Clustering and Pattern Mining of Customer Transaction Data using Apriori Algorithm

Sonali L. Mortale, Manisha J. Darak

Abstract: ***Clustering customer transaction data is an important procedure for analyzing customer behavior in retail and e-Commerce. Clustering of trading data with finding patterns using Apriori algorithm will helps to develop a market strategy and increases the profit. The system uses Apriori algorithm for finding pattern. The input of Apriori algorithm is the output of Customer Transaction Clustering Algorithm. In a system the customer transaction data is presented by using transaction tree and the distance between them is also calculated. Cluster the customer transaction data by using customer transaction clustering algorithm. The system selects frequent customer as representatives of customer groups. Finally, the system forwards the output of clustering to Apriori algorithm for finding patterns.***

Keywords: ***Clustering, Apriori Algorithm, Customer Transaction Clustering Algorithm, Transaction Tree.***

I. INTRODUCTION

Since it is usually the first step towards analyzing the behavior of customers in these companies, customer segmentation is the best solution for retail and e-commerce companies. Early works use common variables, such as customer demographics and lifestyle but common variables are difficult to collect and some collected variables are invalid without revision. With the rapid increase in customer behavior data collected, researchers will now focus on clustering customers from the transaction data [1].

Clustering of customer transaction data is an essential phase for identifying customer activity in retail and ecommerce Trading Companies [1] [2]. Transaction data is a daily transaction of a customer where a transaction record contains a set of products (items) purchased by the customer in one basket. The main purpose of this paper is to find the optimal number of clusters and these cluster result used for finding patterns by using the Apriori algorithm. In clustering, the system use large amount of raw and unorganized data as an input and determine similarities in input data.

Basically, transaction data is information about customers ' daily transactions [1] [5]. It contains information about what type of product or set of products is purchased by buyers. There are three common problems with data clustering. One of them is how to show data about customers and customer transactions. Second, how to calculate the distance between different customers, and third is how to divide a customer into a number of customer groups [1].

The system applies Apriori algorithm for finding pattern. Apriori algorithm is useful in developing frequent sets of elements and corresponding Association rules. It defines sets of elements that are subsets of at least transactions in the database. The input of Apriori algorithm is the output of Customer Transaction Clustering Algorithm. In a system the customer transaction data is presented by using transaction tree and the distance between them is also calculated. Cluster the customer transaction data by using customer transaction clustering algorithm. The system selects frequent customer as representatives of customer groups. Finally, the system forwards the output of clustering to Apriori algorithm for finding patterns.

Use the transaction tree distance to compare customers at all levels of system items (product) tree. However, the customer's transaction information is very large, even after the data is compressed by the transaction tree. So the customer transaction information clustering speed is very important. So the system proposes a Customer Transaction Clustering with apriori algorithm for finding patterns.

II. REVIEW CRITERIA

X. Chen, Y. Fang, M. Yang, F. Nie, Z. Zhao and J.Z. Huang [1], the author PurTreeClust presented for large customer transaction data. The purchase tree is constructed for each customer from the customer transaction data. A new distance metric is defined to effectively calculate the distance from two purchase trees. They cluster the purchase trees and rank the purchase trees as candidate representative. This paper also proposes gap statistic based method for evaluation of clusters.

Yiling Yang Xudong Guan Jinyuan You [2], this paper study the problem of category data clustering, especially transaction data which is characterized by high dimensionality and large capacity. They are very effective and fast, and are scalable, so they can be used to increase the height of the cluster histograms, they propose a CLOPE algorithm.

V. L. Migueis, A. S. Camanho, and J. F. e Cunha [3], this paper proposes a retail segmentation method based on the customer's lifestyle. The typical shopping basket clustering technique used in the past is transaction records. They deduce a lifestyle corresponding to each typical shopping basket.

