ISSN 2348 - 2249
Volume-2, Issue-1, Jan-March, 2014
International Journal for Multi Disciplinary Engineering and Business Management
(IJMDEBM)

SURVEY ON DATA MINING APRIORI ALGORITHM

Reeti Trikha
Student, Computer Science Department
Punjab Technical University
India
reeti.trikha31@gmail.com

Jasmeet Singh
Computer Science Department
Punjab Technical University
India
jasmeetgurm@gmail.com

Abstract **- Data Mining is known as Knowledge Discovery in Databases. Basically it refers to the extraction of previously unknown and useful information from data in large databases. Association rule mining is a widely used approach in data mining. It helps in revealing interesting relationships. Decision making and understanding the behavior of the customer has become challenging problem for organizations. So for this a technique has been introduced in data mining known as Market Basket Analysis. The algorithm used for learning association rules is Apriori Algorithm. This algorithm finds the frequent pattern based on support and confidence measures. These measures limit the generation of interesting patterns. It's a simple algorithm but having many drawbacks. This paper shows a Survey on working of Apriori algorithm using Tanagra.**

Keywords **- Data Mining,KDD, Association, Apriori, Market Basket Analysis, Support, Confidence, Tanagra.**

I. INTRODUCTION

The main purpose of data mining [4] is to disclose the hidden information from the database. As we know the growth of data volume is increasing day by day in almost every sector like banking, marketing, medicine, website store design, telecommunication, manufacturing, transportation etc.,. So data mining proposes a different technique for deletion of repetitive data and conversion of data to more usable form. Basically data mining has two objectives:

PREDICTION: It helpsin predicting future values or unknown values of selected variables.
DESCRIPTION: It helps in describing general properties of the existing data in form of patterns.

Data mining is also known as Knowledge Discovery [2] in Databases. Basically they are synonyms for each other. Data Mining is a part of the knowledge discovery process. So, discovery of patterns depends upon the data mining tasks that have been employed. Working of KDD process has been shown in the figure 1.1 below:

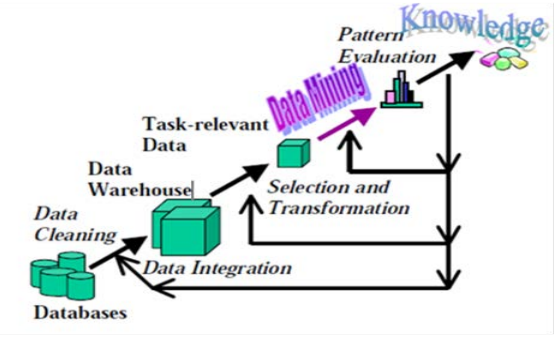

Fig 1.1: KDD Process

Usually it involves three processes as:
First is called PREPROCESSING. This process is executed before data mining techniques are actually applied on the data. It performs data cleansing, data integration, data selection and transformation. Second is called DATA MINING, the main process of KDD process, in which different algorithms are applied to extract hidden information. Then comes third process called as POST PROCESSING, which evaluates the mining result according to requirement of user and domain knowledge. If results are satisfactory, knowledge can be presented, otherwise more processes have to run till we get satisfactory results.

Data mining has discovered many techniques amongst which association rule mining is very important. The main goal of Association Rule Mining [5] is to detect relationships between databases. This is a common task which is helpful in various data mining projects. This introduces one of its best algorithms for mining known as Apriori. The Apriori algorithm [8] discover the frequent patterns from database which support the minimum threshold i.e. it discovers those patterns whose support and confidence must satisfy the minimum support and confidence.

II. ASSOCIATION RULE MINING

2.1 INTRODUCTION

Association rule learning is a method of data mining which helps in discovering interesting relations between variables in a large database. The discovery of frequent patterns is done among a large set of data items. From frequent pattern we

ISSN 2348 - 2249
Volume-2, Issue-1, Jan-March, 2014
International Journal for Multi Disciplinary Engineering and Business Management
(IJMDEBM)

mean those patterns which support minimum support and threshold [3]. For example, a super market analysis indicates that if a customer buys bread and butter together, then there is high probability of buying milk with it. This kind ofinformation helpsbusiness organisations to analyse thebehaviour of theircustomers. Shopping centres are using this approach these days to increase their sales by placing the items next to each other. Association Rule Mining has been used in various applications like banking where it helps banks in detecting those customers which are making profit for their organisation. It also helps in detecting frauds to the bank.

