An Efficient Approach for Extraction of Actionable Association Rules

Prashasti Kanikar
Assistant Professor
Mukesh Patel School of Technology Management & Engineering
JVPD Scheme Bhaktivedanta swami Marg
Vile Parle (w), Mumbai- 400 056

Ketan Shah, PhD.
Associate Professor
Mukesh Patel School of Technology Management & Engineering
JVPD Scheme Bhaktivedanta swami Marg
Vile Parle (w), Mumbai- 400 056

ABSTRACT
Traditional association mining often produces large numbers of association rules and sometimes it is very difficult for users to understand such rules and apply this knowledge to any business process. So, to find actionable knowledge from resultant association rules, the idea of combined patterns is explored in this paper. Combined Mining is a kind of post processing method for extracting actionable association rules from all possible association rules generated using any algorithm like Apriori or FP tree. In this approach, first the association rules are filtered by varying support and confidence levels, then using the interestingness measure Irule , it is decided whether it is useful to combine the association rules or individual rules are more powerful. For experimental purpose, the Combined Mining approach is applied on a survey dataset and the results prove that the method is very efficient than the traditional mining approach for obtaining actionable rules. The scheme of combined association rule mining can be extended for combined rule pairs and combined rule clusters. The efficiency can be further improved by the parallel implementation of this approach.

General Terms
Association Rule Mining

Keywords
Association Rule Mining, Data Mining, Knowledge Discovery in Databases, Pattern Mining.

1. INTRODUCTION
In data mining area, in order to discover the knowledge, the general framework suggested is called as knowledge discovery in databases(KDD). In the context of KDD, the extraction of rules in forms of association rules is a technique that is used frequently. But now, the focus has been shifted from 'valid' and 'understandable' knowledge to actionable knowledge for decision making. A pattern can be called as actionable if a user can act upon it for his advantage. Study of actionable patterns help in delivering expected outcomes in business processes.

The approach used to extract the patterns is known as association rule mining. As large numbers of association rules are often produced by association mining algorithms, sometimes it can be very difficult for decision makers to not only understand such rules, but also find them a useful source of knowledge to apply to the business processes. In other words, association rules can only provide limited knowledge for potential actions. Therefore, there is a strong and challenging need to mine for more informative and comprehensive knowledge.

Generally ,enterprise data mining applications, such as mining public service data and telecom fraudulent activities, involve complex data sources, particularly multiple large scale, distributed, and heterogeneous data sources embedding information about business transactions, user preferences, and business impact. In these situations, business people certainly expect the discovered knowledge to present a full picture of business settings rather than one view based on a single source. With the accumulation of ubiquitous enterprise data, there is an increasing need to mine for such informative knowledge in complex data.

To present associations in an interesting and effective way, and in order to find actionable knowledge from resultant association rules, the idea of combined patterns is used. Combined patterns comprise combined association rules, combined rule pairs and combined rule clusters. The resultant combined patterns provide more interesting knowledge and more actionable results than traditional association rules.

1.1 Contribution of this paper
Here , we focus on post-processing of association rules with interestingness measure Irule. The contributions of the project work are:
i) Generation of association rules for given support and confidence values.
ii) Extraction of actionable association rules and rejection of less important combined association rules.
iii) Computing the percentage reduction achieved through combined mining approach using Irule measure.
iv) A comparative study of results obtained. The results are used to compare and discuss the behavior of interestingness measure (Irule) on varying support and confidence for a survey dataset.

2. LITERATURE REVIEW
Pattern mining is a data mining method that involves finding existing patterns in data. In this context *patterns* often means association rules. To select interesting rules from the set of all possible rules, constraints on various measures of significance and interest can be used. The best-known constraints are minimum thresholds on support and confidence.

The **support** of an itemset A is defined as the proportion of transactions in the data set which contain the itemset. The **confidence** of a rule is defined as conf(A=>B) = support (A U B)/support(A).