Manuscript published on 30 September 2019
***** Correspondence Author

Sonali L. Mortale*, Department of Computer Engineering, Siddhant College of Engineering Sadumbre, Pune, India. Email: sonali.mortale@gmail.com

Prof. Manisha J. Darak, Department of Computer Engineering, Siddhant College of Engineering Sadumbre, Pune, India. Email: darakmanisha9@gmail.com

Customers are assigned to segments based on their similarity to the general basket. They identify actions to strengthen the relationship between the company and its customers.

Q. Wu, X. Chen, J. Z. Huang and M. Yang [4], a new subspace-weighted co-clustering (SWCCC) algorithm is proposed.

In this method, a set of sub-space weights is introduced into the weighted gene object on the sample cluster. The subspace weight is automatically calculated during the clustering process. Important genes can be identified from subspace weights.

M. Pawlik and N. Augsten [5], this paper, propose a robust tree Editing distance algorithm called RTED. The asymptotic complexity of RTED is less than or equal to the highest contention complexity for any input instance. We present a class of LRH (Left-Right-Heavy) algorithms, including the RTED and fastest tree-Editing distance algorithms, which are presented in the literature. There is a suggestion LRH's algorithm for run-time complexity.

Kishana R. Kashwan and C.M.Velu [6], the research paper, developed a real-time and online system for a specific super market to predict sales in various seasonal cycles; the model received input from the sales data records and automatically updates the segment at the end of the day business statistics.

Htun Zaw Oo, Nang Saing Moon Kham [7], the author discuss implementation of the system for pattern discovery using association rules as a method for web-based mining. Analysis of such clusters leads to the discovery of strong related rules. They acquired all the important Association rules between items in a large database of transactions. A relationship between different page requests was found. The extracted rule support and trust values are taken into account in order to gain the attention of web visitors. Therefore, the number of hits can be increased by analyzing the attitude of visitors. The approach discussed in this paper, helps the web designers to improve their website usability.

Shengrui Wang, Ernest Monga, Andre Mayers and Tengke Xiong [8], the categorical data for hierarchical clustering algorithms is proposed, which leads to the formation of DHCC. In this paper, consider the task of clustering categorical data from the viewpoint of optimization, and propose an effective procedure for initializing and improving the partitioning of clusters. The initialization of the split is based on a multiple response analysis (MCA).It also devises strategies to determine when to end the split process.

X. Chen, X. Xu, Y. Ye, and J. Z. Huang [9], in this paper, propose (TW-K-mean), automatic two-step variable weight clustering algorithm for multi-view data, which can simultaneously calculate the weight of a view and its individual variables. In this algorithm, the view weights are assigned to each view to identify the compactness of the view, and the variable weights are assigned to each variable in the view to support a class that determines the distance function with the quantity variable weight.

Kavita M. Gawande .Mr. Subhash K. Shinde , Mrs. Dipti Patil [10], here, the purchasing pattern of the food product of the customer was offered using the data mining technology. KMedoids is clustering algorithms used for food items. The output of the clustering is used as the input of the association rule mining apriori algorithm for frequent pattern matching.

III. METHODOLOGY

A detailed description of the proposed system is as follows:

Transaction Dataset: The system uses customer transaction data. The data is collected from an offline mobile sales store and consumers through field survey campaign. These transactions include products purchased by customers.

Preprocessing of data: Perform the preprocessing on dataset. Preprocessing the transaction data of customers.

Product (Item) Tree Generation: The item tree contains a number of nodes. In which product or item is represented by child node. An internal node represents category of particular item.

Transaction Tree Generation: The transaction tree consists of several nodes. The child nodes represent the items that the customer purchased, and the internal node represents the category of the specific item.

Transaction Tree Distance: Customers do not purchase similar products; because of this, the distance between any two transaction trees will have a high distance value. Within the tree editing distance, it is very difficult to restore the cluster structure. The transaction tree distance metric is used to solve this problem. The distance to the transaction tree compares customers to all levels of the product tree. The distance can be calculated by following formula:

$$d(\varphi_i, \varphi_j, \gamma) = \sum_{l=1}^{H(\Phi)} w_l d^l(\varphi_i, \varphi_j) \qquad (1)$$

Transaction Tree Clustering: The cluster the transaction tree, Transaction Tree Clustering Algorithm is used.

