ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Issue 02, February 2018

Data Analysis with Apriori Algorithm Using Rule Association Mining

Praveen Kumar. G [1], Dr. S K Mohan Rao [2]
1 Research Scholar, Department of Information Technology, Sreenidhi Institute of Engineering &Science, Hyderabad, India.
2. Professor & Principal, Gandhi Institute for Technology, Bhubaneswar, India.

Abstract: At the present a day's Data mining has a lot of e-Commerce applications. The key problem is how to find useful hidden patterns for better business applications in the retail sector. For the solution of these problems, The Apriori algorithm is one of the most popular data mining approach for finding frequent item sets from a transaction dataset and derive association rules. Rules are the discovered knowledge from the data base. Finding frequent item set (item sets with frequency larger than or equal to a user specified minimum support) is not trivial because of its combinatorial explosion. Once frequent item sets are obtained, it is straightforward to generate association rules with confidence larger than or equal to a user specified minimum confidence. The paper illustrating apriori algorithm on simulated database and finds the association rules on different confidence value.

Keywords: Data Mining, e-Commerce, apriori algorithm, association rules, support, confidence, retail sector.

I. Introduction

Today retailer is facing dynamic and competitive environment on global platform and competitiveness retailers are seeking better market campaign [i]. Retailer are collecting large amount of customer daily transaction details. This data collection requires proper mechanisms to convert it into knowledge, using this knowledge retailer can make better business decision. Retail industry is looking strategy where they can target right customers who may be profitable to their business. Data mining is the extraction of hidden predictive information from very large databases. It is a powerful technology with great potential to help organizations focus on the most important information in their data warehouses [2] [4] [6]. Data mining tools predict future trends and behaviours, helps organizations to make proactive knowledge-driven decisions [7] [5]. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools have the answer of this question. Those traditionally methods were lot of time consuming to resolve the problems or decision making for profitable business. Data mining prepare databases for finding hidden patterns, finding predictive information that experts may miss Because it lies outside their expectations. from the last decade data mining have got a rich focus due to its significance in decision making and it has become an essential component in various industries [3][7][5]. the field of data mining have been prospered and posed into new areas such as manufacturing, insurance, medicine, retail etc [3][2][4][6]. Hence, this paper reviews the various trends of data mining and its relative applications from past to present and discusses how effectively can be used for targeting profitable customers in campaigns.

II. Related Work

Association rule mining is interested in finding frequent rules that define relations between unelated frequent items in databases, and it has two main measurements: support and confidence values. the frequent item sets is defined as the item set that have support value greater than or equal to a minimum threshold support value, and frequent rules as the rules that have confidence value greater than or equal to minimum threshold confidence value. These threshold values are traditionally assumed to be available for mining frequent item sets. Association rule mining is all about finding all rules whose support and confidence exceed the threshold, minimum support and minimum confidence values.

Association rule mining proceeds on two main steps. The first step is to find all item sets with adequate supports and the second step is to generate association rules by combining these frequent or large item-sets [8] [9][lO].In the traditional association rules mining [11][12], minimum support threshold and minimum confidence threshold values are assumed to be available for mining frequent itemsets, which is hard to be set without specific knowledge; users have difficulties in setting the support threshold to obtain their required results. Setting the support threshold too large, would produce only a small number of rules or even no rules to conclude. In that case, a smaller threshold value should be guessed (imposed) to do the mining again, which may or may not give a better result, as by setting the threshold too small, too many results would be produced for the users, too many results would require not only very long time for computation but also for screening these rules. That would explain the need to develop an algorithm to generate a minimum support, and minimum confidence values depending on the datasets in the databases.

To use association rule mining without support threshold [13][14][15][16], another constraint such as similarity or confidence pruning is usually introduced. However, the coincidental itemset problem had not been directly considered by any of these researches. There are some researches that are relevant to the coincidental itemset problem, and proposed an additional measure [17] in order to improve the support- confidence framework.