2.1 Recent approaches for Pattern Mining

Zhiwen Yu, Xing Wang ,Hau-San Wong and Zhongkai Deng have proposed a new pattern mining algorithm based on local distribution. The major contribution is a local distribution clustering algorithm based on the normalized cut approach using the local patterns from challenge datasets [4]. Shigeaki Sakurai, Youichi Kitahara and Ryohei Orihara proposed the sequential interestingness as a new evaluation criterion that evaluates a sequential pattern corresponding to the interests of analysts [5]. Unil Yun and John J. Leggett's approach is to push the weight constraints into the sequential pattern growth approach while maintaining the downward closure property [6].

2.2 Recent approaches of association rule mining

Association Rule Mining has always been an attractive area for researchers. Yanchang Zhao, Huaifeng Zhang, Fernando Figueiredo, Longbing Cao,Chengqi Zhang have proposed a technique to discover combined rules on multiple databases and applied to debt recovery in the social security domain[4]. Lingjuan Li, Min Zhang have proposed the data mining strategy based on cloud computing from the theoretical view and practical view [8]. Ashish Mangalampalli and Vikram Pudi have proposed the method that deals with the detection of brain tumor in the CT scan brain images.The preprocessing technique applied on the images eliminates the inconsistent data from the CT scan brain images[9].

3. COMBINED MINING

Traditional association rule mining can only generate simple rules. But the simple rules are often not useful, understandable and interesting from a business perspective. Thus, Zhao et al. [1] proposed combined association rules mining, which generated through further extraction of the learned rules. In other words, to present associations in an effective way, and in order to discover actionable knowledge from resultant association rules, the idea of combined patterns is used. A pattern could be drawn using multi-feature, multi-source or multi-method approach. In multi-feature method, the data can be transactional or categorical or geographical.In this paper the focus is on generation of Combined Rules under multi-feature category.

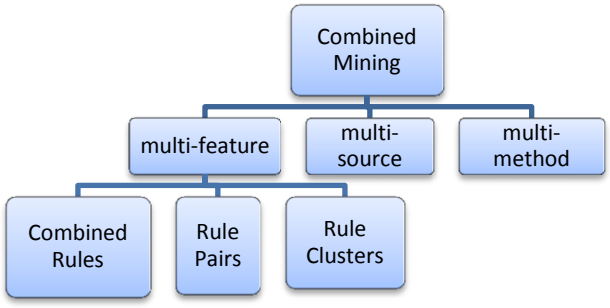

Fig. 1: Classification of Combined Mining Approach

3.1 Traditional Vs Combined Mining

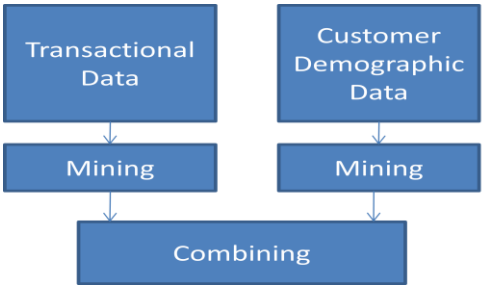

Fig. 2: Representation of traditional mining approach

Suppose our data is located at two places. At one place we have transactional data and at other place we have customer demographic data. According to traditional approach, the data sets are mined for association rules first and then the results are combined.

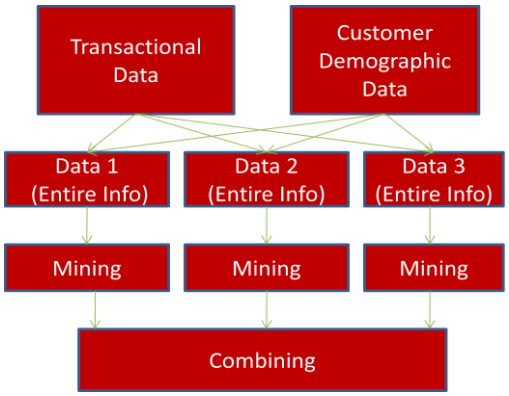

Fig. 3: Representation of combined mining approach

In the same scenario ,according to combined approach, the data is classified based on some similarity measure (region or state or language etc.) and then stored at various locations. The difference is ,here, all the groups have all the data attributes. Then all the groups are mined for association rules and finally the results are combined.

