Analysis of Customer Behavior using Clustering and Association Rules

P.Isakki alias Devi,
Research Scholar,
Vels University,Chennai – 117,
Tamilnadu, India.

S.P.Rajagopalan
Professor of Computer Science & Engineering,
GKM College of Engineering & Technology,
Chennai-63, Tamilnadu, India.

ABSTRACT
The analysis of customer behavior is used to maintain good relationship with customers. It maximizes the customer satisfaction. We can also improve customer loyalty and retention. The aim of this paper is to develop a very useful trend for launching products with configurations for customers of different gender based on past transactions. Based on the previous transactions of the customers, prediction is done and data is estimated with the help of clustering and association rules. This paper proposes an effective method to extract knowledge from transactions records which is very useful for increasing the sales. Customer details are segmented using k-means and then Apriori algorithm is applied to identify customer behavior. This is followed by the identification of product associations within segments. This paper aims to develop a new trend and launch a new series of products using the previous transactions of the customers.

General Terms
Clustering analysis, Association rules

Keywords
Customer Relationship Management, Data mining, Clustering, Association rules, K-means, Apriori algorithm.

1. INTRODUCTION
Customer Relationship Management is used to make more efficient business-customer relationships in order to maximize client satisfaction and thereby improve customer loyalty and retention [1]. Data mining techniques are useful to analyze the customer behavior. Data Mining grew from the persistent growth of techniques used to interrogation masses of data [9]. Prediction is done using the relevant data set taken from the database on the basis of the attributes. Customer data and customer relationships should be stored and maintained. Customer loyalty and retention are improved when customer satisfaction is more. Customer satisfaction and loyalty are attracted more attention in recent years [10].

2. CUSTOMER RELATIONSHIP MANAGEMENT
Data mining creates the ideal environment for making customer relationship management (CRM). The approaches of how data analytics can be used to make various CRM functions like customer segmentation, communication targeting, retention, and loyalty much more effective are present and implemented before. CRM is considered as the database marketing of an organization with the database of customers [8]. CRM identifies the most profitable customers and provides the highest level of service to them. Customer-centric approach helps to improve customer satisfaction and thereby maximize customer retention.

Good customer relationship management software helps you to understand customer expectations and provides services. Software must be absolutely secure and hassle-free to use [3].

3. PROPOSED WORK
Predictive analytics analyzes historical data to make good predictions. Such predictions rarely take the form of absolute statements, and are expressed as values that correspond to the behavior taking place in the future [2].

The customers with similar purchasing behavior are first grouped by means of clustering techniques. Finally, for each cluster, an association rules are used to identify the products that are frequently bought together by the customers.

3.1 Clustering Analysis
Clustering analysis is a data mining technique that maps data objects into unknown groups of objects with high similarity. Clustering is the task of segmenting a heterogeneous population into a number of more homogenous clusters [11]. Clustering algorithms are classified into partitional or hierarchical.

The k-means clustering is used to cluster observations into groups of related observations without any prior knowledge of those relationships. In this paper, we apply the *k* -means algorithm to segment customers based on transactions records. This algorithm aims to assign a set of *n* data objects to *k* clusters in order to achieve a high intracluster similarity and a low inter cluster similarity.

The algorithm clusters observations into k groups, where k is provided as an input parameter. It then assigns each observation to clusters based upon the observation's proximity to the mean of the cluster. The cluster's mean is then recomputed and the process begins again. The algorithm arbitrarily selects k points as the initial cluster centers. Finally, this algorithm aims at minimizing an objective function, in this case a squared error function. The objective function

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point and the cluster centre, is an indicator of the distance of the n data points from their respective cluster centres. The algorithm is composed of the following steps:

1. Place K points into the space represented by the objects that are being clustered. These are represented as initial group centroids.
2. Assign each object to the group which has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat the steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

We must define the initial seeds in the first iteration of the algorithm. We have to run *k* -means with different seeds and to choose the ones that have produced the lowest value of a criterion function, usually the square-error. Concerning the definition of the number of clusters, several heuristics are available, as it is not possible to theoretically determine the optimal value of the number of clusters [4].