2.2 CONCEPTS OF ASSOCIATION RULE MINING

Suppose we have I as aset of items, Das a set of transactions ,then association rule is an implication of the X =>Y, where X,Y are subsets of I , and X,Y do not intersect. Accordingly, each rule comprises of two measures- support and confidence.
*Support:*It's the*probability that* both Xand Y will come together in a transaction.

Support(X, Y) = n (XUY)/N
N=Total no .of transactions.

*Confidence:*It's the probability that follows a condition that is a transaction having X also containsY.

Confidence(X, Y) = support (XUY)/support(X)

Itemset: It is a set of items in a transactional database.

k-itemset : It is a set of k-items in a transactional database.

Consider the following Transactional database Table-I:

TABLE I
TRANSACTIONAL DATABASE

TransactionId	Pencil	Sharpener	Eraser
1	1	1	0
2	0	0	1
3	0	0	0
4	1	1	1
5	0	1	0

In Table I, 1 represents the presence of item and 0 represents the absence of items. Now let's calculate the following:
Consider X=Pencil and Sharpener, Y = Eraser.

It's a 3-itemset where,

Support (X) =1/5
=0.2(20%)

Confidence(X→Y)=0.2/0.4
=0.5(50%)

Support means that pencil, sharpener and eraser all purchasedtogether while confidence says that whenever pencil and sharpener are purchased together, there is alsothe possibility that eraser is purchased with them.

2.3MARKET BASKET ANALYSIS

Market Basket Analysis [10] is one of the best examples of association rule mining. It helps in analysing the behaviour of the customer towards the items placed for its purchase by the organisation.Such kind of analysis helps organisations for introducing new strategies which can benefit their organisation. Market Basket analysis assumes that we have a large set of items, e.g. pencil, eraser, sharpener, scale, etc. The probability of buying a particular item or few items or all items will vary from customer to customer. This gives an idea what items people buy together. This strategy is used by marketers to position items and control the way a typical customer traverses their store.

GOALS OF MARKET-BASKET ANALYSIS:

1. It helps in generating association rules.
2. It suggests a way to eliminate duplicate data.

III. APRIORI ALGORITHM

Apriori is one of the association rule mining algorithm which is used to discover all frequent itemsets from transactional database [6].Apriori is used in many applications. To understand the working of Apriori algorithm we need to be familiar with the following terms:

Transaction: It's an entry in the database which contains a set of items.

Itemset: It's a set of items in a transactional database.

Candidate Itemset (Li): Items which are only used for the processing initially. They can be all possible combination of itemsets.

Minimum Support: It's a condition which helps to eliminate the non-frequent items from database. It helps in generating frequent items.

Frequent Itemset (Large Itemset (Li)): The itemsets which contain frequent items are known as frequent itemsets.

Apriori uses iterative approach known as breadth first search (level-wise search), where k-1 itemsets are used to generate k itemsets. Apriori works on the following main principle:

ISSN 2348 - 2249
Volume-2, Issue-1, Jan-March, 2014
International Journal for Multi Disciplinary Engineering and Business Management
(IJMDEBM)

"All frequent itemsets should have their frequent subsets". This means if {c,d,e} is a frequent itemset , then all its subsets {c,d},{c,e},{d,e},{c},{d} and {e} must also be frequent.

Apriori algorithm uses following two concepts:

(i) Joining and
(ii) Pruning

Apriori Algorithm Steps:

(i) Given support threshold's'. Now in first pass, we find the items that satisfy given threshold. The set discovered is called L1, frequent itemset.
(ii) Now pairs of items in L1, act as candidate pairs C2 for second pass. The pairs in C2 which support s are frequent pairs, L2.
(iii) The candidate triples, C3 are those sets which have their sets in L2. On the third pass, count the occurrence of triples in C3; those with a count of at least's' are frequent triples, L3.
(iv)Proceed as above till we are left with only frequent itemsets.

A general outline of the Apriori algorithm is:

- Firstly, count all itemsets of size K.
- Prune the non-frequent itemsets of size K.
- Generate new candidates of size K+1.
- Again prune the non-frequent itemsets of size K.
- Repeat from step 1, until we are left with no more candidate itemsets,
- When we get all frequent itemsets, generate rules.

Let's consider an example to show how Apriori algorithm actually works:
Consider a database, D consisting of 9 transactions. Suppose we require min. support count 2. We need to find frequent itemsets first and then association rules will be generated. Table II shows transactional databasehaving 9 transactions.