3.2 Interestingness Measures

3.2.1 Lift

A few years after the introduction of association rules, researchers started to realize the disadvantages of the confidence measure by not taking into account the baseline frequency of the consequent. Therefore, the lift (also called interest) measure was introduced:

$$I = \frac{P(X \cap Y)}{P(X)*P(Y)}$$

Since *P(Y)* appears in the denominator of the interest measure, the interest can be seen as the confidence divided by the baseline frequency of *Y*.

3.2.2 Irule

Based on traditional supports, confidences and lifts, a new measure is designed for measuring the interestingness of combined association rules.

*I*rule indicates whether the contribution of U (or V) to the occurrence of T increases with V (or U) as a precondition. Therefore, "*I*rule < 1" suggests that $U \cap V \rightarrow T$ is less interesting than $U \rightarrow T$ and $V \rightarrow T$. The value of *I*rule falls in [0,+∞). When *I*rule > 1, the higher *I*rule is, the more interesting the rule is. Therefore, this measure is more useful than the traditional confidence and lift [18].

$$I\text{rule}\,(U \wedge V \rightarrow T) = \frac{Lift\,(U \wedge V \rightarrow T)}{Lift\,(U \rightarrow T)\,Lift\,(V \rightarrow T)}$$

3.3 Steps for combined mining

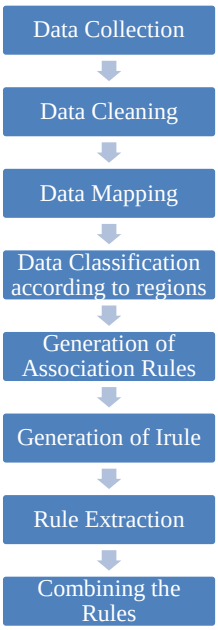

Fig. 4: Steps for Combined Mining

4. DATA SET DESCRIPTION

The data set considered here is a survey database [11] to find out quality rules in order to make decisions to make the travelling convenient for the people of that particular region. The data set consists of following attributes and the answers of Questions asked to people of that area.

Table I : Dataset considered

ATTRIBUTE	DESCRIPTION
Gender	Male or Female
Age	Given in years
Marital	Marital Status of respondant
Level of Education	Highest Level of Education Achieved
Gross Income	Gross Income of respondant
Region	Region where respondant lives
Question 1	Method of Transport to do Main Shop
Question 2	Time Taken to Travel to do Main Shop
Question 3	Ease of Travel to Main Shop

5. EXPERIMENTATION PERFORMED

5.1 Data Mapping

The textual data is mapped to numeric values in order to make the computations smoother.

Table II : Mapping details

Gender 1-female 2- male **Marital status** 3- single 4- married 5- divorced 6- widowed 7- separated **Gross income** 31- Less than £2600 32- £2600 to less than £5200 33- £5200 to less than £10400 34- £10400 to less than £15600 35- £15600 to less than £20800 36- £20800 to less than £28600 37- £28600 to less than £36400 38- £36400 or more 41- Refused 42- Unknown **Method of travel** 51- Car 52 -Public Transport 53- Don't Shop 54- On Foot 55- Other 56- Bicycle 57- Motorbike/ Moped	**Time taken** 61- 5 mins or less 62- 6-10 mins 63- 11-20 mins 64 -21-30 mins 65 -31-45 mins 66- 45 + mins 70- N/A **Ease** 71- Very Easy 72 -Fairly Easy 73- Fairly Difficult 74- Very Difficult 80- N/A **Region** 101- The North 102 -South West 103 -Midlands/East Anglia 104 -London 105 -South East 106 -Wales 107 -Scotland

5.2 Data Classification

After mapping , the data is classified among the regions.