3.2 Apriori algorithm for Association Rules

Association rules are useful for the analysis of customer data [13]. An association rule can be represented in the form *X => Y*, indicating that when product X is purchased, product Y is also purchased. X is known as the antecedent and Y is the consequent, such that X triggers the purchase of Y. The algorithms which can be used for market basket analysis are for a generic problem of association: the Apriori algorithm, Frequent-pattern growth algorithm [5] and Eclat algorithm. These algorithms involve two stages. The first stage concerns the discovery of the products which more frequently purchased. The Apriori algorithm is the most well known association rule mining algorithm. At first, we give the following definitions:

Definition 1 : Given a set of items I = { $I_1,I_2,\ldots,I_s$}, and the database of transaction records D = {$t_1,t_2,\ldots,t_n$}, where t_1 = { $I_{i1}, I_{i2},\ldots,I_{ik}$} and I_{ij} €I, an association rule is an implication of the form X=> Y where X,Y C I and X ∩ Y = Φ .

Definition 2: The support (s) for an association rule X=>Y is the percentage (%) of transactions in the database that contains X U Y. support (X =>Y) = P (X U Y), where P is the probability.

Definition 3: The confidence or strength (Φ) for an association rule (X=>Y) is the ratio of the number of transactions that contain X U Y to the number of transactions that contains X. That is confidence (X=>Y) = P (Y|X).

Associations among requirements with high support and confidence suggest standardization. The algorithm uses the following property: If an itemset satisfies the minimum support threshold, so do all its subsets. The key of Apriori algorithm is to generate the large itemsets and then to generate association rules. Association rules are adopted to discover the interesting relationship and knowledge in a large dataset [6].

The problem is decomposed into two sub problems. One is to find those itemsets whose occurrences exceed a predefined threshold in the database which are called frequent or large itemsets. The second problem is to generate association rules from those large itemsets with the constraints of minimal confidence. Suppose one of the large itemsets is Lk, Lk = {I1, I2, … , Ik}, association rules with this itemsets are generated in the following way: the first rule is {I1, I2, … , Ik-1}=> {Ik}, by checking the confidence this rule can be determined as interesting or not. Then other rule are generated by deleting the last items in the antecedent and inserting it to the consequent, further the confidences of the new rules are checked to determine the interestingness of them. Those processes are iterated until the antecedent becomes empty. The Apriori algorithm finds the frequent sets *L* in Database *D*.

First we can stop generating supersets of a candidate once we determine that it is infrequent, since no superset of an infrequent itemset can be frequent. Second, we can avoid any candidate that has an infrequent subset. These two observations can result in significant pruning of the search space.

Find frequent set L_{k-1}
Join Step.
C_k is generated by joining $L_{k-1}$with itself
Prune Step.
Any ($k-1$) -itemset that is not frequent cannot be a subset of a frequent *k* -itemset, hence should be removed.
where
(C_k: Candidate itemset of size *k*)
(L_k: frequent itemset of size *k*)

Apriori Pseudocode
Apriori (*T*, ε)

$L_1 \leftarrow \{$ *Large 1-itemsets that appear in more than ε transactions*$\}$
$k \leftarrow 2$
While $L_{k-1} \neq \varnothing$
$C_k \leftarrow$ *Generate* (L_{k-1})
for transactions $t \in T$
$C_t \leftarrow$ Subset (C_k,*t*)
For candidates $c \in C_t$
$\text{count}[c] \leftarrow \text{count}[c] + 1$
$L_k \leftarrow \{c \in C_k | \text{ count}[c] \geq \varepsilon\}$
$k \leftarrow k + 1$
Return $\bigcup_k L_k$

4. PROPOSED DESIGN

This paper develops a method to design retail promotions, informed by product associations observed in the same groups of customers. Clustering and association rules are used to identify customer behavior. We can easily predict the sales also. Table 1 shows the customer data. The details can be fetched from many databases.