TABLE II

TID	List of items
T100	l1,l2,l5
T100	l2,l4
T100	l2,l3
T100	l1,l2,l4
T100	l1,l3
T100	l2,l3
T100	l1,l3
T100	l1,l2,l3,l5
T100	l1,l2,l3

We scan database to identify the number of occurrence of specific item. After that we will compare candidate support count with minimum support count as shown in Fig-3.1 below:

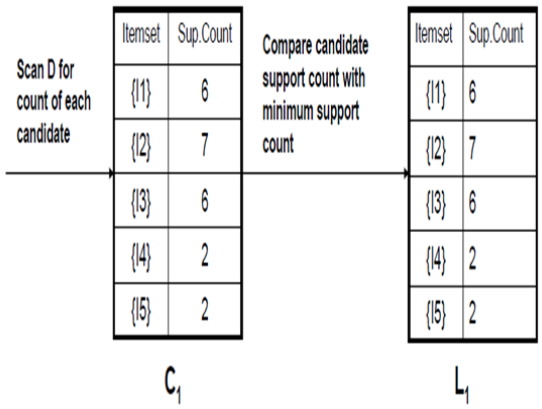

Fig-3.1: Generating 1-frequent itemset

To discover the set of frequent 2-itemsets, L2, the algorithm will use L1. It joins L1 to generate candidate set of 2-itemsets.

Again the transactions in D are scanned and the support counts are compared. The frequent sets of items, L2 are discovered as shown in Fig-3.2

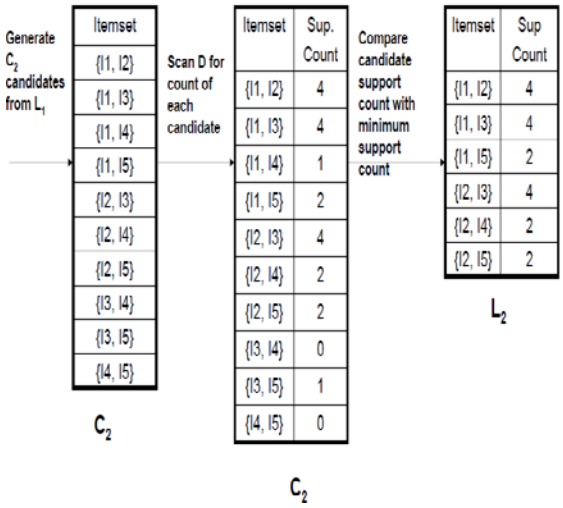

Fig-3.2: Generating 2- frequent itemset

Similarly C3 is computed. We compute L2 and join L2. In this way we are left with frequent itemsets in the end as shown in Fig-3.3

ISSN 2348 - 2249
Volume-2, Issue-1, Jan-March, 2014
International Journal for Multi Disciplinary Engineering and Business Management
(IJMDEBM)

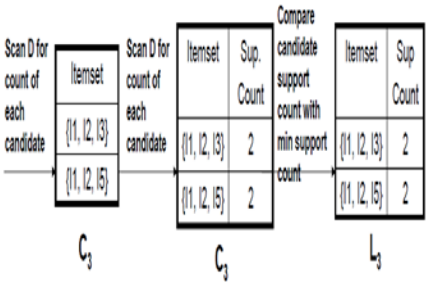

Fig-3.3: Generating 3- frequent itemset

Pseudo CodeForAprioriAlgorithm [1]:
Ck :Set of Candidate itemsets of size k
Lk: Set of Frequentitemsets of size k
L1 = {Frequent items};
For (k=1; Lk! =null; k++) do begin
Ck+1=candidates generated from Lk;
For each transaction in database do
Increment the count of all candidates in Ck+1 that are contained in t
Lk+1 = Candidates in Ck+1 with minimum support
End
Return Lk;
Apriori algorithm is the classical and simplest algorithm to implement the concept of association rule mining. But there are some disadvantages as follows:
Drawbacks of AprioriAlgorithm:
1. It consumes lot of time for scanning database.
2. It increases space complexity by generating a large number of non-frequent itemsets.
3.It makes several iterations while mining data.
4. Generation of large amount of frequent itemset,also makes it inefficient to use.
5. It does notconsiders the significance of item to the user or business. Itsfocus is only on the presence and absence of the item.

