A Parallel Approach to Combined Association Rule Mining

Zaid Makani
B.Tech Computers, 4th year
Mukesh Patel School of Technology Management and Engineering
JVPD Scheme Bhaktivedanta Swami Marg
Vile Parle (W), Mumbai-400056

Sana Arora
B.Tech Computers, 4th year
Mukesh Patel School of Technology Management and Engineering
JVPD Scheme Bhaktivedanta Swami Marg
Vile Parle (W), Mumbai-400056

Prashasti Kanikar
Assistant Professor
Mukesh Patel School of Technology Management and Engineering
JVPD Scheme Bhaktivedanta Swami Marg
Vile Parle (W), Mumbai-400056

ABSTRACT
Data Mining carried out using traditional methodologies of Support-Confidence framework and Association Rule Mining yield an enormous number of inefficient rules or patterns in a certain amount of time. In this paper, a parallel approach to Combined Mining has been implemented that not only generates rules which are "actionable" but also does so in a time period that is lesser than that of the traditional approach. These implementations are carried out on datasets at different locations consisting of multiple related data items and are independent of each other. The results of an Apriori algorithm is fed as an input to Combined Mining so as to generate more useful patterns for the process of business decision making. The results highlight that the two objectives of "Actionability" and "Efficiency with respect to time" are achieved using parallel approach to Combined Mining rather than compromising on one of them as in serial approach.

General Terms
Association Rule Mining, Data Mining, Pattern Mining

Keywords
Combined Mining, Parallel Combined Mining, Sequential Combined Mining

1. INTRODUCTION
The field of Data Mining has been extensively researched upon to extrapolate improved and more favorable results. It involves extraction of Association Rules or patterns obtained by carrying out the process of Association Rule mining on a 'Pre Processed' dataset and yielding a set of values on which post processing is carried out. This entire process of carrying out Pre-processing + Data Mining + Post- Processing is known as Knowledge Discovery in Databases or KDD.

Figure 1: KDD Process flow

As here the focus is on "Actionability", association rules can be called as actionable if a user can act upon it for his advantage. These help is attaining likely results that provide immense value to day to day business processes.

An approach to improve the patterns generated by traditional mining algorithms, a method called 'Combined Mining' was proposed by Longbing Cao, Huaifeng Zhang, Yanchang Zhao, Dan Luo, and Chengqi Zhang. [2] Its main aim is to 'Increase the % Reduction of the rules generated' which is done on the basis of certain criteria called "Interestingness Measures". This % Reduction gives an idea of how many unwanted rules have been successfully rejected, i.e. more the % Reduction, better the results and vice versa. There are a large amount of measures available today, like Irule, Lift, and Rule Interest which can be applied on the rules generated by the traditional approach to extract more "actionable rules". The rules which do not satisfy the particular criteria stand rejected.

There is a tradeoff between 'Time' and '% Reduction' when the traditional approach and the Sequential/Serial Method of Combined Mining are compared. It does give more % Reduction however time required is also more. To overcome this tradeoff, the parallel approach to Combined Mining has been implemented which gives a high % Reduction as well as does so in a lesser amount of time.

Contributions of this paper:

This paper highlights the following aspects in relation to the field of Pattern Mining and Association Rule Mining:

i. The role of combined mining in improving the "usefulness" of rules generated by extraction of more actionable rules using interestingness measures called Lift and Irule.
ii. Implementation of the sequential approach to Combined Mining method to generate rules and calculate the % Reduction and time taken for multiple values of Support-Confidence.
iii. Implementing the Parallel approach to Combined Mining to calculate the % Reduction and time taken for multiple values of Support-Confidence.
iv. A comparative study on traditional approach to mining, sequential approach and parallel approach in terms of time taken and % Reduction for varying levels of Support-Confidence.