Table 1. Customer Data

Item_id	Item	Quantity	Gender	Date
1	Shampoo	5	F	28/12/2011
2	Hair Conditioner	5	F	28/12/2011
3	Sugar	12	M	28/12/2011
4	Flour	12	F	28/12/2011
5	Toothpaste	2	M	29/12/2011
6	Toothbrush	4	F	30/12/2011
7	Shampoo	5	M	30/12/2011
8	Hair Conditioner	6	F	30/12/2011
9	Toothpaste	21	F	30/12/2011
10	Toothbrush	20	M	30/12/2011
11	Biscuits	15	M	31/12/2011
12	Chocolates	30	M	31/12/2011
13	Shampoo	3	F	31/12/2011
14	Hair Conditioner	5	F	31/12/2011

All the details are stored in .csv file format. Using Weka tool, the table is analysed. Customer data contains item_id, item, quantity, gender and date. Clustering and Apriori algorithm are applied using Weka tool. Figure 1 shows the count of each attribute values.

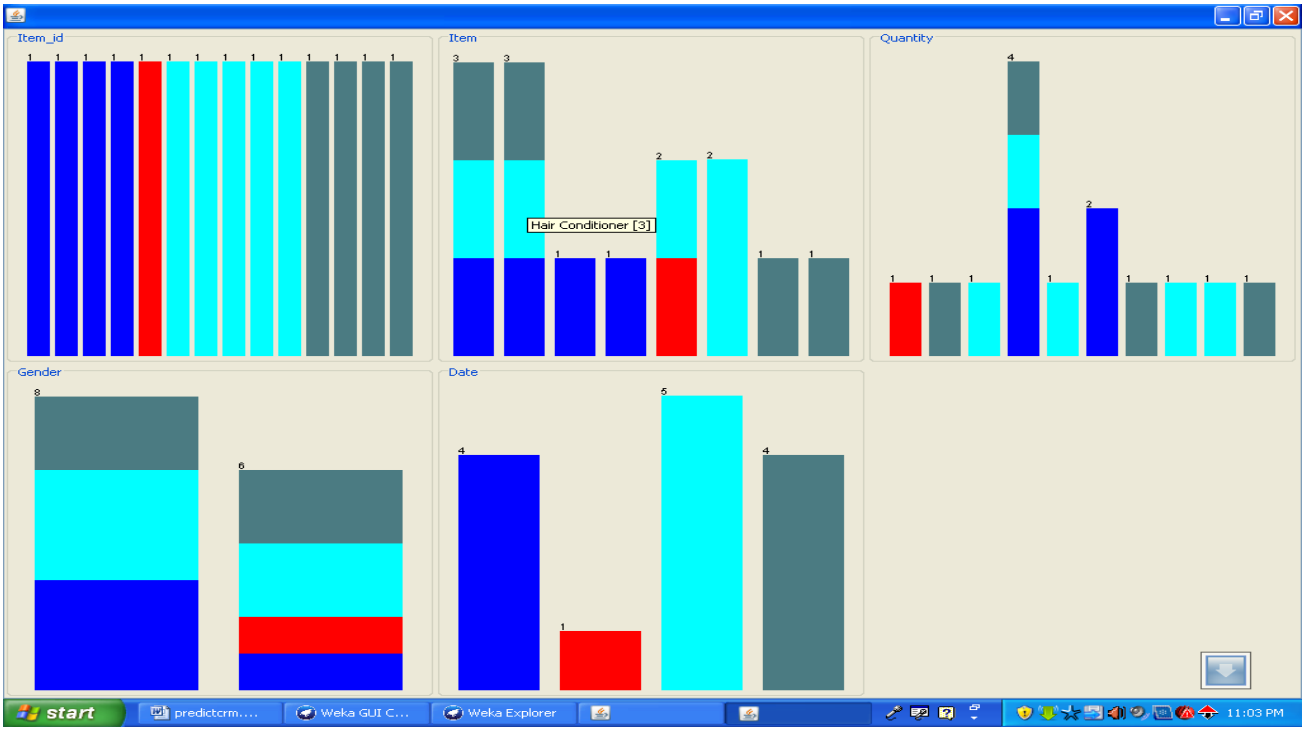

Fig 1: Count of each attribute values

4.1 K-means Algorithm

Simple k-means algorithm is then applied. The following informations are the result of Weka tool.