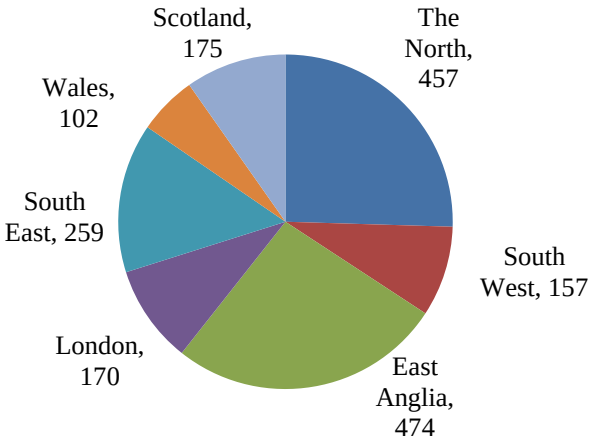

Fig. 5: Region based classification of 1794 records

5.3 Algorithm

infile is the input file that contains association rules for any support and confidence level and poolfile is the file that contains all possible association rules and their support and confidence values.

While the infile has next line
read a line of infile
separate the antecedent part
find antecedent1 and antecedent2
find consequent
find rule lift
for antecedent1 and consequent combination, search in pfile
read the corresponding lift and store it as ldenom
for antecedent2 and consequent combination, search in poolfile
read the corresponding lift and store it as rdenom
compute dlift as ldenom*rdenom
compute Irule as lift/dlift
if(Irule>1)
print rule status as EXTRACTED
print the rule and all related variables
else
print rule status as REJECTED
print the rule and all related variables

print the summary report.

Here, table III shows the results using tarditional approach. Initially the data is located at two places. This data is mined indepently for association rules at both locations. Results do not show any rule for data1 and there is only 1 rule for data2. Hence,its not possible to make decisions out of the only one rule found. So, on the basis of results we can say that there is a need for improved approach.

Table III: Results using traditional approach

Support(%)	Confidence(%)	dataset	Rules generated	CM applied	Poolfile with S5C5	Rules extracted	Rules rejected	Reduction achieved	Time(millisec)
10	10	data1	20	0	84	-	-	-	-
10	10	data2	58	12	160	**1**	11	91	312

Here, table IV shows the results using Combined Mining approach. Initially the data is classified based on regions into data groups group1 to group7. Then this data is mined for association rules and finally the combined approach is used for extracting the actionable rules.

Table IV: Results using Combined Mining approach

Support(%)	Confidence(%)	dataset	Rules generated	CM applied	Poolfile with S5C5	Rules extracted	Rules rejected	Reduction achieved	Time(millisec)
10	10	G1	186	55	991	11	44	80	2777
20	20	G1	20	0	-	-	-	-	-
10	10	G2	362	129	1544	22	107	82	9578
20	20	G2	60	16	1544	2	14	87	1297
30	30	G2	10	0	-	-	-	-	-
10	10	G3	242	81	1148	4	77	95	4438
20	20	G3	38	8	1148	0	8	100	547
30	30	G3	8	0	-	-	-	-	-
10	10	G4	200	50	1107	8	42	84	3109
20	20	G4	20	0	-	-	-	-	-
10	10	G5	322	106	1486	23	83	78	8187
20	20	G5	44	9	1486	1	8	88	781
30	30	G5	12	0	-	-	-	-	-
10	10	G6	220	58	1338	10	48	82	4219
20	20	G6	34	6	1338	0	6	100	484
30	30	G6	6	0	-	-	-	-	-
10	10	G7	268	91	1440	26	65	71	6188
20	20	G7	42	9	1440	4	5	55	703
30	30	G7	6	0	-	-	-	-	-

Please note that , to make the data more readable, in this table instead of group1, G1 is written. The same convention is followed for all other groups.

5.4 Experimental Setup

For performing the experiment, eight machines with intel core 2 processor and 4 GB –RAM are used. All the machines are connected through local area network.