IV. WORKING OF APRIORI USING TANAGRA

Figure-4.1 will be the first screen when Tanagra starts up

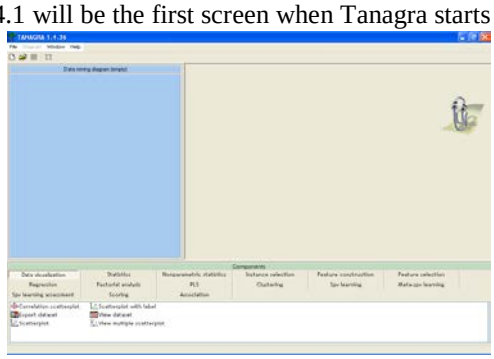

Figure-4.1: Tanagra Home Screen

Click on new it will ask for the dataset file as shown in figure-4.2

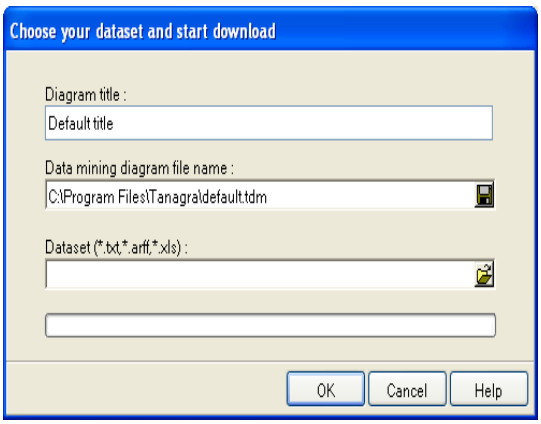

Figure-4.2: Choose Dataset

Choose the dataset by clicking to the dataset icon. It will ask location as in figure-4.3

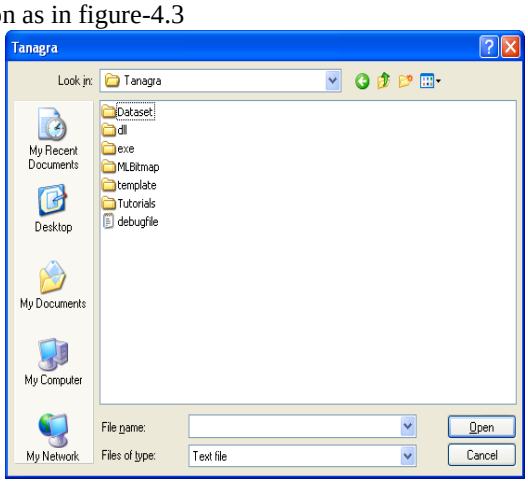

Figure-4.3: Choose Dataset from specified location

After choosing dataset from the folder the path is shown as in figure-4.4

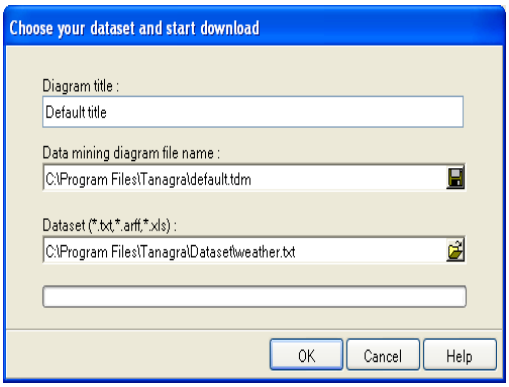

Figure-4.4: Dataset location

Click ok and it will show the Dataset Description as shown in figure-4.5

ISSN 2348 - 2249
Volume-2, Issue-1, Jan-March, 2014
International Journal for Multi Disciplinary Engineering and Business Management
(IJMDEBM)

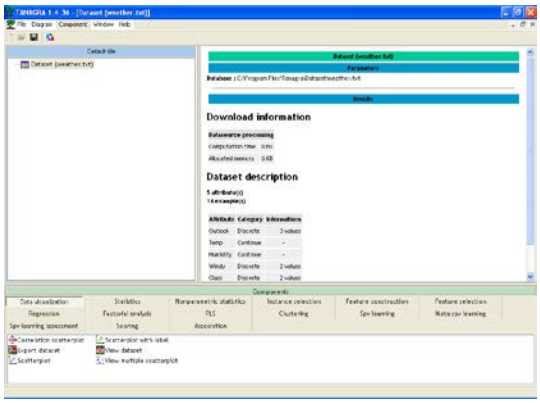

Figure-4.5: Dataset Description

Drag view dataset from lower window to upper left window onto Dataset as shown in figure-4.6

