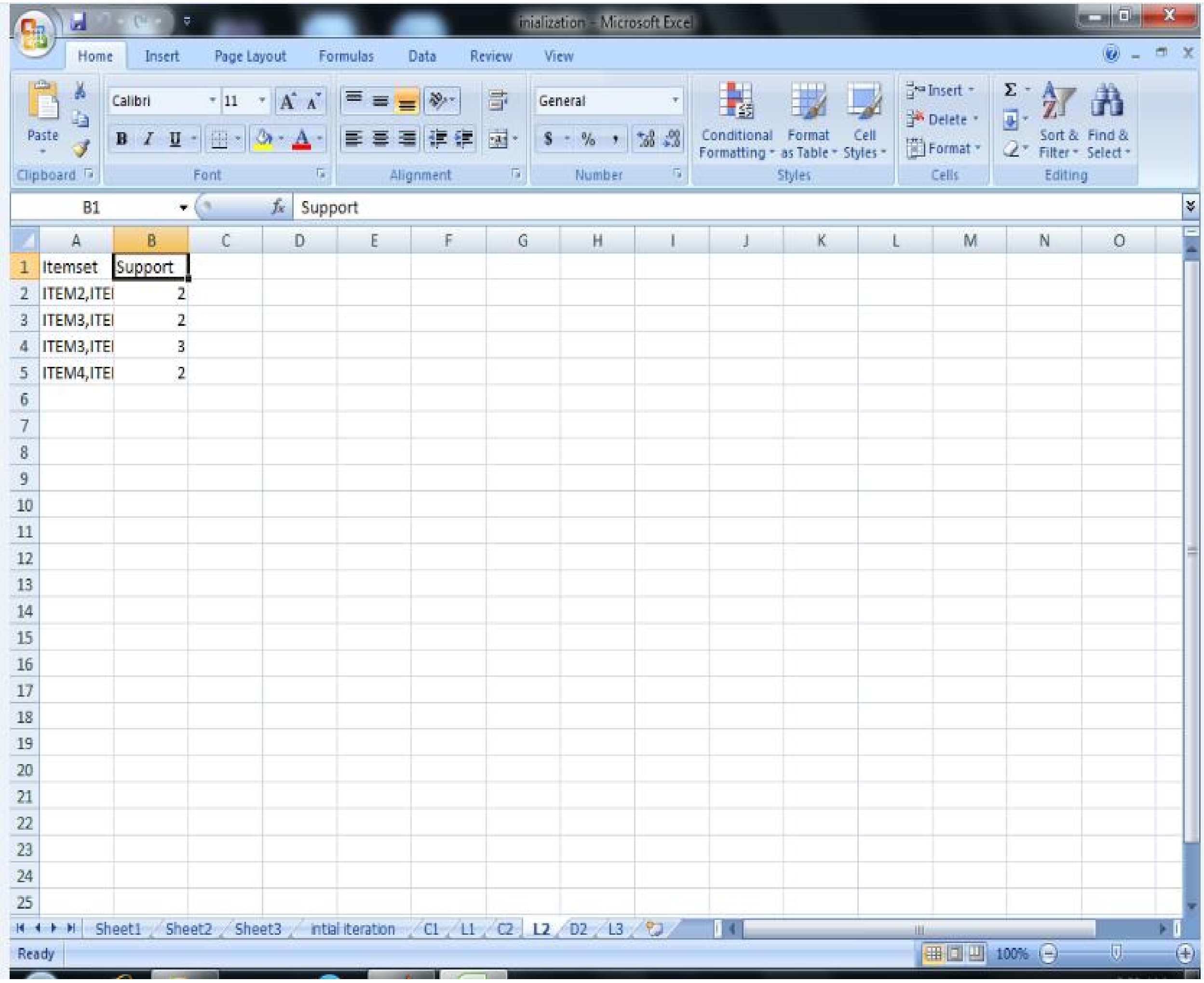


B1 Support

	A	B
1	Itemset	Support
2	ITEM2,ITE	1
3	ITEM2,ITE	2
4	ITEM2,ITE	1
5	ITEM3,ITE	2
6	ITEM3,ITE	3
7	ITEM4,ITE	2

Fig5: Item combined

As illustrate in figure items are combined on the basis of Apriori algorithm. After combination of items, association rules generate.