2. LITERATURE SURVEY
Data Mining is one of the most imperative and crucial steps in the KDD Process. It encompasses several tasks that can be termed as "Data Mining Techniques". The most researched

and indispensable technique is known as "Pattern Mining", a method to discover hidden patterns in data. These patterns represent relations between the elements present in the database, also termed as 'Association Rules'. Therefore, to discover unique, significant, valuable and more importantly "interesting" information, Association Rule Mining is carried out based on two important criteria or thresholds called "Support" and "Confidence".

Support of an item X is the proportion of transactions that contain X.

Support (X) =No. of transactions that contain X

Confidence of a rule X→Y indicates the frequency of association or reliability

Confidence(X) =P(X ∩ Y)/P(X)

2.1. Research on Interestingness Measures:

Jianhua Liu, Xiaoping Fan and Zhihua Qu have stated that the classical model of the Support-Confidence framework poses many issues in extracting the best possible patterns in data. The major contribution of the authors is the proposal of a new sufficiency measure to discover more interesting patterns in data. Several such interestingness measures such as Lift and Irule have been brought into the limelight by Longbing Cao, Huaifeng Zhang, Yanchang Zhao, Dan Luo, and Chengqi Zhang. [2]

2.2. Research on Combined Mining:

Longbing Cao, Huaifeng Zhang, Yanchang Zhao, Dan Luo, and Chengqi Zhang have introduced the concept of "Combined Mining" to improve upon the approach of traditional mining followed so far. They propose combined mining as a general approach to mining for informative patterns combining components from either multiple data sets or multiple features or by multiple methods as per requirement. [2] Also, Prashasti Kanikar and Ketan Shah have implemented an efficient approach for extraction of actionable association rules using the serial approach proving experimentally a part of what was proposed in the paper on "Combined Mining". [12]

3. COMBINED MINING

3.1. INTERESTINGNESS MEASURES

The concepts of Lift and Irule have been emphasized by Prashasti Kanikar and Dr. Ketan Shah. [12] To throw some light on how they work, the concepts have been elucidated here with the help of an example.

3.1.1. Lift

Consider a sample dataset with four transactions containing items as listed below:

TID 1	A	B	C
TID 2	A	C	
TID 3	A	D	
TID 4	B	E	F

Table 1: Sample dataset with Transactions and their respective items

Using the Apriori Algorithm, few of the association rules generated are as follows: support=25% and confidence= 50%:

('A', 'C') ==> ('B'), 50% confidence
('A', 'B') ==> ('C'), 100% confidence
('B', 'E') ==> ('F'), 100% confidence
('B', 'F') ==> ('E'), 100% confidence

To calculate lift of a rule:

Lift(X→Y) =P(X ∩ Y) / P(X)* P(Y)

Rule	P(X ∩ Y)	P(X) * P(Y)	Lift
{A,C}→B	1/4	2/4 * 2/4	1
{A,B}→C	1/4	1/4 * 2/4	2
{B,E}→F	1/4	1/4 * 1/4	4
{B,F}→E	1/4	1/4 * 1/4	4

Table 2: Example to calculate lift of given rules

3.1.2. Irule

For the same examples as in Table 1, once we have calculated the lift, we now calculate the Irule using the formula:

Irule (PQ→R) = Lift (PQ→R) / Lift (P→R) * Lift (Q→R)

Rule	Lift (PQ→R)	Lift (P→R)	Lift (Q→R)	Irule
{A,C}→B	1	0.666	1	1.5
{A,B}→C	2	1.333	1	1.5
{B,E}→F	4	2	4	0.5
{B,F}→E	4	2	4	0.5

Table 3: Example to calculate lift of given rules

Since the rules {A, C} →B and {A, B} →C have an Irule > 1 therefore they are accepted.

However, rules {B, E} →F and {B, F} →E have an Irule < 1 therefore they are rejected.