III. DESIGN

Data mining concepts:
Associations and item-sets: It is denoted as
A---->B

If A is true then B will also true. Example: If Deepavali comes, the sales of fireworks go up.
A= Deepavali (The festival of light in India) Week.
B= sales of fireworks go up.
Using this association rule can predict that if A is true then B also true.
For any rule if A ---->B ----A,then>B----Aand>B are called an "interesting item-set"

Example:
People buying shoes also buy socks.
People buying shocks also buy shoes.

Sales Transaction Table: From the basket analysis of the set of products in a single transaction. Discovering for example, that a customer who buys shoes is likely to buy socks Shoes ----Socks>

Transactional Database: The set of all sales transactions is called the population. The representation of the transactions in one record per transaction. The transaction is represented by a data Tuple.

TRANSACTION DATA

TRANSACTION	DATA
TXl	Shoes, Socks, Tie
TX2	Shoes, Socks, Tie, Belt, Shirt
TX3	Shoes, Tie
TX4	Shoes, Socks, Belt

TABLE 1.

Socks ----Tie>
Sock is the rule antecedent
Tie is the rule consequent

Support and Confidence: Any given association rule has a support level and a confidence level.

Support it the percentage of the population which satisfies the rule or in the other words the support for a rule R is the ratio of the number of occurrence of R, given all occurrences of all rules. The support of an association pattern is the percentage of task - relevant data transactions for which the pattern is true.

Support (A ----B)>= P (A U B) (1)

The transaction table given above is showing the item sets Purchased by the customer in a period of time. The support for the item sets Bread and noodles means a customer who purchased bread also purchased the noodles is given below. The support for ten transactions where bread and noodles occur together is two.

Support for {Bread, Noodles} = 20/10 0= 0.20.

This means the association of data set or item set, the bread and butter brought together with 20 percent support.

Confidence for Bread ----Noodles> = 2/8 = 0.25

This means that a customer who buy bread then there is a confidence of 25 percent that it also buy butter.
Mining for frequent item-sets

IV. METHODOLOGY

Mining for frequent item-sets: To mining the frequent item we use following Improved Apriori algorithm.

The Improved Apriori Algorithm: Improved Apriori is a seminal algorithm for finding frequent item-sets using candidate generation [18]. Mining for association among items in a large database of sales transaction is an important database mining function.

Given minimum required support s as interestingness criterion:

1. Search for all individual elements (I-element item-set) that have a minimum support of s.

2. **2.Repeat:-**

   l. From the results of the previous search for i-element item-set, search for all i+ 1 element item- Sets that have a minimum support of item-set.

   2. This becomes the set of all frequent (i+ 1) item- Sets that are interesting

   3.Until item-set size reaches maximum. By using the consumer database given in table no.2

Let's illustrate the process of Apriori with an example, let takes the consumer database which is showing the number of item- sets purchased by the consumers from a bakery shop.

Let Minimum support is 0.3.

Single item like bread, butter etc. in the given database every item occurs three of more time than the minimum support or minimum support threshold value is 0.3. Now focus on interestingness of the single item-sets, so database contains many items like bread, butter, milk etc

So interestingness 1- element item-sets

{Bread}, {Butter}, {milk}, {ice-cream}, {noodles}

Step 1

Cl

Item	support
Bread	0.8
Butter	0.7
Ice-Cream	0.5
Milk	0.5
Noodles	0.5

L1

item	**support**
Bread	0.8
Butter	0.7
Ice-Cream	0.5
Milk	0.5
Noodles	0.5

In the step 1 no Item thrown away because Its support values are greater than minimum support values.

Now the support for two element item- sets.