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10

Relation: data-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last

Instances: 14 Attributes: 5

1. Item_id
2. Item
3. Quantity
4. Date

Ignored:

5. Gender

Test mode: Classes to clusters evaluation on training data

=== Model and evaluation on training set ===

kMeans

======

- Number of iterations: 2
- Within cluster sum of squared errors: 37.0
- Missing values globally replaced with mean/mode

Table 2 contains the clustering model for gender which contains cluster centroids.

Table 2. Clustering Model for Gender

		Cluster#	
Attribute	**Full Data (14)**	**0 (9)**	**1 (5)**
Item_id	1	3	1
Item	Shampoo	Toothpaste	Shampoo
Quantity	5	12	5
Date	30/12/2011	30/12/2011	28/12/2011

Clustered Instances

- 9 (64%)
- 5 (36%)

Class attribute: Gender

Classes to Clusters:

- 0 1 <-- assigned to cluster
- 4 4 | F
- 5 1 | M
- Cluster 0 <-- M
- Cluster 1 <-- F

Incorrectly clustered instances: 5.0 35.7143 %

Instances: 14

Attributes: 5

1. Item_id
2. Item
3. Quantity
4. Gender

Ignored:

Date

Test mode: Classes to clusters evaluation on training data

=== Model and evaluation on training set ===

kMeans

======

- Number of iterations: 3
- Within cluster sum of squared errors: 25.0
- Missing values globally replaced with mean/mode

Table 3 contains clustering model for date which contains cluster centroids.

Clustered Instances

- 5 (36%)
- 5 (36%)
- 3 (21%)

- 1 (7%)

Table 3. Clustering Model for Date

		Cluster#			
Attribute	**Full Data (14)**	**0 (5)**	**1 (5)**	**2 (3)**	**3 (1)**
Item_id	1	5	1	2	3
Item	Shampoo	Tooth brush	Shampoo	Hair Con diti-oner	Sugar
Quantity	5	2	5	5	12
Gender	F	M	F	F	M

Class attribute: Date

Classes to Clusters:

- 0 1 2 3 <-- assigned to cluster
- 0 2 1 1 | 28/12/2011
- 1 0 0 0 | 29/12/2011
- 2 2 1 0 | 30/12/2011
- 2 1 1 0 | 31/12/2011
- Cluster 0 <-- 29/12/2011
- Cluster 1 <-- 30/12/2011
- Cluster 2 <-- 31/12/2011
- Cluster 3 <-- 28/12/2011

Incorrectly clustered instances: 9.0 64.2857 %

Attributes: 5

1. Item_id
2. Quantity
3. Gender
4. Date

Ignored:

Item

Test mode: Classes to clusters evaluation on training data

=== Model and evaluation on training set ===

kMeans

======

- Number of iterations: 3
- Within cluster sum of squared errors: 12.000000000000004
- Missing values globally replaced with mean/mode

Table 4 contains clustering model for item which contains cluster centroids.

Table 4. Clustering Model for Item

		Cluster#							
Attribute	**Full Data (14)**	**0 (3)**	**1 (2)**	**2 (1)**	**3 (1)**	**4 (1)**	**5 (2)**	**6 (3)**	**7 (1)**
Item_id	1	5	7	14	3	13	1	6	4
Quantity	5	2	5	5	12	3	5	4	12
Gender	F	M	M	F	M	F	F	F	F
Date	30/12/ 2011	31/12 /2011	30/12 /2011	31/12 /2011	28/12 /2011	31/12 /2011	28/12 /2011	30/12 /2011	28/12 /2011

Clustered Instances

1. 3 (21%)
2. 2 (14%)
3. 1 (7%)
4. 1 (7%)
5. 1 (7%)
6. 2 (14%)
7. 3 (21%)
8. 1 (7%)