Mining Purchase patterns by association rule mining: To mine the customer purchased items the system uses apriori algorithm for finding patterns.

Recommendation of products: Finally, system recommends the product and gives fast and accurate results.

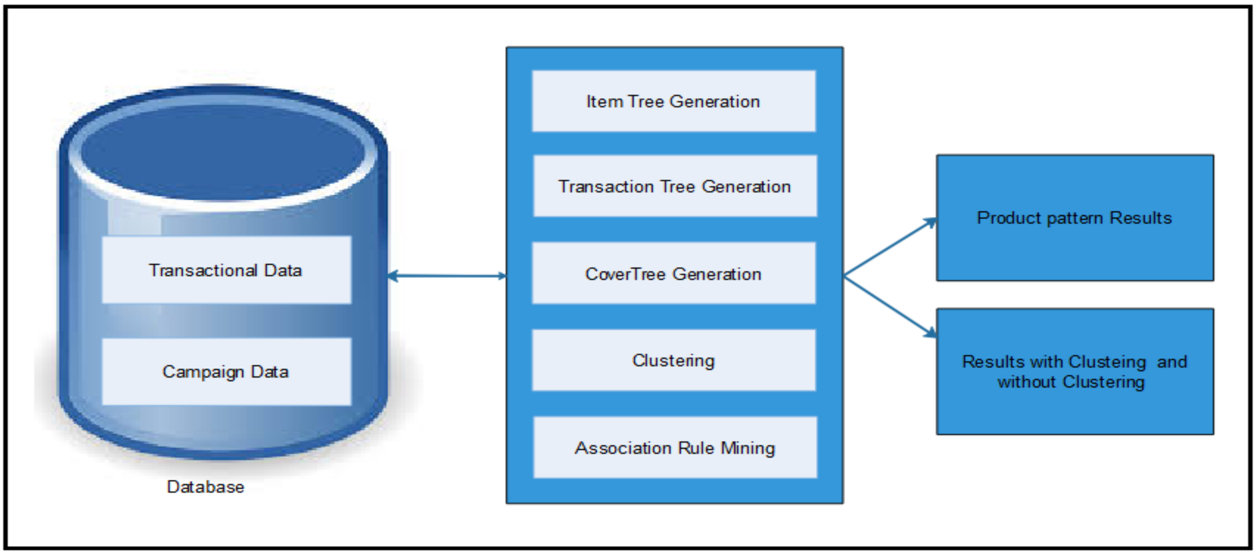


Fig. 1. System Architecture

A. Algorithm

- Algorithm 1: Cust_Tran_Clustering

Clustering of customer transaction data consists of following steps.

1. Generate the Item (product) tree.
2. Generate customer transaction tree for each customer.
3. Calculate the distance between two transaction trees.
4. Estimate the level density of transaction Tree with cover tree: $den^{l}_{CT}(p)$
5. Calculate the separate distance of object p ∈ CS_l: $sdis^{l}_{CT}(p)$.
6. Calculate separate density of object p ∈CS_l: $sden^{l}_{CT}(p)$ $sden^{l}_{CT}(p) = den^{l}_{CT}(p) * sdis^{l}_{CT}(p)$
7. Select 'k' representative trees as 'k' tree having highest separate densities.
8. Perform clustering by assigning each customer to the nearest representative.

- Algorithm 2: Apriori Algorithm

Following are the steps of Apriori Algorithm:

1. Initialize s=1
2. Generate frequent itemset of size '1'.
3. Generate candidate itemset of size 's+1' from frequent itemset of size 's'.
4. Prune candidate itemsets containing subsets of size's' that are infrequent.
5. Count the support of each candidate itemset
6. Remove candidate that are infrequent, keeping only those that are frequent.
7. Repeat steps 3 to 6 until no new frequent itemsets are identified

IV. RESULT AND DISCUSSION

A. Experimental Setup

The system is built by using JAVA and Netbeans framework on windows platform. The system use customer transaction dataset (a) We have collected customer transaction data from offline mobile sales store. (b)We collected data from consumers through field survey campaign.