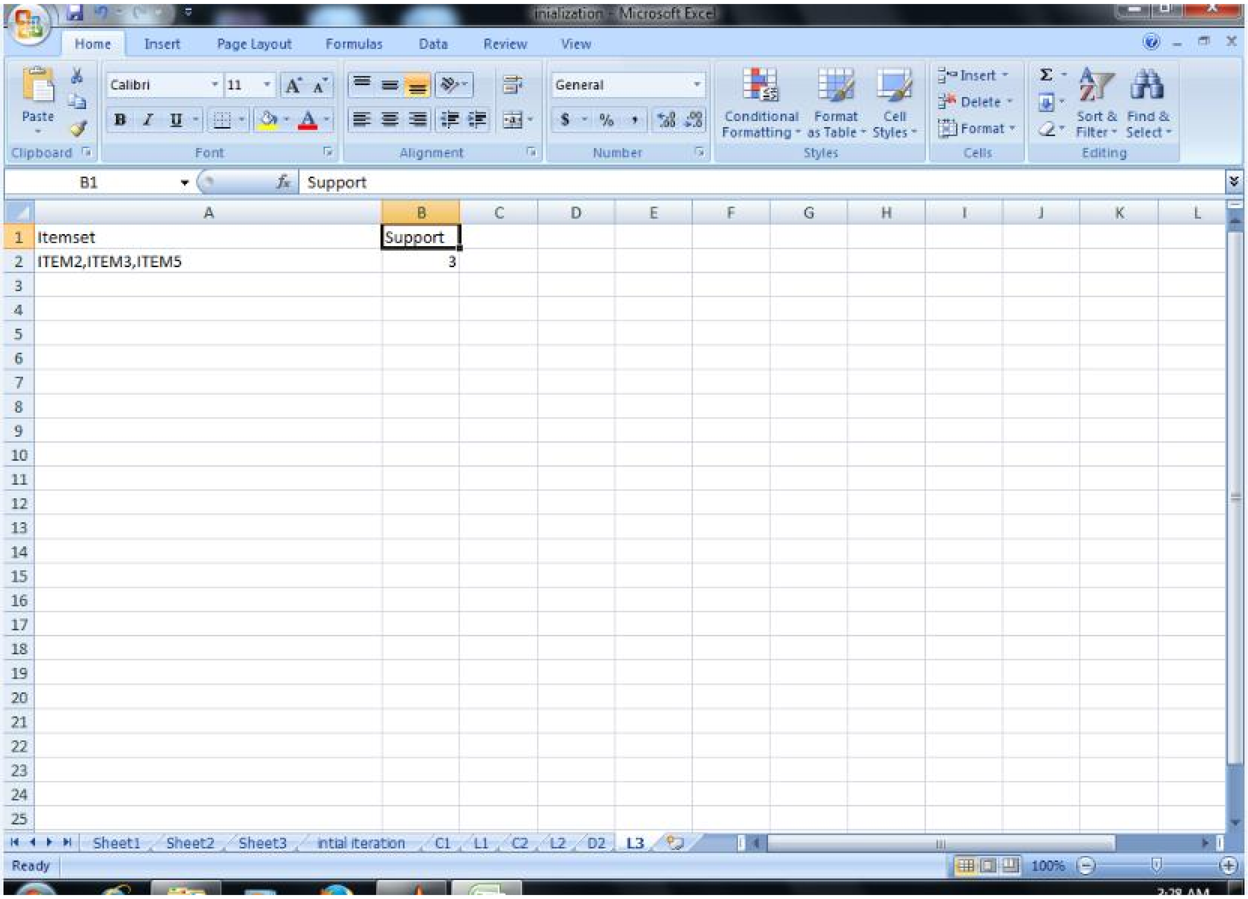


B1 f_x Support

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

Fig6: Item count Again

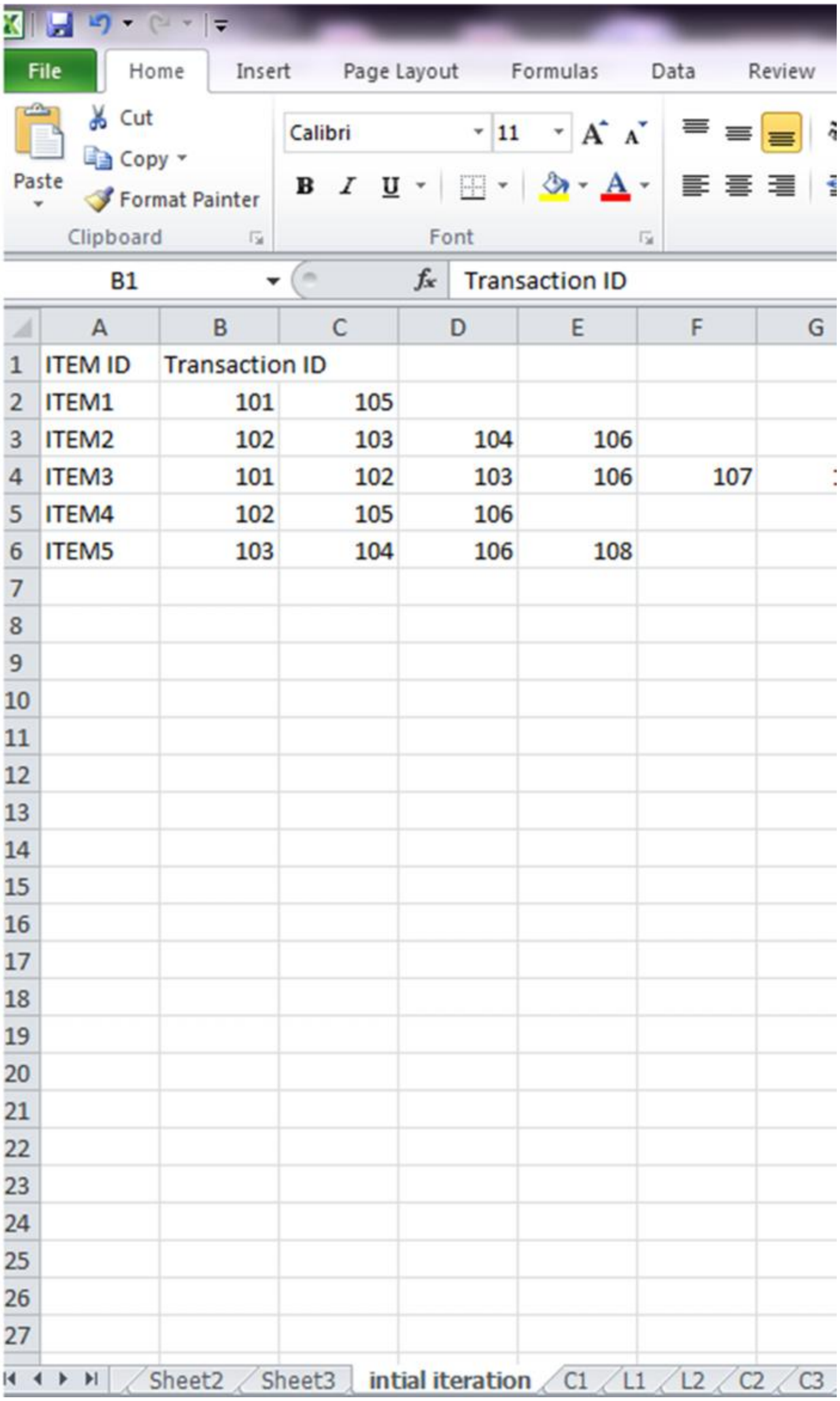
Here two items are combined and item is count on the basis of two items now. The again item which has value less than minimum support count will be deleted from the list.