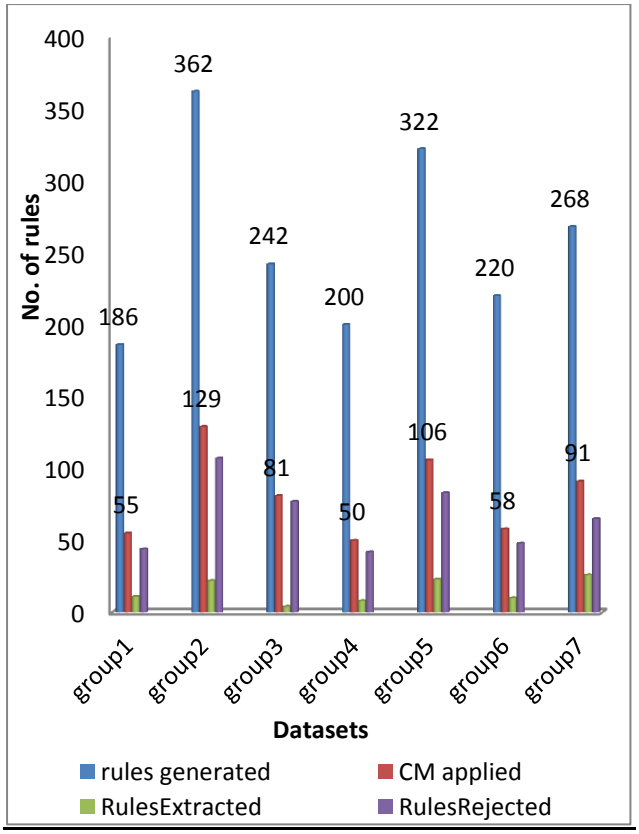

Fig. 6 : Plot of rules generated ,Combined Mining applied,Rules extracted and Rules rejected for Support=10% and confidence=10%

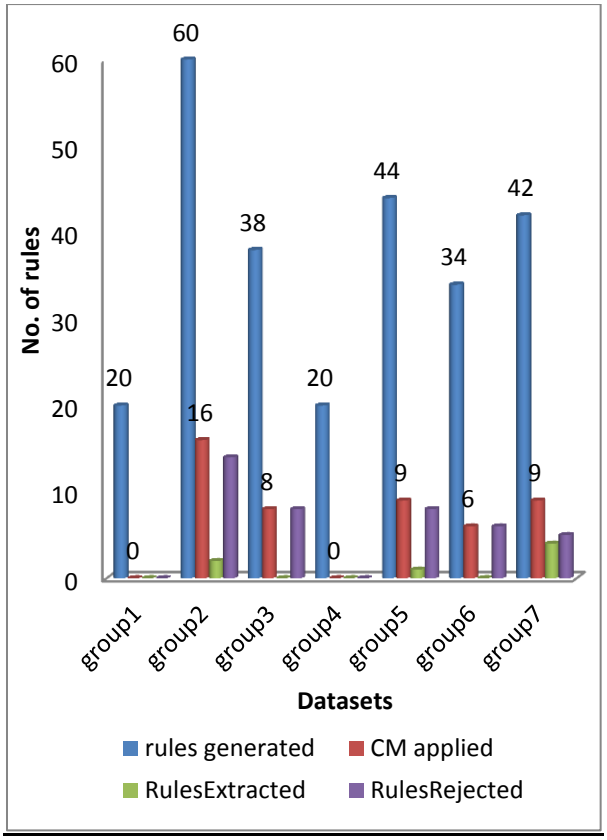

Fig. 8 : Plot of rules generated ,Combined Mining applied,Rules extracted and Rules rejected for Support=20% and confidence=20%