B. Dataset

- **Transaction Data**

At the end seeks to understand the behavior and preferences of consumers when using mobile OS on smartphones. The first step involves developing a database. At the same time, we observed the trends in the mobile OS industry and built a consumer survey campaign and, accordingly, tables in the relational database. The database is built on MySQL. To do this, we collected customer transaction data from off-line mobile shoppers. We made up a list of products from 93 products for our dataset and list of 100 customers and combine 560 transactions.

- **Survey Campaign**

To obtain more relevant data, a field survey method is used to collect information from consumers with mobile smartphones. The survey involved 100 consumers, where the number of male respondents exceeds the number of female respondents. The share of women is 44%, while the share of men is 56%. The majority of respondents, 47%, are between 23 and 32 years old.

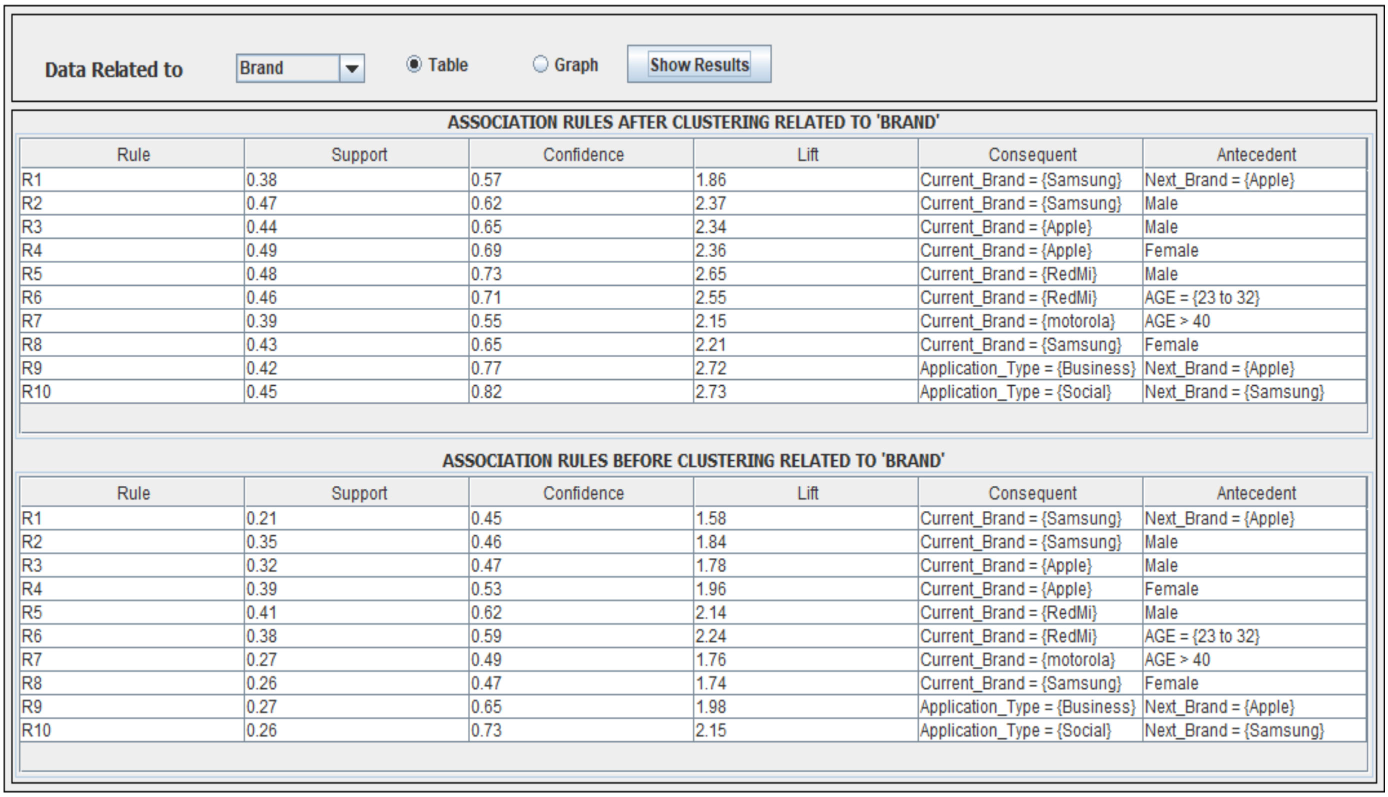
C. Results

1) Confidence and Lift of data related to Brand:

Three rules Three rules are extracted from demographic information, called R1, R2, and R3, etc. the results of the analysis of Table 1 show that Samsung, Apple, Redmi are more attractive among male consumers.

In addition, Redmi seems more desirable for young consumers aged 23-32 years.

Table I- Association Rule Result Related to Brand



Data Related to Brand ● Table ○ Graph Show Results

ASSOCIATION RULES AFTER CLUSTERING RELATED TO 'BRAND'

Rule	Support	Confidence	Lift	Consequent	Antecedent
R1	0.38	0.57	1.86	Current_Brand = {Samsung}	Next_Brand = {Apple}
R2	0.47	0.62	2.37	Current_Brand = {Samsung}	Male
R3	0.44	0.65	2.34	Current_Brand = {Apple}	Male
R4	0.49	0.69	2.36	Current_Brand = {Apple}	Female
R5	0.48	0.73	2.65	Current_Brand = {RedMi}	Male
R6	0.46	0.71	2.55	Current_Brand = {RedMi}	AGE = {23 to 32}
R7	0.39	0.55	2.15	Current_Brand = {motorola}	AGE > 40
R8	0.43	0.65	2.21	Current_Brand = {Samsung}	Female
R9	0.42	0.77	2.72	Application_Type = {Business}	Next_Brand = {Apple}
R10	0.45	0.82	2.73	Application_Type = {Social}	Next_Brand = {Samsung}

ASSOCIATION RULES BEFORE CLUSTERING RELATED TO 'BRAND'

Rule	Support	Confidence	Lift	Consequent	Antecedent
R1	0.21	0.45	1.58	Current_Brand = {Samsung}	Next_Brand = {Apple}
R2	0.35	0.46	1.84	Current_Brand = {Samsung}	Male
R3	0.32	0.47	1.78	Current_Brand = {Apple}	Male
R4	0.39	0.53	1.96	Current_Brand = {Apple}	Female
R5	0.41	0.62	2.14	Current_Brand = {RedMi}	Male
R6	0.38	0.59	2.24	Current_Brand = {RedMi}	AGE = {23 to 32}
R7	0.27	0.49	1.76	Current_Brand = {motorola}	AGE > 40
R8	0.26	0.47	1.74	Current_Brand = {Samsung}	Female
R9	0.27	0.65	1.98	Application_Type = {Business}	Next_Brand = {Apple}
R10	0.26	0.73	2.15	Application_Type = {Social}	Next_Brand = {Samsung}