Itemset	Support
ITEM2,ITEM3,ITEM5	3

Fig7: Generation of Association rules

After the deletion of minimum support count, items are combined after generation of association rules.

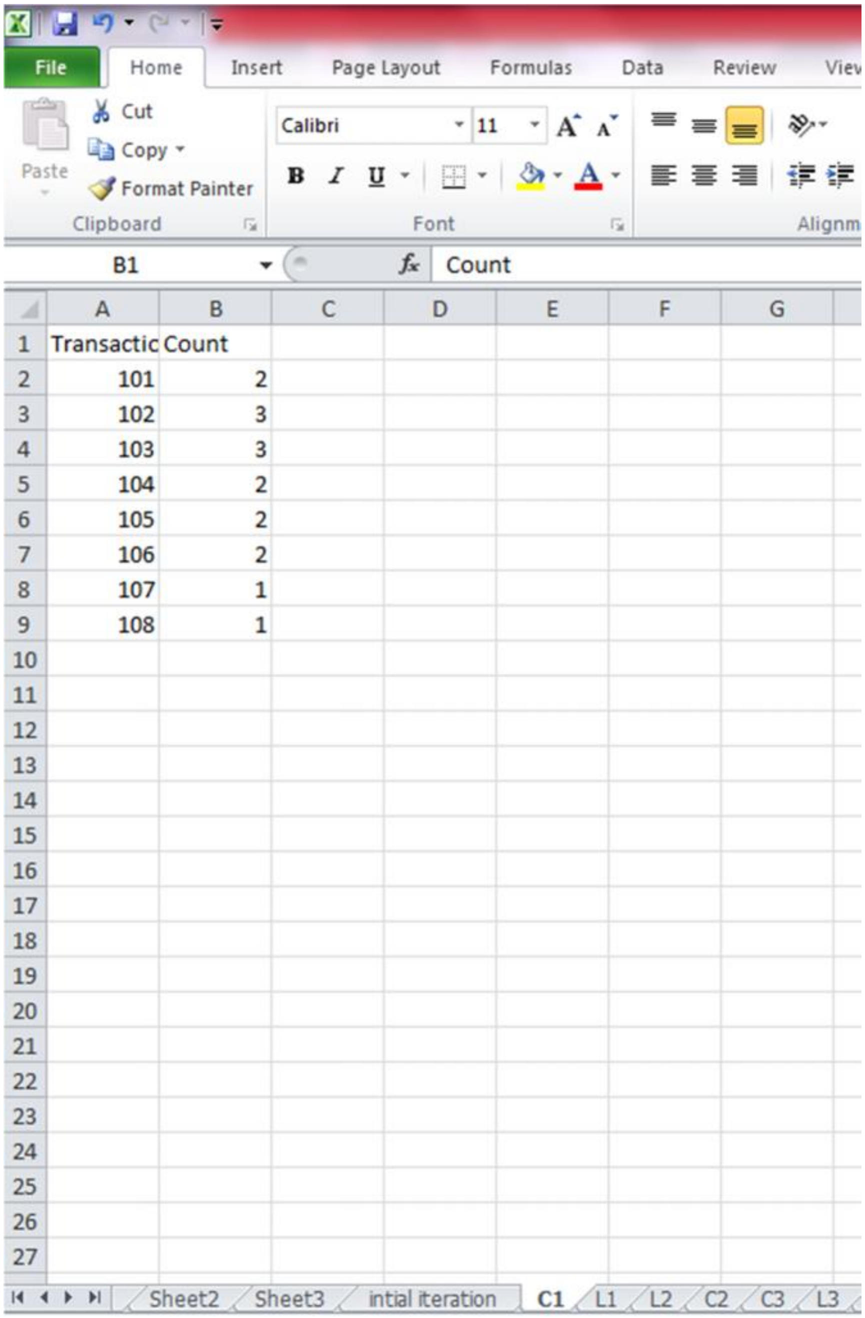


B1 fx Transaction ID

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

Fig8: Transpose of the dataset