Interestingness 2-element item-sets

{Bread, Butter}, {Bread, Milk}, {Bread,noodles},{Bread,ice-cream}, {Butter,Milk}, {Butter,noodles}, {Noodles, ice-cream}

C2→ L2

C2

item	setssupport
{Bread,Butter,Milk}	0.3
{Bread,Milk,lee-cream}	0.1
{Bread,Butter,ice-cream}	0.0
{Butter,Milk,Noodles}	0.1
{Bread"Milk,noodles}	0.0
{Noodles,ice-cream,Bread}	0.2

item	setssupport
{Bread,Butter}	0.5
{Bread,Milk}	0.4
{Bread,noodles}	0.2
{Bread,ice-cream}	0.3
Butter,Milk}	0.4
{Butter,noodles}	0.3
{Noodles, ice-cream}	0.3

L2

item	setssupport
{Bread,Butter}	0.5
{Bread,Milk}	0.4
{Bread,ice-cream}	0.3
Butter,Milk}	0.4
{Butter,noodles}	0.3
{Noodles, ice-cream}	0.3

Here {Bread, Noodles} Item-set thrown away because its support value is less then minimum support.

Interestingness 3-element item-sets

C3 → L3

C3

L3

item-sets	support
{Bread,Butter,Milk}	0.3

Here only one item-sets which satisfy the minimum support value. So after three iteration, only one item-set filtered. {Bread, Butter, Milk}

Item-sets showing that at least three customer buy Bread, Butter and Milk together, its minimum support value which is 0.3.the main advantage of the improved apriori algorithm is that it only takes data from precious iteration not from the whole data. If in the previous iteration some item-sets thrown away just because they are not satisfying the minimum support.Every iteration gives new item-sets which follow the minimum support. After filtering, algorithm gives the maximum number of item-sets which is repeated maximum time. In the above example three customers purchasing the three item-sets together this is Bread, Butter and Milk.

So in the retail sector, finding the behavior of a customer is a profitable approach for finding the tendency of a customer. If a customer X frequently comes in a bakery then data mining can predict the purchasing tendency of the customer X on the basis of its previous records. Data mining can predict that if customers X buying Bread and Butter then he is likely to buy Milk also. So it the main task of the improved apriori algorithm is that it can find the hidden pattern in the data base.

Mining for association rules:

Association rules are the form

A----->B

This implies that if a customer purchase item A then he also purchase item B. For the association rule mining two threshold values are required. As given in the design part.

- Minimum support

- Minimum confidence

The ordering of the items is not important. a customer can purchase item in any order means if he purchased Milk first then Butter and after purchasing both he can buy Bread or after buying Bread he can purchase Milk and Butter. But in the association rules the direction is important.

If A-B is different from B-A

There is general procedure for defining the mining association rules using Improved Apriori algorithm.

- Use apriori to generate frequent item-sets of different sizes

- At each iteration divide each frequent item-set X into two parts antecedent (LHS) and consequent (RHS) this represents a rule of the form LHS?RHS.

- The confidence of such a rule is support(X) / support(LHS)

- Discard all rules whose confidence is less than minimum confidence

The frequent item-sets are {Bread, Butter, Milk} with support is 0.3. There is the generation of the frequent item-set of size three. These item- sets can be divided in to some rules which is given below

{Bread}? {Butter, Milk}
{Bread, Butter} ? {Milk}
{Bread, Milk} ?{Butter}
{Butter}? {Bread, Milk}
{Butter, Milk}? {Bread}
{Milk}?{Bread, Butter}

(2) Now generate the idea of the improved apriori algorithms for generating rules for three item-sets which are frequently used the association rules for three item-sets like Bread, Butter and Milk which holds the minimum support 30 percent. The set of item- set Bread, Butter and Milk can be represented in the different forms of rules. Every rule has a certain confidence. Equation calculates the confidence for each rule. So the confidence of each rule is given in table 3.

TABLE Ill.

ASSOCIATION RULES FOR FREQUENT ITEM-SETS

Rule	**Confidence(%)**
{Bread}-+{Butter, Milk}	37
{Bread,Butter}-+{Milk}	**60**
Bread, Milk}-+{Butter}	**75**
{Butter}-+{Bread, Milk}	**42**
{Butter, Milk}-+{Bread}	**75**
{Milk}-+{Bread, Butter}	**75**

If the mlTIlmUm confidence threshold IS 70 percentages thendiscovered rules are

{Bread, Milk} � {Butter}

{Butter, Milk}� {Bread}
{Milk}� {Bread, Butter}

Because the confidence value of these rules are greater than minimum confidence threshold value which is 70 percent.