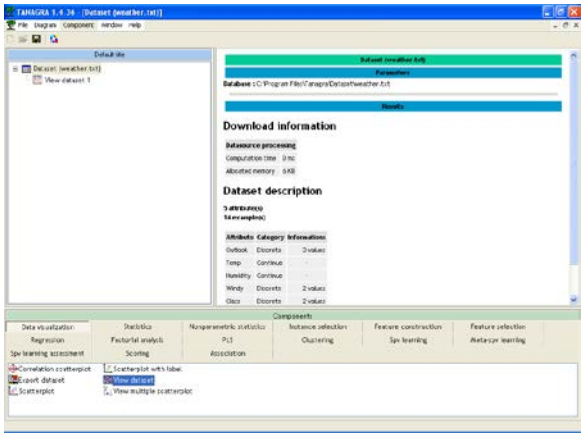

Figure-4.6: View Dataset

Right click on view dataset 1 and click on view, following figure-4.7 appears

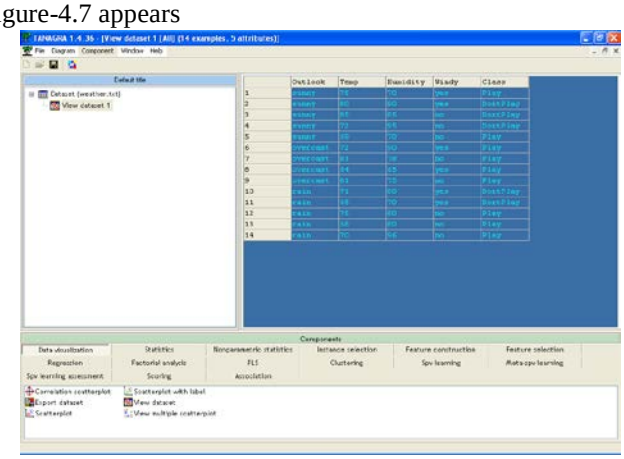

Figure-4.7: Dataset viewing screen

Add define status from toolbar just close to save button as shown in figure-4.8

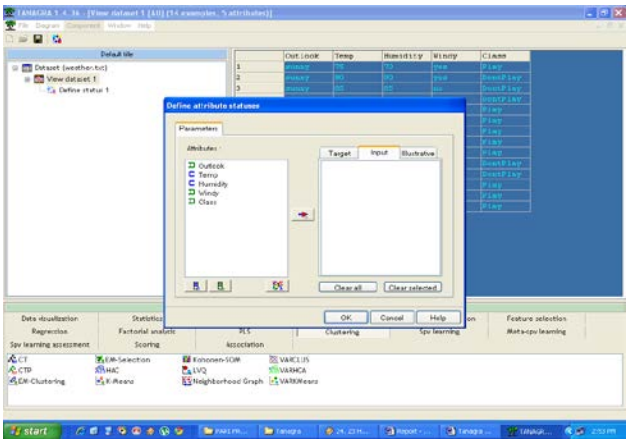

Figure-4.8: Define Status

For Apriori algorithm we need two or more discrete attributes as input and no need for target. So select as shown in figure-4.9. For input and target attributes we can have help in form of tooltip by placing mouse over the respective algorithm.

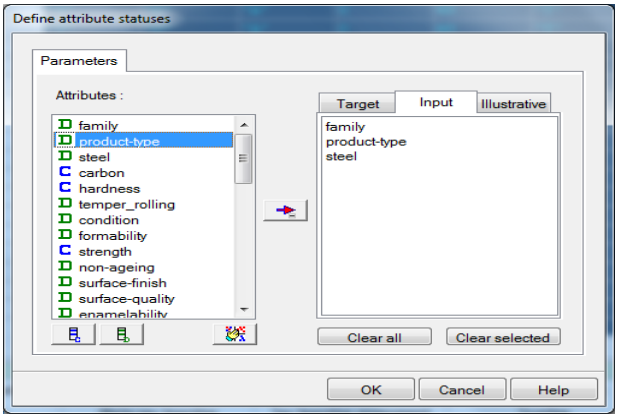

Figure-4.9: Define Status attributes

Click on association from lower window drag Apriori over define status as in figure-4.10

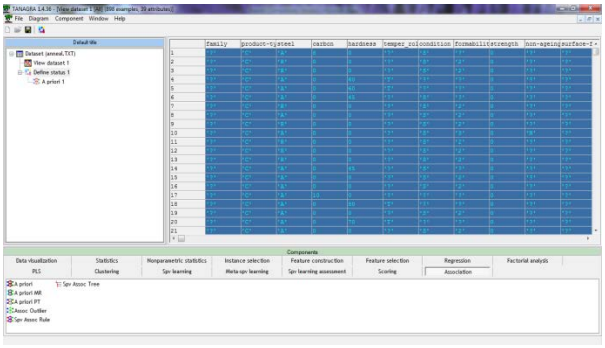

Figure-4.10: Apriori

Right click on Apriori, execute and view the result as in figure-4.11

ISSN 2348 - 2249
Volume-2, Issue-1, Jan-March, 2014
International Journal for Multi Disciplinary Engineering and Business Management
(IJMDEBM)