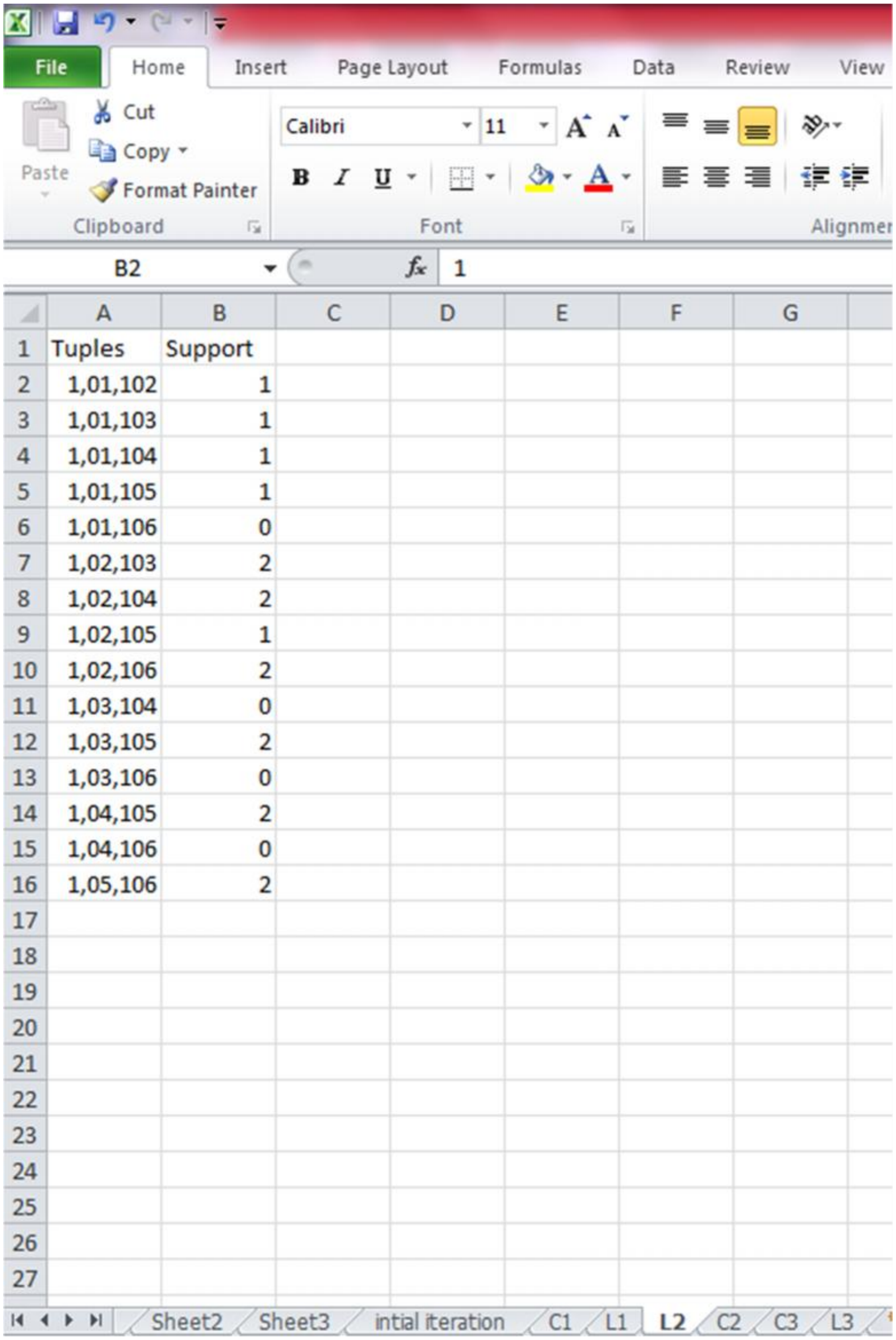
As illustrated in the figure, to reduce the number of transitions, original data base will be transposed.



Transactic	Count
101	2
102	3
103	3
104	2
105	2
106	2
107	1
108	1

Fig9: Item Counts

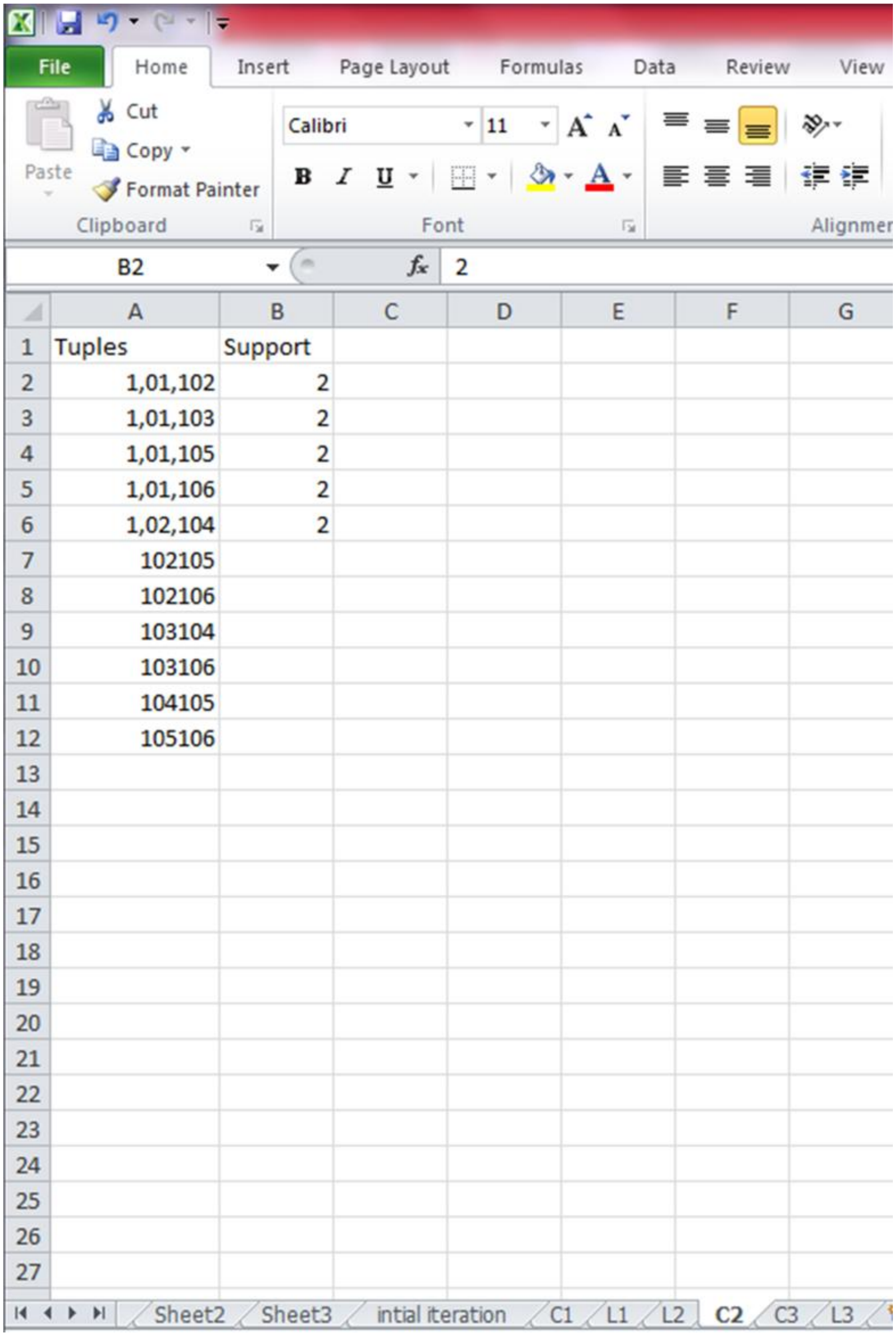
As illustrate in the figure 2, the item count is defined on the basis of Apriori algorithm