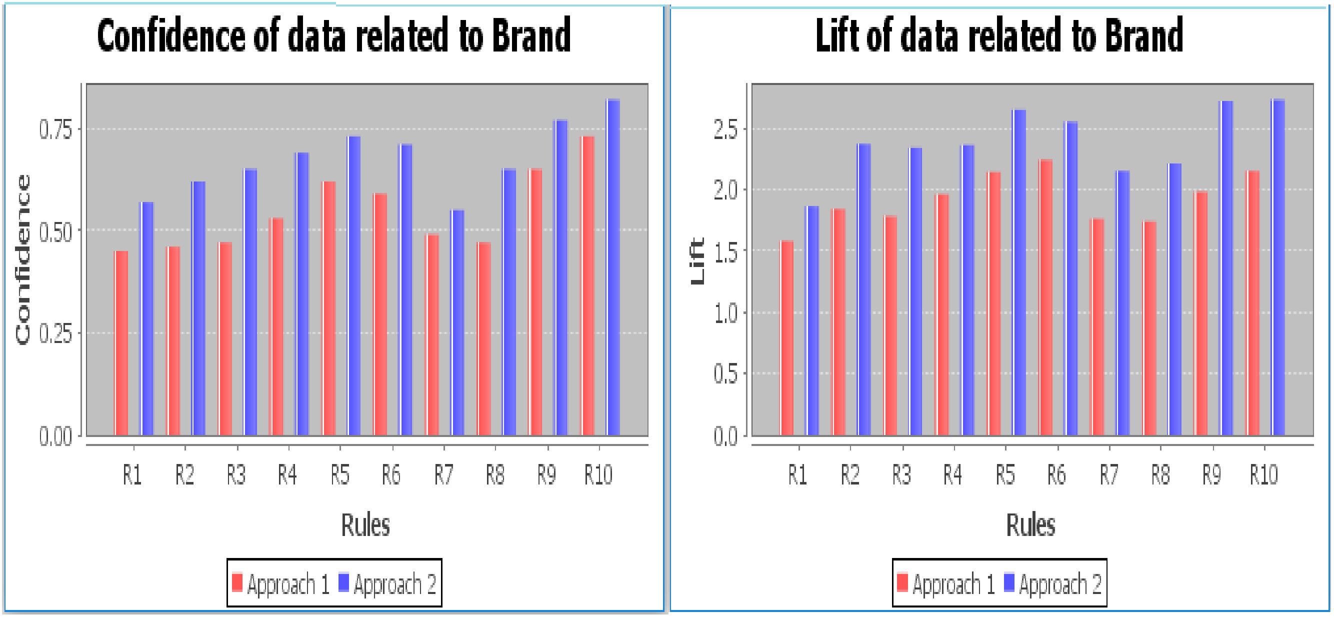


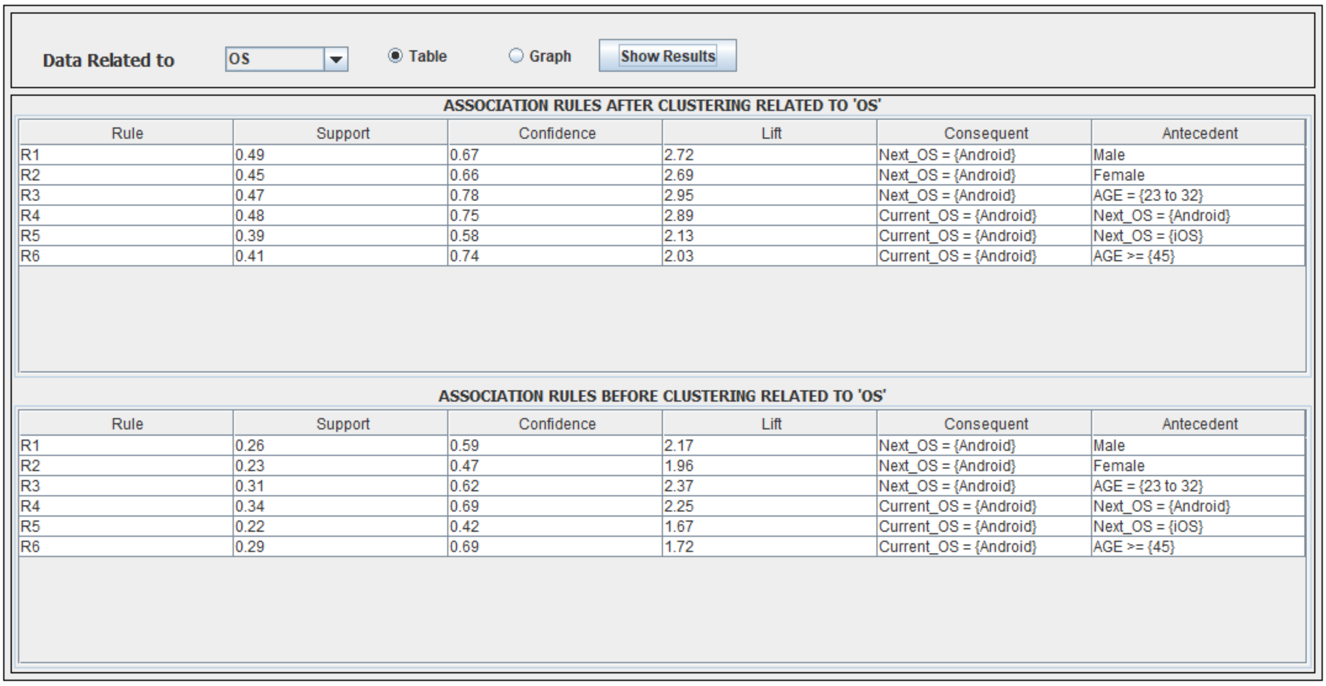
Fig.2. Comparison between the Confidence and Lift of data related to Brand