So in the simple language if a customer buy Bread and Milk he is likely to buy Butter. A customer buy Butter and Milk is likely to Bread. A customer buy Milk is likely to buy Bread and Butter. Suppose changing the value of the minimum confidence means average of all rules as the minimum confidence value is 60. This value adds a rule which is {Bread, Butter} � {Milk}

So association rules which frequently used and follow the minimum confidence. So the research part of this paper is this by changing the value of minimum confidence, gives different association rules. The value of minimum confidence is high then rules filtered more accurately.

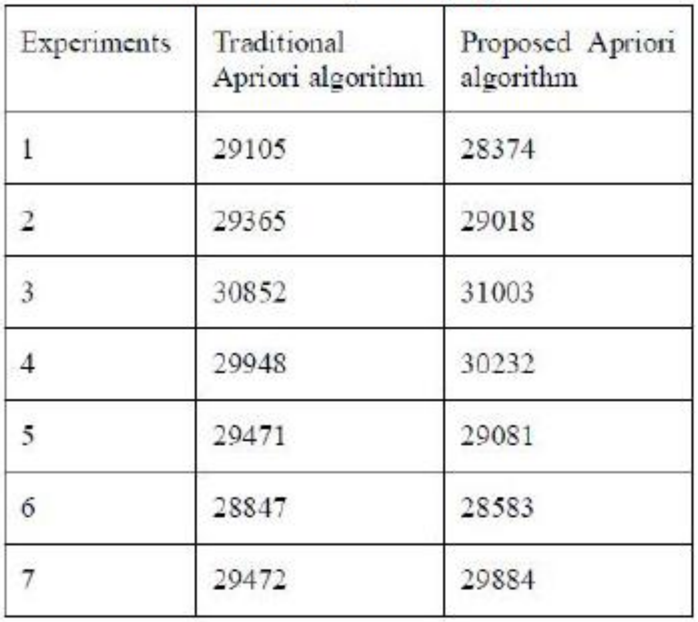
IV. Results Analysis

The implementation of the traditional Apriori algorithm and proposed Apriori algorithm is performed in previous section. This section provides the comparative performance study and results evaluation of the proposed algorithm.

A. Memory Usage

The amount of main memory required to perform data analysis using the algorithm is termed here as memory usages or space complexity. The estimated comparative memory consumption of the implemented algorithm is reported using figure 1 and table 1.

Table 1 Memory Consumption



Experiments	Traditional Apriori algorithm	Proposed Apriori algorithm
1	29105	28374
2	29365	29018
3	30852	31003
4	29948	30232
5	29471	29081
6	28847	28583
7	29472	29884