Tuples	Support
1,01,102	1
1,01,103	1
1,01,104	1
1,01,105	1
1,01,106	0
1,02,103	2
1,02,104	2
1,02,105	1
1,02,106	2
1,03,104	0
1,03,105	2
1,03,106	0
1,04,105	2
1,04,106	0
1,05,106	2

Fig10: Transaction combined

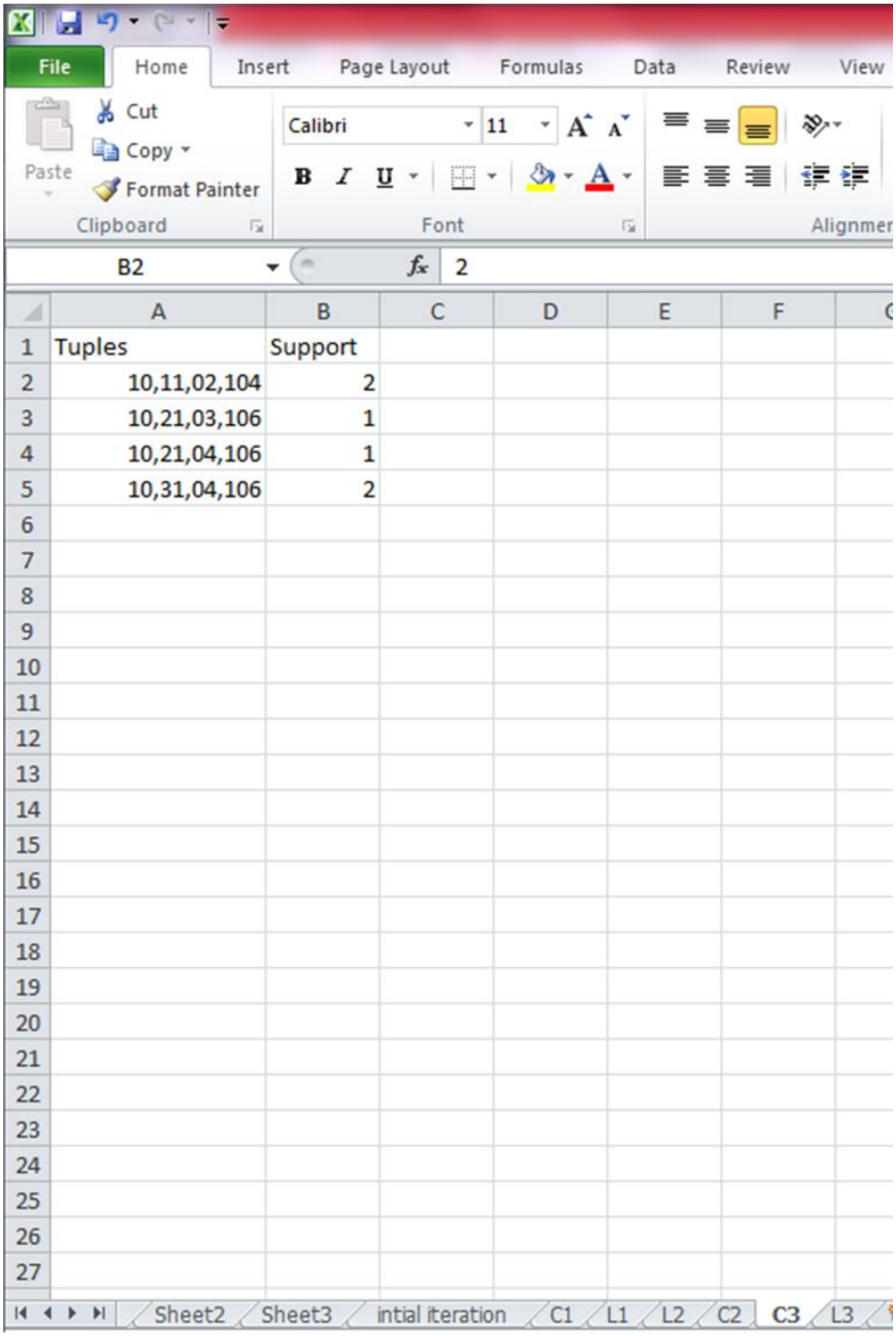
As illustrate in figure items are combined on the basis of Apriori algorithm. The values less than minimum support count will be deleted.



Tuples	Support
1,01,102	2
1,01,103	2
1,01,105	2
1,01,106	2
1,02,104	2
102105	
102106	
103104	
103106	
104105	
105106	

Fig11: Transactions combined

As illustrate in figure items are combined on the basis of Apriori algorithm. After combination of transactions, association rules generate.