3.2. METHODS OF IMPLEMENTATION OF COMBINED MINING

Combined Mining can be classified on the basis of the following two criteria:

i. Procedure adopted
ii. Technique of implementation

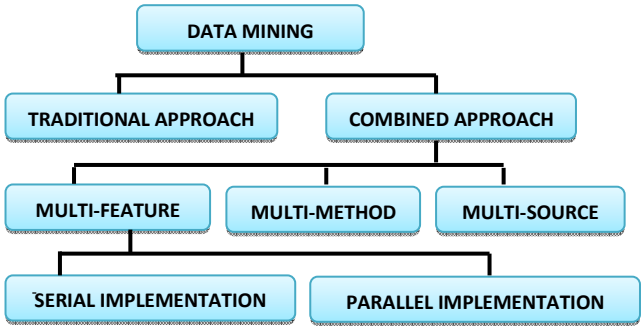

Figure 2: Classification of approaches in Data Mining

In this paper the Multi Feature approach of Combined Mining as proposed by Longbing Cao, Huaifeng Zhang, Yanchang Zhao, Dan Luo, and Chengqi Zhang has been used.

3.3. MULTI FEATURE COMBINED MINING

Multi Feature Combined Pattern Mining comprises of a combined pattern, which is composed of heterogeneous features of different data types, such as ordinal, numerical, binary, and categorical, or of different data categories, such as transactions, customer demographics and time series. A combined pattern is composed of heterogeneous features of different data types.

Steps in the combined approach to mining and traditional approach to mining are shown in Figure 3 below:

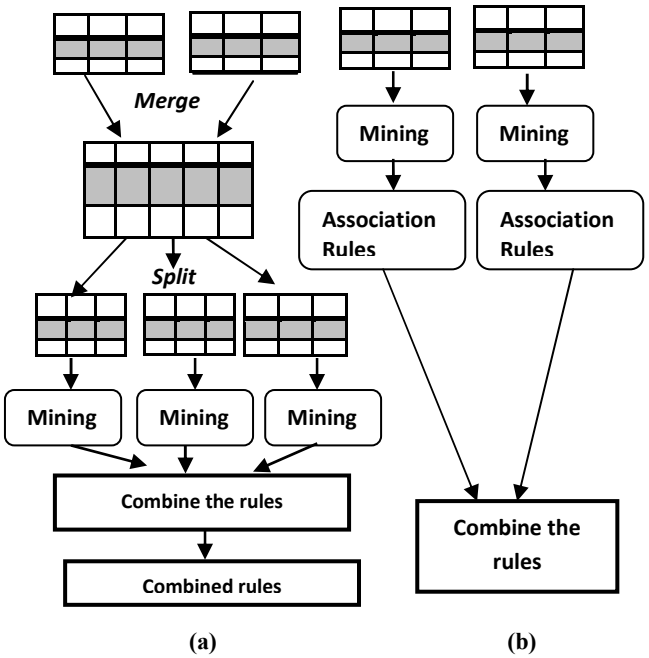

Figure 3: (a) Steps in Combined Mining approach (b) Steps in traditional mining approach

3.3.1. ALGORITHM

*Assume Antecedent1=Ant1 and Antecedent2=Ant2
Start
Accept the datasets, minimum support % and confidence % from the user
Merge the datasets into one (column wise) to recreate the entire database at one location
Split the datasets into separate parts depending on attributes common to all datasets (In this case, Regions)
Repeat following steps for all regions
    Apply Apriori algorithm to compute frequent itemsets, followed by Association Rule Mining to generate Association Rules for the region
    Separate out rules with Antecedent 2
    Apply Irule on them
        Calculate lift of the complete rule
        Calculate lift of antecedent and consequent
        Use formula Irule (rule):
        Lift (rule) / (Lift (Ant1→Consequent)*Lift (Ant2→Consequent))
            If Irule>1, rule is accepted
            Else if Irule<1, rule is rejected
Display results
Stop

4. DEPICTION OF THE DATA SET

The dataset used is a survey conducted on a region basis, used to get an analysis of the forms of commutation and travel that is most convenient to the residents of that region. It can be used to answer the questions like:
i. For a shorter distance, which mode of transport is preferred?
ii. How does the marital status affect ones' purchase of a vehicle?
iii. How much more popular is a vehicle on the basis of gender or age?