2) Confidence and Lift of data related to OS:

The data analysis in Table 2 shows that the two leaders of the mobile OS market are Android and their desirability is very close to each other.

Android users seem to want to continue using Android as their next mobile OS. Therefore, brand loyalty proves to be an important determining factor in the impact of consumer choices, and this is a small customer data set. Communication is an important social theme to attract first-time smartphone users around.

Table II- Association Rule Result Related to OS



Data Related to OS ● Table ○ Graph Show Results

ASSOCIATION RULES AFTER CLUSTERING RELATED TO 'OS'

Rule	Support	Confidence	Lift	Consequent	Antecedent
R1	0.49	0.67	2.72	Next_OS = {Android}	Male
R2	0.45	0.66	2.69	Next_OS = {Android}	Female
R3	0.47	0.78	2.95	Next_OS = {Android}	AGE = {23 to 32}
R4	0.48	0.75	2.89	Current_OS = {Android}	Next_OS = {Android}
R5	0.39	0.58	2.13	Current_OS = {Android}	Next_OS = {iOS}
R6	0.41	0.74	2.03	Current_OS = {Android}	AGE >= {45}

ASSOCIATION RULES BEFORE CLUSTERING RELATED TO 'OS'

Rule	Support	Confidence	Lift	Consequent	Antecedent
R1	0.26	0.59	2.17	Next_OS = {Android}	Male
R2	0.23	0.47	1.96	Next_OS = {Android}	Female
R3	0.31	0.62	2.37	Next_OS = {Android}	AGE = {23 to 32}
R4	0.34	0.69	2.25	Current_OS = {Android}	Next_OS = {Android}
R5	0.22	0.42	1.67	Current_OS = {Android}	Next_OS = {iOS}
R6	0.29	0.69	1.72	Current_OS = {Android}	AGE >= {45}