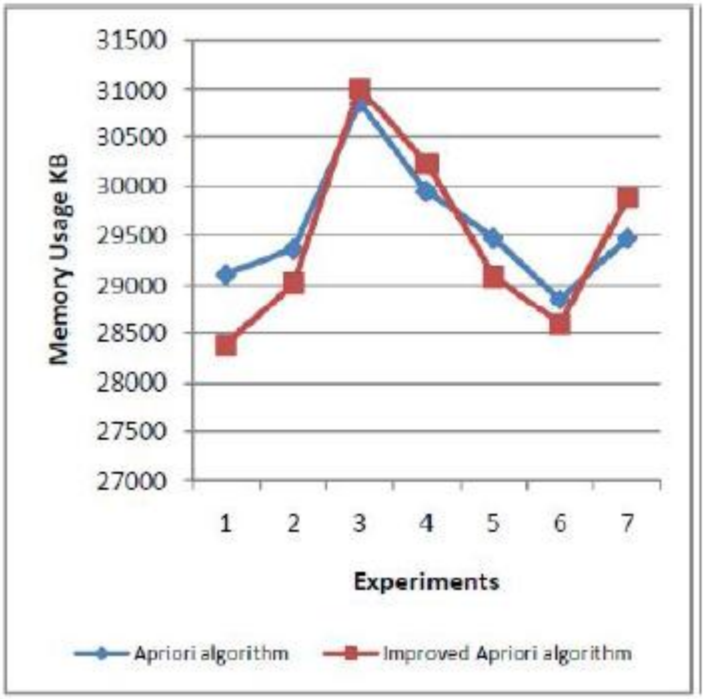


Figure 1 Memory Consumption

The memory consumption or space complexity of the implemented algorithms namely proposed and red line shows the performance of improved A priori and traditional algorithm is denoted using blue line. In most of the experiments the performance of algorithms are fluctuating but it remains adoptable for data analysis in both the cases. In experimentations size of data is increases and their memory consumption is evaluated. During the experimentations that is observed if the number of candidate set generation is large then the memory requirement is higher and otherwise it remains fixed not much fluctuating.

Traditional A priori algorithm is reported using the figure 1.In this diagram the X axis contains different experiments performed with system and Y axis shows amount of main memory consumed in terms of KB (kilobytes). Additionally to represent the performance of algorithms

B. Time Consumption

The amount of time consumed for developing the A priori based association rules using the input datasets is termed here as the time consumption of algorithm or time complexity. The time consumption of the implemented algorithms is reported using the figure 2 and table 2. In this figure X axis shows the different experiments performed with the system and the Y axis shows the amount of time consumed for developing the association rules in terms of seconds. According to the given performance the implemented algorithms proposed .Apriori algorithm consumes less amount of time as compared to the traditional Apriori. But the time consumption is increases as the amount of data for association rule development is increases.

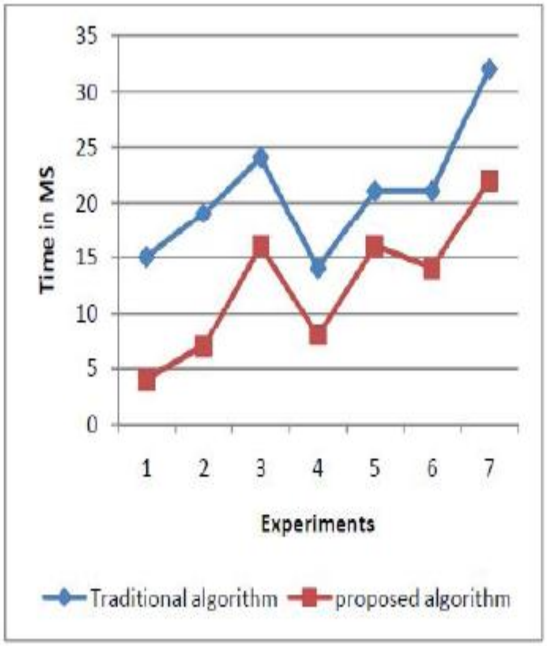
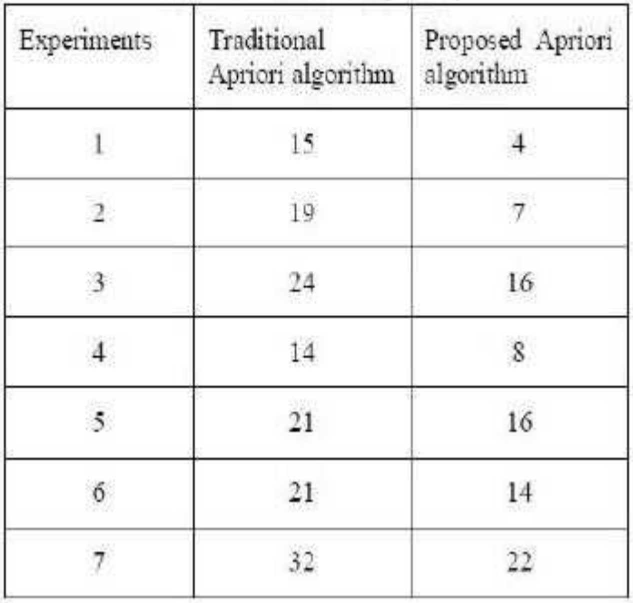