Number of records: 3000

Attribute	Description	Values
Region	Region where person lives	Europe, N. America, Pacific
Marital Status	Marital Status of the person	Married, Single
Gender	Gender	Male, Female
Yearly	Yearly Income	10000-170000
Children	No. of Children	0,1,2,3,4,5
Education	Qualification	Partial College, Bachelors, Partial High School, High School
Occupation	Occupation of the person	Professional, Management,Clerical
Home Owner	Home owner or not	Home Yes, Home No
Cars	No. of cars	0,1,2,3,4
Distance	Distance	0-1, 1-2, 2-, 5-10, 10+ Miles
Age	Age	{25-96}
Bike Buyer	Buys a bike	{Bike Yes, Bike No}

Table 4: Dataset description

5. APPROACHES TO COMBINED MINING

5.1. Serial implementation

In this method of implementation, using the Multi Feature algorithm, the three regions are mined on one single machine one after the other, sequentially. As a result of this, the time taken to mine all three regions is quite large as compared to the traditional approach to mining. It results in an average time of 4012 milliseconds. However, the average % Reduction obtained here is 74.3%. It is considered a lot better than the traditional method where the average % reduction is 37.4% in an average time of 1033 milliseconds.

5.2. Parallel implementation

This implementation involves using the of Client Server model. It involves the following steps:
1. The three regions (databases) are placed on three separate computers. Each of these computers acts as a client.
2. Each of these clients runs the multi feature script on their respective regions (databases).

3. The output generates the combined rules as well as the computation time/execution time of the program.
4. Another computer acts as the server and its purpose is to receive the results and display them to the user.
5. The clients send the results of their execution times to the server via Client-Server scripts.
6. The server accepts these results and displays them to the user

As a result of this, the time taken to mine all three regions is much lesser as compared to traditional approach to mining as well as the serial approach to Mining. The average % Reduction here is 74.3% however average time for computation required is 2437.2 milliseconds.

Benefits of parallel implementation:

i. % Reduction here is 74.3% as opposed to traditional mining where it is 37.4% hence it is a lot better than the traditional method.
ii. Efficiency with respect to time required here is high as it requires 2437.2 milliseconds, whereas the parallel implementation takes 4012 milliseconds.

6. ANALYSIS OF EXPERIMENTATION

The implementation conducted involves thedataset to be classified into the following 3 regions.

1. Europe: 895 records (F0)
2. Pacific: 494 records (F1)
3. North America: 1611 records (F2)

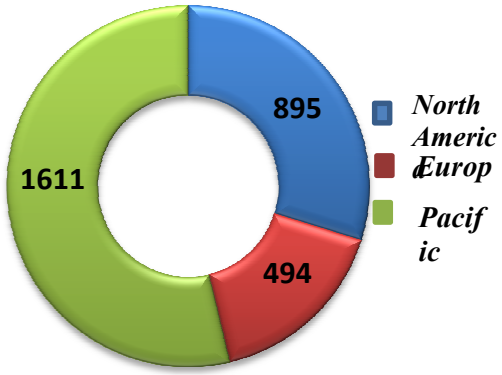

Figure 4: Pie chart depicting region wise distribution of records

EXPERIMENTATION PERFORMED

i. Dataset as depicted above is pre-processed as per requirement.
ii. Traditional approach of mining is applied on it for varying values of Support-Confidence to analyse the difference
iii. Sequential approach of Combined Mining is conducted by classifying the database on the basis of region again for varying levels of Support-Confidence.
iv. The parallel approach to Combined Mining has then been implemented for varying levels of support-confidence.
v. The results of all the above cases have been tabulated to get a clear idea about the '% Reduction' and 'Efficiency with respect to time' as a comparative study.

7. RESULTS AND DISCUSSION

The comparative study of all three techniques has been tabulated below.

7.1. Traditional approach

This has been carried out individually for the two datasets which contain related items at are initially stored at different locations.