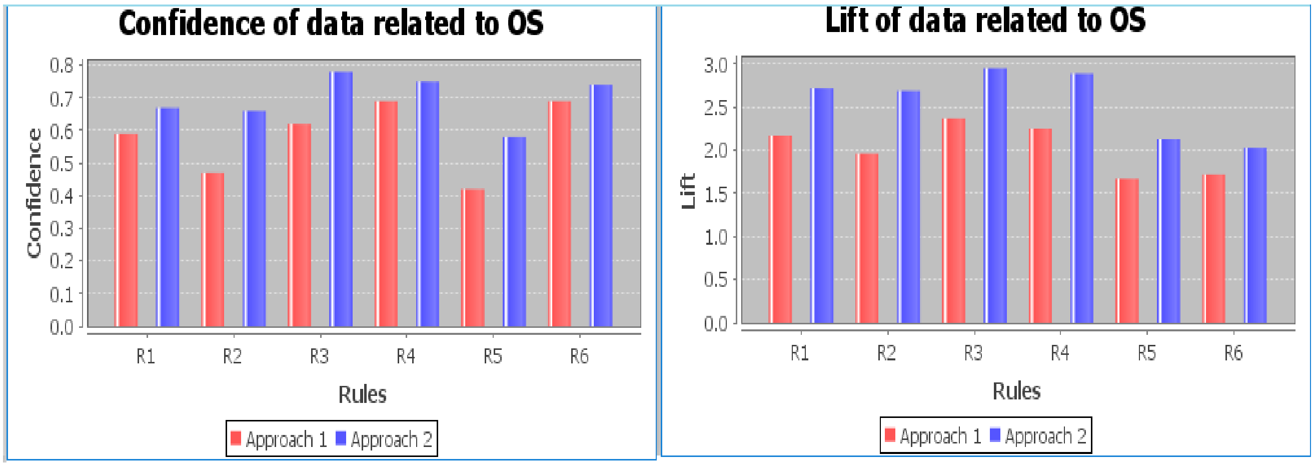


Fig.3. Comparison between the Confidence and Lift of data related to OS

Result graph shows that the proposed system is more accurate than the existing system.

V. CONCLUSION

The system is presenting an Apriori algorithm for finding pattern matching. For this “item tree” is constructed for each customer from the customer transaction data. The input of Apriori algorithm is the output of Customer Transaction Clustering Algorithm. This is beneficial for increasing product sales by identifying relationships between customer purchase pattern combinations. The customer transaction clustering algorithm is used to cluster customer transaction data. The most frequent clients are selected as representatives of client groups. Clustering is performed by assigning the client to the nearest neighborhood. Finally, the clustering results are routed to the apriori algorithm to search for patterns. Finally, the graph shows that the Apriori algorithm with Customer Transaction Clustering is more accurate and efficient than Apriori Algorithm without Clustering.

REFERENCES

1. X. Chen, Y. Fang, M. Yang, F. Nie, Z. Zhao and J. Z. Huang, "PurTreeClust: A Clustering Algorithm for Customer Segmentation from Massive Customer Transaction Data," IEEE vol. 30, no. 3, March 2018.
2. Y. Yang, X. Guan, and J. You, “Clope: a fast and effective clustering algorithm for transactional data,” 2002.
3. V. L. Migueis, A. S. Camanho, and J. F. e Cunha, “Customer data mining for lifestyle segmentation,.” 2012.
4. Q. Wu, X. Chen,J. Z. Huang and M. Yang , “Subspace Weighting Co_clustering of Gene Expression Data”, TCBB, 2017.
5. M. Pawlik and N. Augsten, “RTED: A robust algorithm for the tree edit distance”,Proc.VLDB Endowment ,Vol.5,No.4 ,2011
6. Kishana R. Kashwan and C.M.Velu , “Customer Segmentation Using Clustering and Data Mining Techniques”, IJCTE Vol. 5.No.6 December 2013

7. Htun Zaw Oo, Nang Saing Moon Kham, “Pattern Discovery Using Association Rule Mining on Clustered Data”,IJNTR,ISSN:2454-4116,Volume-4, Issue-2,February 2018,Page 07-11.
8. Shengrui Wang, Ernest Monga, Andre Mayers and Tengke Xiong, “DHCC: Divisive hierarchical clustering of categorical data”, Springer 2012.
9. X. Chen, X. Xu, Y. Ye, and J. Z. Huang, “TW-k-means: Automated Twolevel Variable Weighting Clustering Algorithm for Multi-view Data,” 2013.
10. Kavita M. Gawande,Mr. Subhash K. Shinde, Mrs. Dipti Patil , “Frequent Pattern Mining Based on Clustering and Association Rule Algorithm”, IJARCS,Vol.3,No.3,2012.

AUTHORS PROFILE

Ms. Sonali L. Mortale: PG Student, Pursuing M.E. (Computer Engineering)
Department of Computer Engineering, Siddhant College of Engineering Sadumbre, Pune, India. Email: sonali.mortale@gmail.com

Prof. Manisha J. Darak, Department of Computer Engineering, Siddhant College of Engineering Sadumbre, Pune, India.
Email: darakmanisha9@gmail.com

She is working as Assistant Professor in Computer Engineering Department, Siddhant College of Engineering Sadumbre, Pune, She is currently pursuing her PhD. Her area of interest is data mining.