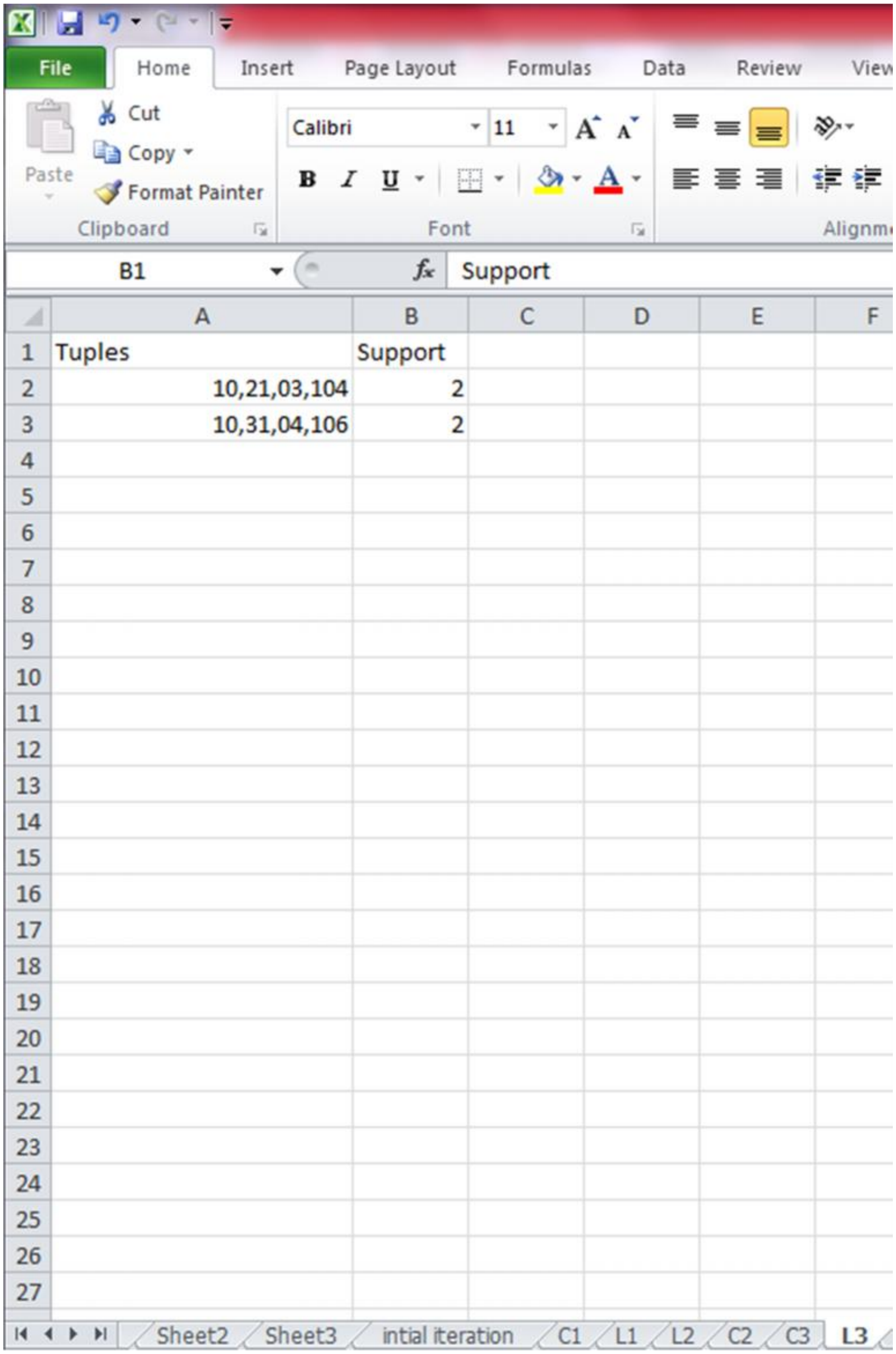


B2 f_x 2

Tuples	Support
10,11,02,104	2
10,21,03,106	1
10,21,04,106	1
10,31,04,106	2

Fig. 12: Transaction results

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



Tuples	Support
10,21,03,104	2
10,31,04,106	2

Fig 13: Association Rule

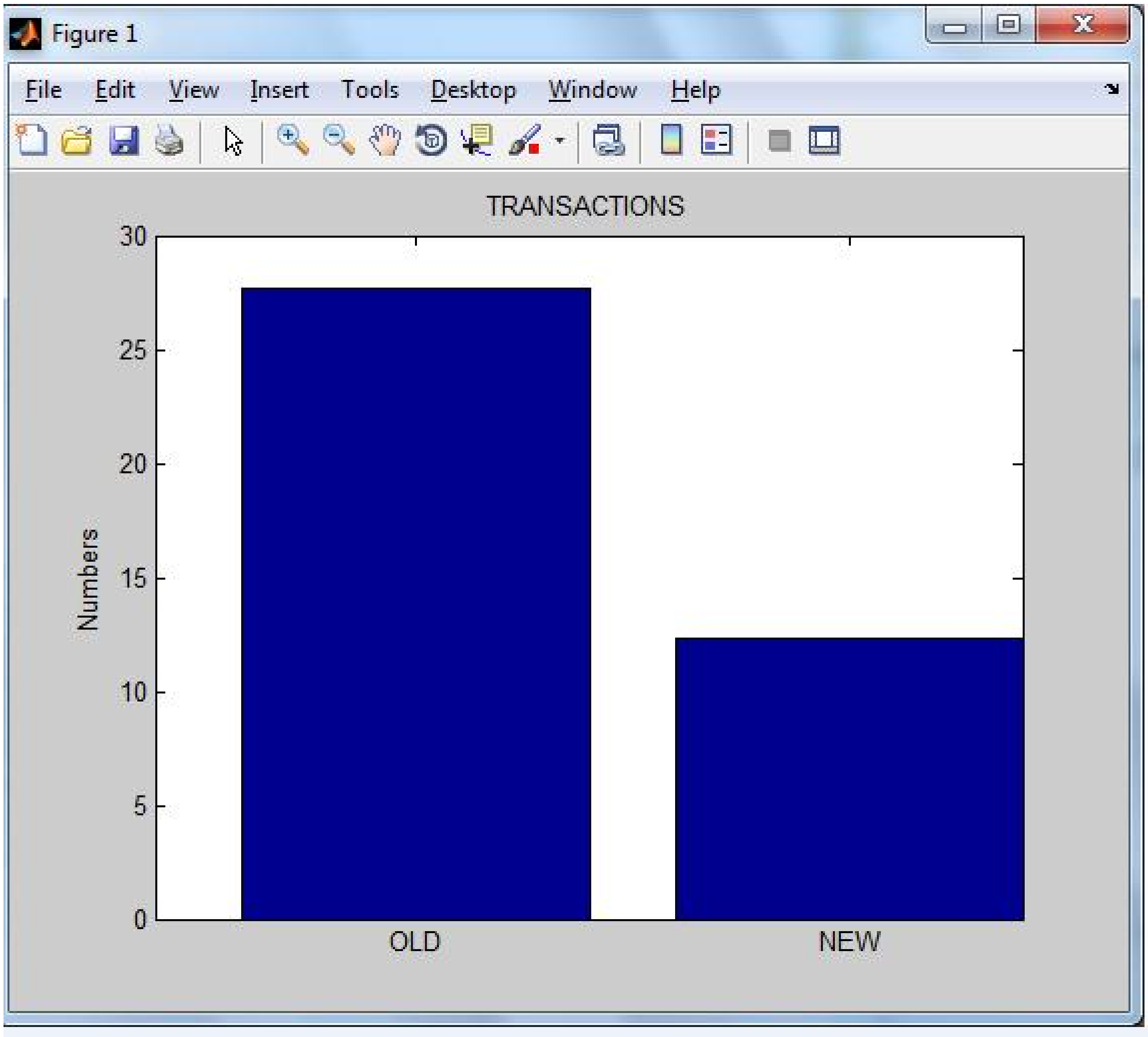


Fig: 14 Transactions Graph

Graph 1 show that old technique has more transactions as compare to new one. So proposed technique which is better than old one to reduce time consumption.