Support-Confidence %	Dataset	Rules Generated	CM Applied	Rules Extracted	Rules Rejected	% Reduction Achieved	Time(ms)
10 – 10	Ds1	14	6	6	-	-	3046
20 - 20	Ds1	28	0	-	-	-	677
30 – 30	Ds1	8	0	-	-	-	441
10 – 10	Ds2	28	12	7	5	41.6	1940
20 – 20	Ds2	18	9	6	-	33.3	726
30 – 30	Ds2	6	3	3	-	-	523

Table 5: Results obtained by Traditional Mining

7.2. Serial approach to Combined Mining

This has been carried out on different regions after the database consisting of ds1 and ds2 has been split into three regions.

Support %	Confidence%	Dataset	Rules Generated	CM Applied	Rules Extracted	Rules Rejected	% Reduction Achieved	Time (ms)
10	**10**	F0	126	21	6	15	71.428	8,836
20	**20**	F0	392	168	55	113	67.261	1,928
30	**30**	F0	14	6	3	3	50	597
10	**10**	F1	126	21	5	16	76.190	5360
20	**20**	F1	240	80	22	58	72.5	1201
30	**30**	F1	28	12	1	11	91.66	382
10	**10**	F2	434	105	52	53	50.476	14008
20	**20**	F2	90	30	13	17	56.666	2733
30	**30**	F2	98	42	7	35	83.33	1069

Table 6: Results obtained by serial approach to Combined Mining

7.3. Parallel approach to Combined Mining

This approach has been the primary focus of this paper. It proves how it can overcome the drawbacks of the other approaches to give more efficient results.

In this, the implementation has been carried out 3 region based datasets, each deployed on a separate machine. They acts as clients interact with the server to return the computation time each of them takes for variable levels of support-confidence.

Support %	Confidence%	Rules Generated	CM Applied	Rules Extracted	Rules Rejected	% Reduction Achieved	Time (ms)
10	**10**	126	21	6	15	71.42	8,220
20	**20**	392	168	55	113	67.26	1,503
30	**30**	14	6	3	3	50	566
30	**80**	3	1	1	0	0	608
40	**80**	126	21	5	16	76.19	366

Table 7: Results obtained by parallel approach to Combined Mining

7.4. EFFICIENCY COMPARISON

This table is a comparative analysis on the three techniques in terms of 'Efficiency with respect to time taken'.

Support%	Confidence%	Traditional approach	Serial approach to Combined Mining	Parallel approach to Combined Mining
10	10	3629.7 ms	21,281 ms	8220 ms
20	20	1340.9 ms	5124.6 ms	1503 ms
30	30	670.5 ms	1583.7 ms	566.9 ms
30	80	778.4 ms	1168.99 ms	608.2 ms
40	80	795.2 ms	1013.4 ms	366.2 ms

Table 8: Comparison of the three approaches with respect to time taken

8. GRAPHICAL ANALYSIS

For all the tables listed above, graphs are plotted to analyse pictorially how the two forms of Combined Mining can be understood with respect to parameters like:

i. Varying values of Support-Confidence
ii. % Reduction achieved
iii. Time taken

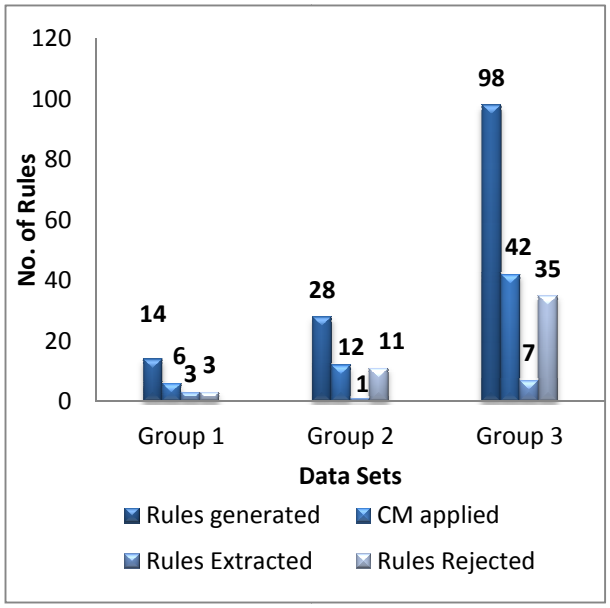

Figure 5: Plot of rules generated, CM applied, Rules extracted, Rules Rejected for Support=30% and Confidence=30%