Figure 2 Time Complexity

Table 2 Time Complexity



Experiments	Traditional Apriori algorithm	Proposed Apriori algorithm
1	15	4
2	19	7
3	24	16
4	14	8
5	21	16
6	21	14
7	32	22

C. Transaction Vs. Rules

In order to represent the effectiveness of the proposed technique the comparison of both the implemented algorithms is performed which number of input transactions and developed association rules in both the conditions. The performed experimentation and their results are provided using figure 3 and table 3.

In this diagram the number of input transactions are given in X axis and the number of generated rules are provided in Y axis. To shows the performance red line shows the performance of proposed technique and blue line shows the performance of traditional Apriori algorithm. In most of the time similar numbers of rules are generated with the less amount of time as compared to the traditional algorithm. The obtained observational results are also demonstrated the proposed technique provides high quality rules as compared to traditional approach

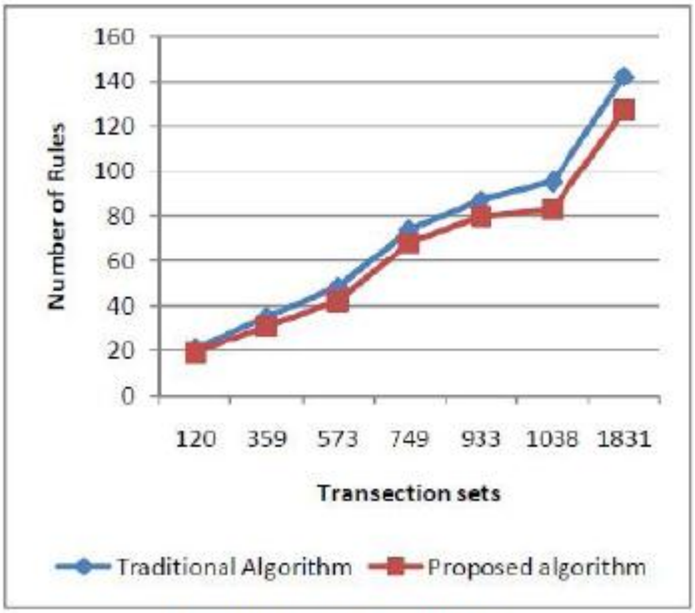
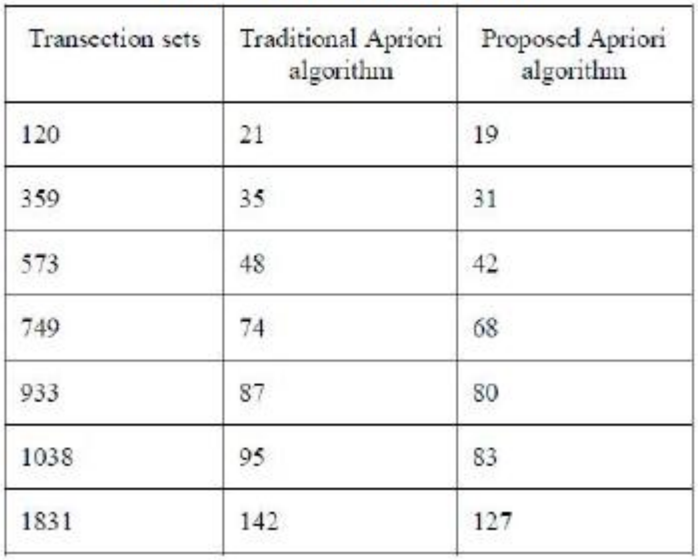