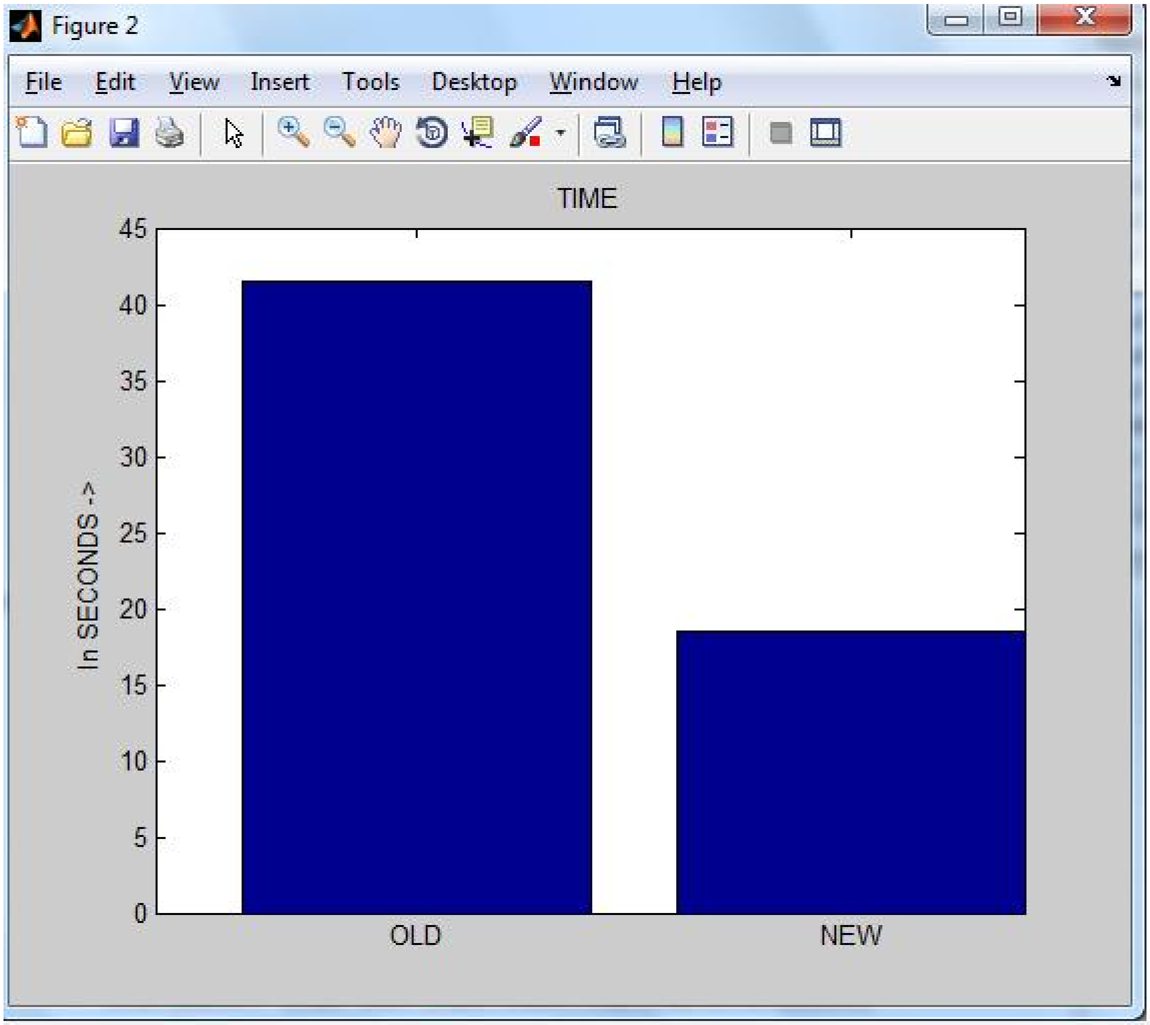


Fig: 15 Timing Graph

Graph 2 shows proposed method takes less time as compared old one

Chapter 5

Conclusion and future Work

The Apriori algorithm is the classical algorithm for the association rules are generated from the large dataset to reduce the data analysis effort. The apriori algorithm needs large number of database scan to generate the association rule, which will reduce the efficiency of apriori algorithm. To enhance the efficiency of apriori algorithm novel technic is been used in this research work. Novel technic based on firstly taking the transpose of the transactions and items .after taking the transpose of the transactions and items neglect those transactions that are used only single item .The transactions that has only one item means those items cannot create the association rules thus when we reduce the those transactions that has used only one items neglect from dataset after every scan of database that is those transactions that is included in candidate generation step or join step. Thus if when we reduce the transactions that can't be create association rules than the number of database scan or searching of items in the transaction taking lesser time as compared to old one in which for every items that is used single time in the database, scanning the every transaction for every single item due to number of transactions scan or database scan are larger. For accomplishing the data mining task many researchers used hardware device .in future the large number of transactions divides equally to the processor will for better enhancement of this algorithm

References

D. Gunaseelan, P. U. (2012). An Improved Frequent Pattern Algorithm for Mining Association Rules . *International Journal of Information and Communication Technology Research* .

Jaishree Singh1, H. R. (2013). Improving Efficiency of Apriori Algorithm Using Transaction Reduction . *International Journal of Scientific and Research Publications* .

Jiawei Han and Micheline Kamber, 2. *Data Mining : Concepts and Techniques.*

Kamber, J. H. *Data Mining : Concepts and Techniques.*

kumar, a. (2007). *Effect of Partition prime algorithm on Data Mining.* Research Scholar, Mewar University.

Ms. Rina Raval1, P. I. (2013). Survey on several improved Apriori algorithms. *IOSR Journal of Computer Engineering (IOSR-JCE)* .

Qiang Yang, Y. (2011). Application of Improved Apriori Algorithm on Educational Information. *Fifth International Conference on Genetic and Evolutionary Computing.*

R.Santhi, K. a. (2011). USING HASH BASED APRIORI ALGORITHM TO REDUCE THE CANDIDATE 2- ITEMSETS FOR MINING ASSOCIATION RULE. *Journal of Global Research in Computer Science* .

Singla, A. (2013). A Survey on Parallel Partition Prime Multiple Algorithm . *International Journal of Engineering and Advanced Technology (IJEAT)* .

WanjunYu, X. F. (2008). The Research of Improved Apriori Algorithm for Mining Association Rules.