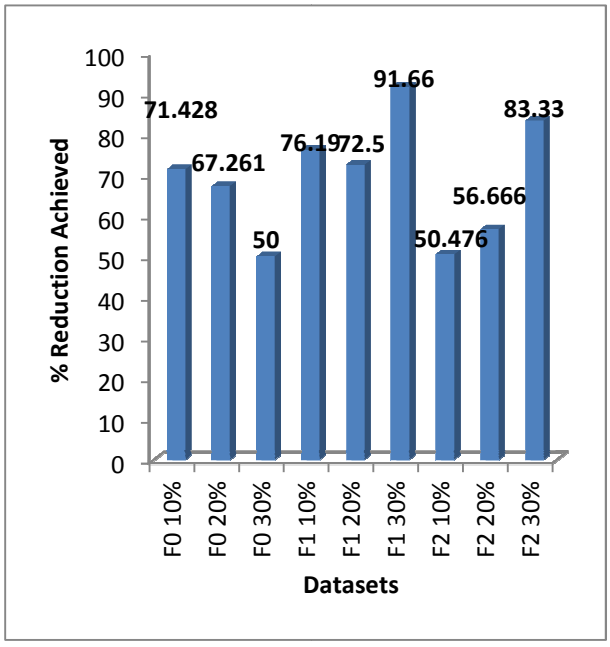

Figure 6: Plot of % Reduction achieved in case of Combined Mining

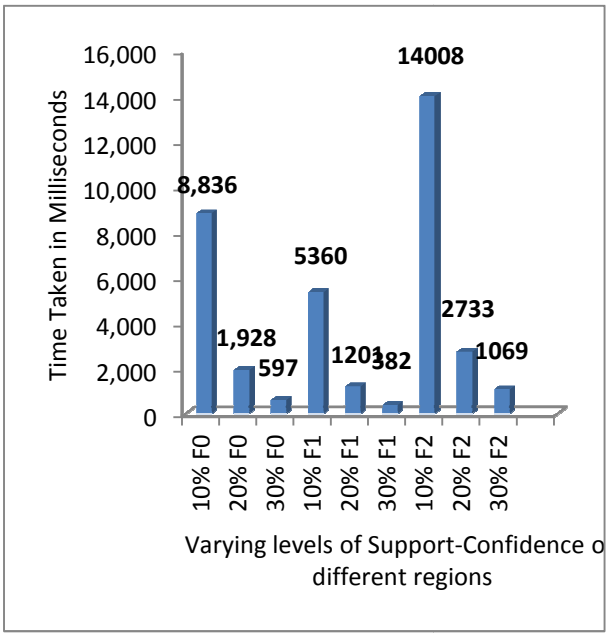

Figure 7: Plot of 'Time taken' in case of serial approach to Combined Mining

In Figure 7, a graph of Time taken versus Varying levels of support and confidence have been plotted. It indicates that the time consumed by the 'Serial approach to Combined Mining' is relatively more.