Figure 3 Transection Vs. Rule

Table 3 Transection Vs Rule



Transection sets	Traditional Apriori algorithm	Proposed Apriori algorithm
120	21	19
359	35	31
573	48	42
749	74	68
933	87	80
1038	95	83
1831	142	127

V.CONCLUSIONS

This paper is an attempt to use data mining as a tool used to find the hidden pattern of the frequently used item-sets. An Apriori Algorithm may play an important role for finding these patterns from large databases so that various sectors can make better business decisions especially in the retail sector. Apriori algorithm may find the tendency of a customer on the basis of frequently purchased item-sets. There are wide range of industries have deployed successful applications of data mining. Data mining in retail industry can be deployed for market campaigns, to target profitable customers using reward based points. The retail industry will gain, sustain and will be more successful in this competitive market if adopted data mining technology for market campaigns.

REFERENCES

[1]. 'The 6 biggest challenges retailer Face today", www.onStepRetail .com. retrieved on June 20II

[2]. Berry, M. J. A. and Linoff, G. Data mining techniques for marketing, sales and customer support, USA: John Wiley and Sons,1997

[3]. Andre Bergmann, "Data Mining for Manufacturing: Preventive Maintenance, Failure Prediction, and Quality Control".

[4]. Fayyad, U. M; Piatetsky-Shapiro, G.; Smyth, P.; and Uthurusamy, R. 1996. Advances in Knowledge Discovery and Data Mining. Menlo Park, Calif.: AAAI Press.

[5]. Dr. Gary Parker, vol 7, 2004, Data Mining: Modules in emerging fields, CD-ROM.

[6]. Jiawei Han and MichelineKamber (2006), Data Mining Concepts and Techniques, published by Morgan Kauffman, 2nd ed.

[7]. Literature Review: Data mining, http://nccur.lib.nccu.edu.twlbitstream/1 I 9/3523 I/S/35603 I OS.pdf, retrieved on June 2012.

[8]. H. Mahgoub,"Mining association rules from unstructured documents" in Proc. 3rd Int. Conf. on Knowledge Mining, ICKM, Prague, Czech Republic, Aug. 25- 27, 2006, pp. 167-172.

[9]. S. Kannan, and R. Bhaskaran "Association rule pruning based on interestingness measures with clustering". International Journal of Com puter Science Issues, IJCSI, 6(1), 2009, pp. 35-43.

[10]. M. Ashrafi, D. Taniar, and K. Smith "A New Approach of Eliminating Redundant Association Rules". Lecture Notes in Computer Science, Volume 31S0, 2004, pp. 465 -474.

[11]. P. Tang, M. Turkia "Para llelizing frequent itemset mining with FP-trees". Technical Report titus.compsci.ualr.edu/-ptang/papers/par-fi.pdf, Department of Computer Science, University of Arkansas at Little Rock, 2005.

[12]. M. Ashrafi, D. Taniar, and K. Smith "Redundant Association Rules Reduction Techniques". L ecture Notes in Computer Science, Volume 3S09, 2005, pp. 254 -263.

[13]. M. Dimitrijevic, and Z. Bosnjak "Discovering interesting association rules in the web log usage data". Interdisciplinary Journal of Information, Knowledge, and Management, 5, 20I0, pp.191 -207.

[14]. R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo.: Fast discovery of association rules- In Advances in Knowledge Discovery and Data Mining (1996).

[15]. Z. HONG-ZHEN, C. DIAN-HUI, and, Z. DE-CHEN "Association Rule Algorithm Based on Bitmap and Granular Computing". AIML Journal, Volume (5), Issue (3), September, 2005.

[16]. K. Yun Sing "Mining Non-coincidental Rules without a User Defined Support Threshold". 2009.

[17]. C. Yin-Ling and F. Ada Wai-Chee "Mining Frequent Itemsets without Support Thres hold: With and without Item Constraints". 2004.

[18]. Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. In: Proceedings of the 20thVLDB conference, pp 4S7-499