Yang, S. (2012). Research and Application of Improved AprioriAlgorithm to Electronic Commerce. *11th International Symposium on Distributed Computing and Applications to Business, Engineering & Science* .

Yang, S. (2012). Research and Application of Improved Apriori Algorithm to Electronic. *11th International Symposium on Distributed Computing and Applications to Business, Engineering & Science.*

Ms. Rina Raval,(2013) Survey on several improved Apriori algorithms *IOSR Journal of Computer Engineering*

Varun Gupta,(2012) Impact Analysis of Requirement Prioritization on Regression Testing *2nd World Conference on Innovation and Computer Sciences.*

Li Hanguang, Ni Yu (2012) Intrusion Detection Technology Research Based on Apriori Algorithm , *International Conference on Applied Physics and Industrial Engineering*

Ma Xiaochun.(2005) The Research and Application of Data Mining in Network Intrusion Detection System *Xi an:Northwestern Polytechnical University* .

Yaqiong Jiang a ,Jun Wangb *(2011)* An Improved Association Rules Algorithm based on Frequent Item Sets *Advanced in Control Engineeringand Information Science.*

Swe Swe Nyein (2011) “*Mining Contents in Web Page Using Cosine Similarity*”.

K.Rajkumar (2011).“*Dynamic Web Page Segmentation Based on Detecting Reappearance and Layout of Tag Patterns for Small Screen Devices*”.

Shuang Lin, Jie Chen, Zhendong Niu(2012) .“*Combining a Segmentation-Like Approach And A Density-Based Approach In Content Extraction*” TSINGHUA SCIENCE AND Technologyissnll1007-0214ll05/18llpp256-264 Volume 17.

Ford Lumban Gaol (2010) *“Exploring The Pattern of Habits of Users Using Web Log Squential Pattern “Second International Conference on Advances in Computing, Control, and Telecommunication Technologies.*

Manne suneetha (2011) *“Clustering of Web Search Results using Suffix Tree Algorithm and Avoidance of Repetition of same Images in Search Results using L-Point Comparison Algorithm” PROCEEDINGS OF ICETECT*

Vivek Arvind. B Swaminathan. J Viswanathan. K. R.(2010*)”An Improvised Filtering Basedintelligent Recommendation Technique For Web Personalization “Department of R&D, DMI College Of Engineering for his support and effective guidance*

Rui Xie (2012) “Lexicon Construction: A Topic Model Approach” International Conference on Systems and Informatics (ICSAI)

Miguel Dar´ıo Duss´an-Sarria_(2009) “*A recommendation-based web content mining model for an university community*” international conference on Knowledge discovery and data mining

Gopal Pandey, Swati Patel, Vidhu Singhal, Akshay Kansara (2013)" *A Process Oriented Perception of Personalization Techniques in Web Mining"* nternational Journal of Science and Modern Engineering (IJISME) ISSN: 2319-6386, Volume-1, Issue-2.

Ashish Jain, Rajeev Sharma, Gireesh Dixit (2013) "*Page Ranking Algorithms in Web Mining, Limitations of Existing methods and a New Method for Indexing Web Pages*" International Conference on Communication Systems and Network.