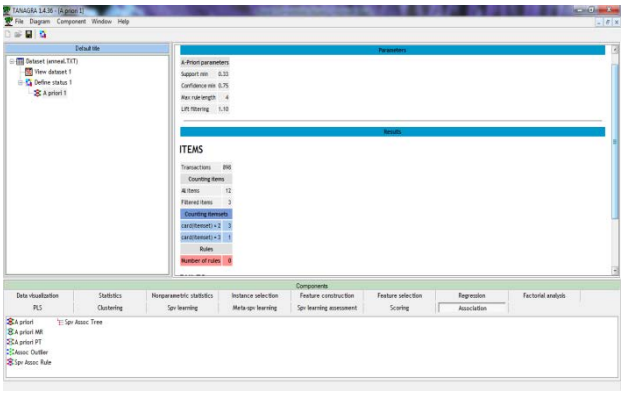

Figure-4.11: Apriori Result

V. CONCLUSION AND FUTURE WORK

Association rule mining is useful indiscovering useful patterns from transactional database. This algorithm is used for the implementation ofassociation rule mining. But this classical algorithm has several limitations like scanning time, memory optimization, candidate generation,etc.

Our classical Apriori treats all the items in database equally by focusing only on the presence and absence of an item within the transaction. It ignores the significance of that item to user or business. So, Apriori algorithm efficiency can be improved further to provide valuable information to customer as well as business.

REFERENCES

[1]Ms. ArtiRathod ,Mr. AjaysinghDhabariya& Mr. Chintan Thacker,(Sep. 2013),"A review on Association Rule Mining and Improved Apriori Algorithms",International Journal of Scientific Research in Computer Science , vol. 1 ,no. 11.

[2] Frawley, W., Piatetsky- Shapiro, G., Matheus, (1992), "Knowledge Discovery in Databases: An Overview", AI Magazine, fall 1992, pp. 213-228.

[3] Chun-Jung Chu, Vincent S. Tseng and Tyne Liang (November (2008), "Mining temporal rare utility Itemsets in large databases using relative utility thresholds", International Journal of Innovative Computing, Information and Control, Vol. 4, no. 11.

[4]Jiawei Han and Michelin Kamber-book second edition,"Data Mining Concepts and Techniques".

[5],Irena Tudor, Universitatea Petrol-Gaze din ploeiesti, "Association Rule Mining as a Data Mining Technique", Bd.Bucuresti 39, ploeiesti,Catedra de Informatica, Vol-LX,No.1,2008.

[6] MamtaDhanda, (July 2011),"An efficient approach to extract frequent patterns from transactional database",International Journal of Engineering Science &Technology, vol. 3, no. 7.

[7],Wei Zhang ,Honghzi Liao and Na Zhao (2008), "Research On The Frequent Pattern Growth Algorithm about Association Rule Mining", International Seminar on Business and Information Management.

[8] Zhang Shu-mao and DU Ying-shuang (Sep 2008),"The anaysis and improvement of Apriori Algorithm", HAN Feng Journal of Communication and Computer, vol. 5, no. 9.

[9] Du Ping and GaoYongping (2010), "A New Improvement of Apriori Algorithm For Mining Association Rules", International Conference on Computer Application and System Modeling.

[10] Dr.Dhanabhakyam, Dr.M.Punithavalli, Dr.SNS college of Arts and Science,"The Survey on Data Mining Algorithm for market basket analysis", Global Journal of Computer Science and Technology (IJCSEIT), Vol.11,Issue 11 Version 1.0, July 2012,ISSN:0975-4172.

[11] Sheila A.Abaya,"Association Rule Mining based on Apriori algorithm in minimizing candidate generation", International Journal of Scientific & engineering Research, Vo1-3,Issue-7,July 2012,ISSN 2229-5518.

[12] Raorane A.A ,Kulkarni R.V and Jitkar B.D"Association Rule- Extracting Knowledge Using Market Basket Analysis" ,Dept. of Computer Science,Vivekanand College, TarabaiParkkolhapur, Research Journal of Recent Sciences, Vo11(2)19-27,Feb. 2012.