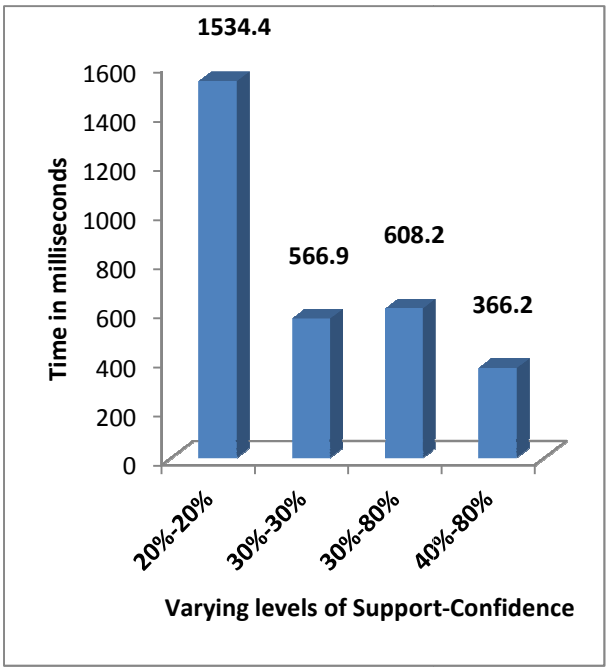

Figure 8: Plot of Time taken in case of parallel approach to Combined Mining

In Figure 8, a graph of Time taken versus Varying levels of support and confidence have been plotted. It indicates that the time consumed by the parallel approach to Combined Mining' is relatively much lesser.

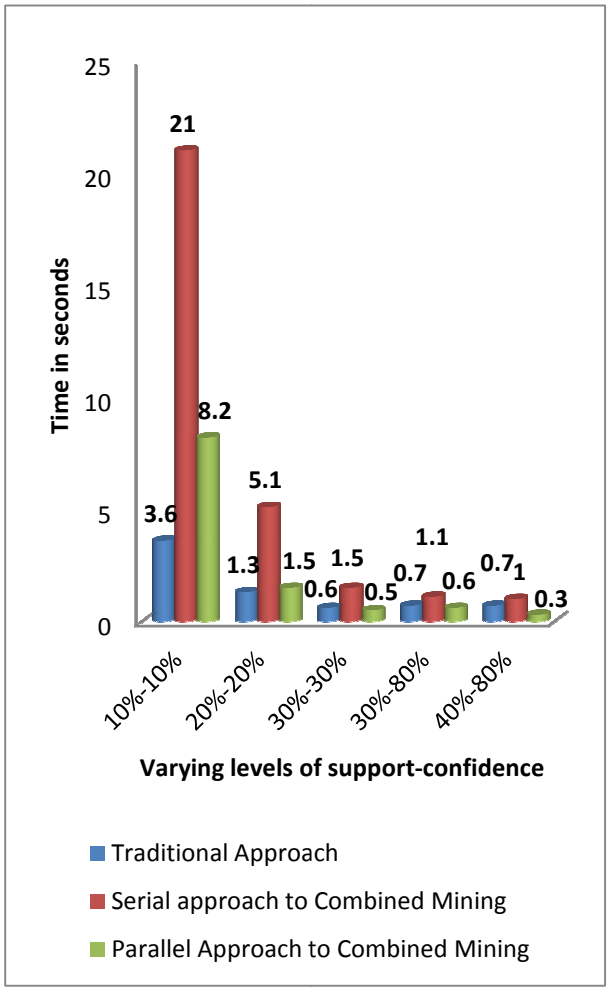

Figure 9: Comparison of time taken in traditional approach, serial approach to Combined Mining and parallel approach to Combined Mining

In Figure 9, a graph of Time taken versus Varying levels of support and confidence have been plotted for all three approaches. It indicates that as the Support-Confidence level increases, the time consumed by the parallel approach to Combined Mining decreases. Therefore, Efficiency with respect to time is achieved through the parallel approach.

9. CONCLUSION

The two approaches of Combined Mining- serial and parallel have been applied on a survey dataset. Certain Interestingness measures like Lift and Irule are used to determine the acceptance of the rule. The results show that, traditional mining shows an average reduction of 37.4% in an average time period of 1033 milliseconds whereas the serial approach of Combined Mining shows an average reduction of 74.3% in an average time period of 4012 milliseconds. Parallel approach on the other hand shows an average reduction of 74.3%, however in an average time of 2437.2 milliseconds. Hence, the parallel approach can be considered the most suitable approach to generate 'Actionable' Association rules in minimal time. However, if the datasets are geographically located very far, it takes more time to process and obtain the results.