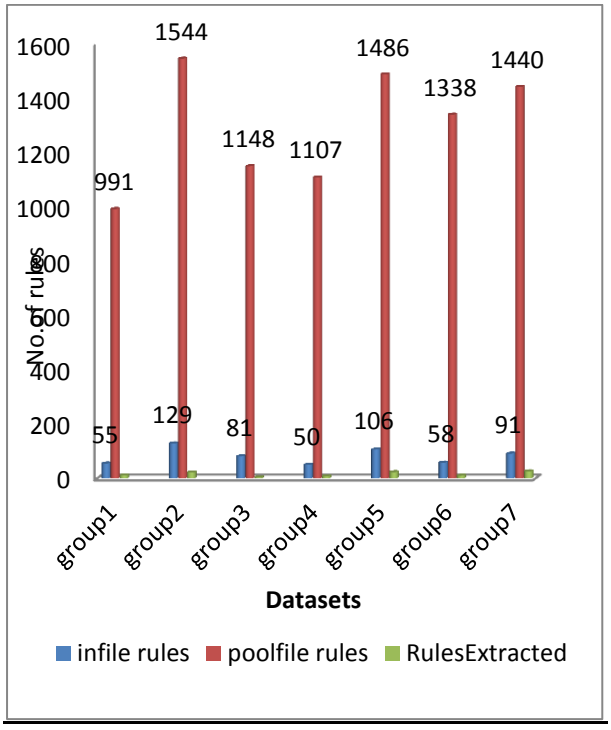

Fig. 7 : Plot of input file rules,pool file rules and rules extracted for Support=10% and confidence=10%

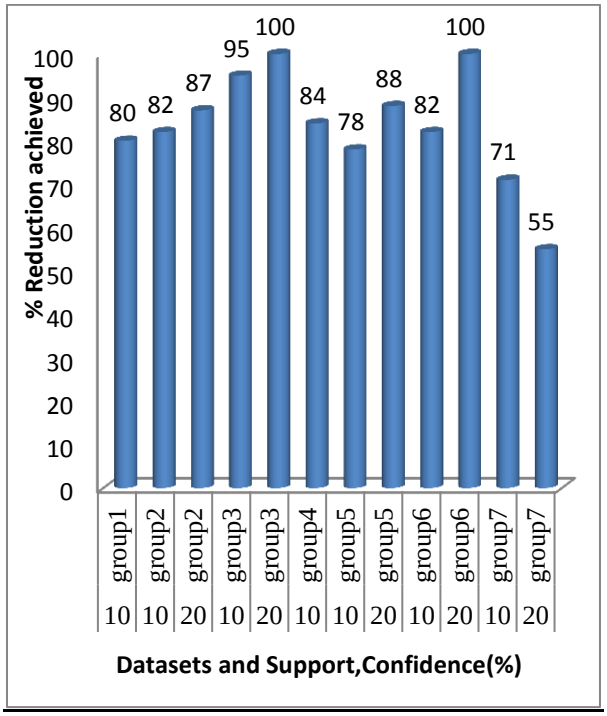

Fig. 9 : Overall plot of reduction achieved for variety of datasets

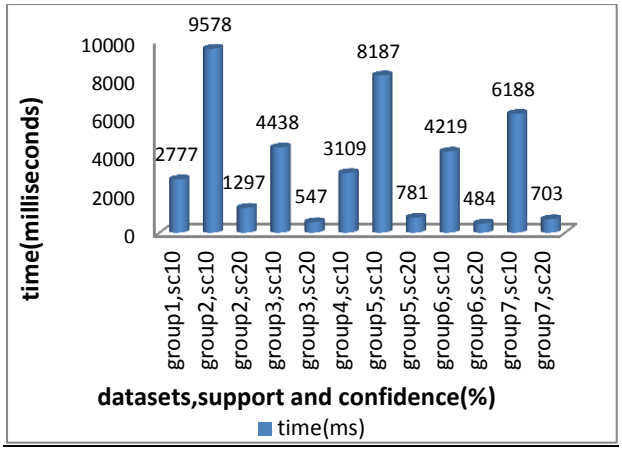

Fig. 10 : Overall plot of time required for Combined Mining for variety of datasets, support and confidence values

Fig. 6 and 8 show the plots for total rules generated ,no. of rules on which Combined Mining is applied,Rules extracted and Rules rejected for various support and confidence levels.Fig. 9 shows the reduction achieved for all 7 data groups and fig.10 shows that time required for rule extraction depends on sizes of input file as well as poolfile.

6. CONCLUSIONS

The idea of combined patterns is applied on survey data sets. The concepts of combined association rules are applied using the interestingness measure Irule. The measure Irule helps us to determine whether the two rules should be combined or not. According to results, the traditional approach shows the reduction of rules by 8.33% while combined approach shows a minimum reduction of 55% and maximum reduction of 100%. Hence, the combined approach is more efficient than the traditional association rule mining approach. The derived combined patterns are more useful and actionable than traditional simple association rules.

7. FUTURE SCOPE

The concept of combined association rules can be extended for combined rule pairs and combined rule clusters. A combined rule pair is composed of two contrasting rules and combined rule clusters are built from combined association rules. The combined patterns provide more interesting knowledge and more actionable results than traditional association rules. To improve the efficiency of combined rule generation scheme, the data could be distributed across multiple machines and then the computations can be carried out in parallel.

8. REFERENCES

[1] Y. Zhao, H. Zhang, L. Cao, C. Zhang, and H. Bohlscheid, “Combined pattern mining: From learned rules to actionable knowledge,” in *Proc. AI*, 2008, pp. 393–403.

[2] Longbing Cao, Huaifeng Zhang,Yanchang Zhao, Dan Luo and Chengqi Zhang, “Combined Mining: Discovering Informative Knowledge in Complex Data”, IEEE Transactions on Systems, Man and Cybernetics—part B: CYBERNETICS, vol. 41, no. 3, june 2011, pp. 699-712.

[3] Za¨ıane, O.R., Antonie, M.-L.” On pruning and tuning rules for associative classifiers”, KES 2005. LNCS (LNAI), vol. 3683, pp. 966–973.

[4] Zhiwen Yu, Xing Wang ,Hau-San Wong and Zhongkai Deng,”Pattern mining based on local distribution”, IEEE,2008,pp. 584-588.

[5] Shigeaki Sakurai, Youichi Kitahara, and Ryohei Orihara,” Sequential Pattern Mining based on a New Criteria and Attribute Constraints”, IEEE, 2007,pp. 516-521.

[6] Unil Yun, and John J. Leggett,” WSpan: Weighted Sequential pattern mining in large sequence databases”, 3rd International IEEE Conference Intelligent Systems, September 2006, pp. 512-517.

[7] P. S. Wang, "Survey on Privacy Preserving Data Mining", JDCTA: Journal of Digital Content Technology and its Applications, Vol. 4, No. 9, pp. 1 -7, 2010.

[8] Lingjuan Li, Min Zhang , ”The Strategy of Mining Association Rule Based on Cloud Computing” , International Conference on Business Computing and Global Informatization , 2011,pp.475-478.

[9] Ashish Mangalampalli, Supervised by Vikram Pudi,” Fuzzy Associative Rule-based Approach for Pattern Mining and Identification and Pattern-based Classification”, *WWW 2011*, March 28–April 1, 2011, Hyderabad, India,pp. 379-383.

[10] T. Brijs, K. Vanhoof, G. Wets,” Defining Interestingness for Association Rules”, International Journal "Information Theories & Applications, Vol.10,pp. 370-375.

[11] http://www.rsscse.org.uk/stats4schools

[12] Goulbourne, G., Coenen, F. and Leng, P. (2000), “Algorithms for Computing Association Rules Using a Partial-Support Tree”, Journal of Knowledge-Based Systems, Vol (13), pp141-149.

[13] Coenen, F., Goulbourne, G. and Leng, P., (2003). “Tree Structures for Mining association Rules”, Journal of Data Mining and Knowledge Discovery, Vol 8, No 1, pp25-51.

[14] W. J. Frawley, G. Piatetsky-Shapiro & C. J. Matheus ,”Knowledge discovery in databases: an overview", Knowledge Discovery in Databases (1991), pp1-27.

[15] U. M. Fayyad, G. Piatetsky-Shapiro & P. Smyth, ”From data mining to knowledge discovery", Advances in Knowledge Discovery and Data Mining (1996), pp. 1-34.

[16] Agrawal, Rakesh; Imielinski, Tomasz; Swami, Arun,” Mining Association Rules Between Sets of Items in Large Databases”, SIGMOD Conference 1993,pp.207-216

[17] Prashasti Kanikar, Dr. Ketan Shah, “Extracting Actionable Association Rules from Multiple Datasets”, International Journal of Engineering Research and Applications (IJERA), May 2012.