10. FUTURE SCOPE

One of the possible directions of research in this field is an implementation of the Multi Method Combined Mining algorithm which involves the use of multiple algorithms to obtain suitable results. Also the incremental approach to Combined Mining, where in Incremental Pair Patterns and Pair Clusters are formed based on their usefulness and significance. It obviates the need to scan the entire database at once, thereby allowing feed 'increments' of the database to a given algorithm.

10. ACKNOWLEDGEMENT

Our sincere gratitude to our mentor and guide, Professor Prashasti Kanikar for all her guidelines and support provided during the course of writing this paper. We would like to express our deepest acknowledgment to the Head of the Department, Dr. Dhirendra Mishra whose constant support and encouragement has led to the successful completion of this paper. We would also like to thank the Dean of our college, Dr. S.Y.Mhaiskar for having provided us with the resources and the opportunity for writing this paper.

11. REFERENCES

[1] WanjunYu, XiaochunWang, Fangyi Wang, Erkang Wang, Bowen Chen, "The Research of Improved Apriori Algorithm for Mining Association Rules", 11th IEEE International Conference on Communication Technology Proceedings, nov 2008, pp. 513-516.

[2] Longbing Cao, Huaifeng Zhang, Yanchang Zhao, Dan Luo, Chengqi Zhang, "Combined Mining: Discovering Information Knowledge in Complex Data", IEEE Transactions on Systems, Man and Cybernetics – Part B: Cybernetics, Vol. 41, No. 3, June 2011, pp.699-712.

[3] Ioan Daniel Hunyadi, "Performance Algorithms in generating Association Rules", International Journal of Mathematics and Computers in Simulation, Issue 3, Vol. 5, 2011.

[4] Zhiwen Yu, Xing Wang, Hau-San Wong and Zhongkai Deng, "Pattern mining based on local distribution", IEEE, 2008, pp. 512-517.

[5] Yanchang Zhao, Huaifeng Zhang, Longbing Cao, Chengqi Zhang and Hans Bohlscheid, "Combined Pattern Mining: From Learned Rules to Actionable Knowledge", in *Proc AI*, 2008, pp. 393-403.

[6] Stephane Lallich, Olivier Teytaud and Elie Prudhomme, "Association rule interestingness: Measure and Statistical Validation"

[7] T. Brijis, K. Vanhoof, G.Wets, "Defining interestingness for Association Rules", International journal Information theories and Applications", Vol. 10.

[8] Jianhua Liu, Xiaoping Fan, Zhihua Qu, "A New Interestingness Measures of Association Rules", Second international Conference on Genetic And Evolutionary Computing, 2008.

[9] Prashasti Kanikar, Dr. Ketan Shah, "Extracting Actionable Association Rules from Multiple Datasets" International Journal of Engineering Research and Applications (IJERA) Vol.2, Issue 3, May-Jun 2012.

[10] Prashasti Kanikar, Dr. Ketan Shah, "An Efficient approach for Extraction of Actionable Association Rules", International Journal of Computer Applications, Volume 54 –No 11, September 2012.

[11] Christophier J. Mathens, Philip K. Chan Gregory Piatetsky- Shapiro, "System for Knowledge discovery in Databases", IEEE TKDE special issue on Learning and Discovery in Knowledge – Based Databases, 1993.

[12] Fengzhang Han," Mining Actionable Combined Patterns with Composite Items", Journal of Convergence Information Technology, Volume 6, Number 4, April 2011.

[13] Pang-Ning Tan, Vipin Kumar, "Interestingness Measures for Association Patterns: A perspective".

[14] Yanxi Liu, "Study on Application of Apriori Algorithm in Data Mining", Computer Modeling and Simulation, 2010. ICCMS '10. Second International Conference, Vol. 3, Jan 2010, pp. 111-114

[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.