Class attribute: Item

- Classes to Clusters:
- 0 1 2 3 4 5 6 7 <-- assigned to cluster
- 0 1 0 0 1 1 0 0 | Shampoo
- 0 0 1 0 0 1 1 0 | Hair Conditioner
- 0 0 0 1 0 0 0 0 | Sugar
- 0 0 0 0 0 0 0 1 | Flour
- 1 0 0 0 0 0 1 0 | Toothpaste
- 0 1 0 0 0 0 1 0 | Toothbrush
- 1 0 0 0 0 0 0 0 | Biscuits
- 1 0 0 0 0 0 0 0 | Chocolates
- Cluster 0 <-- Biscuits
- Cluster 1 <-- Toothbrush
- Cluster 2 <-- No class
- Cluster 3 <-- Sugar
- Cluster 4 <-- Shampoo
- Cluster 5 <-- Hair Conditioner
- Cluster 6 <-- Toothpaste
- Cluster 7 <-- Flour

Incorrectly clustered instances: 7.0 50 %

Figure 2 shows the cluster for Gender attribute. Figure 3 shows the cluster for Item attribute.

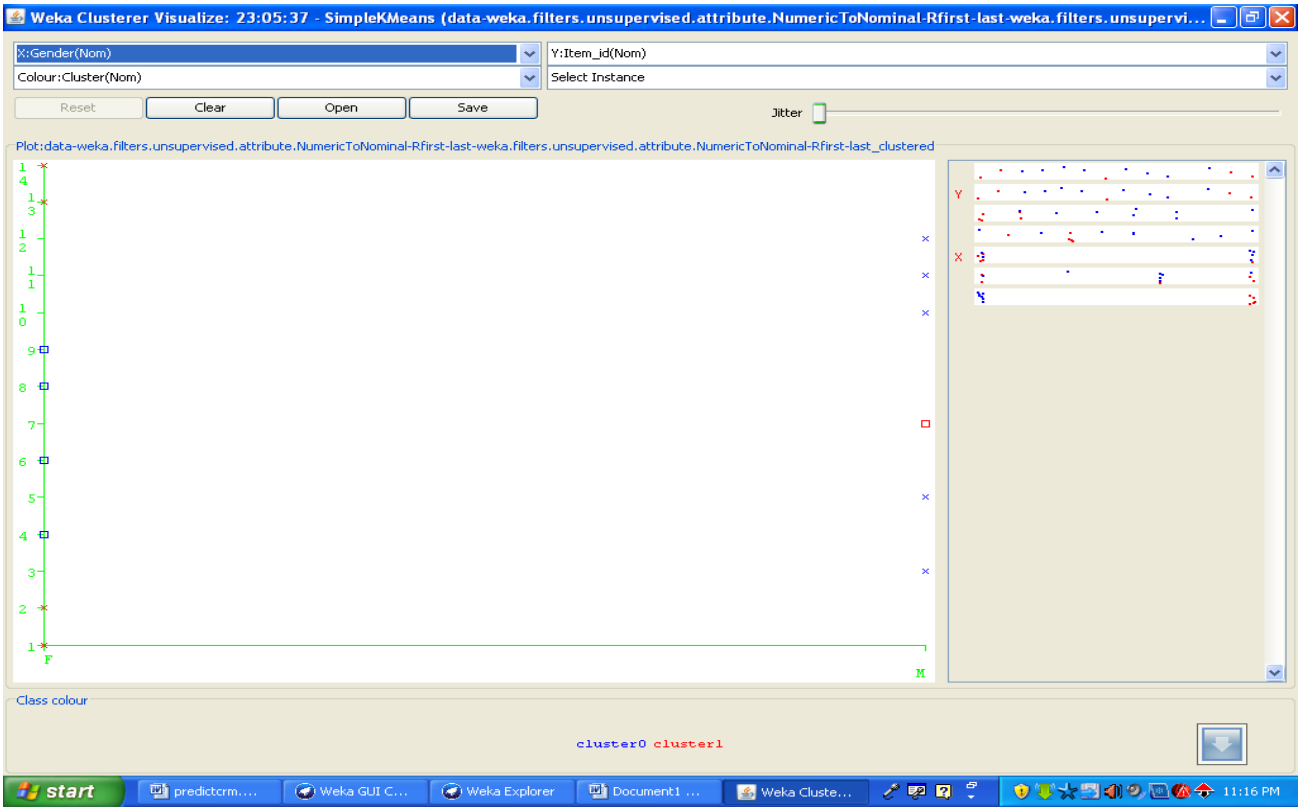

Fig 2: Cluster for Gender Attribute

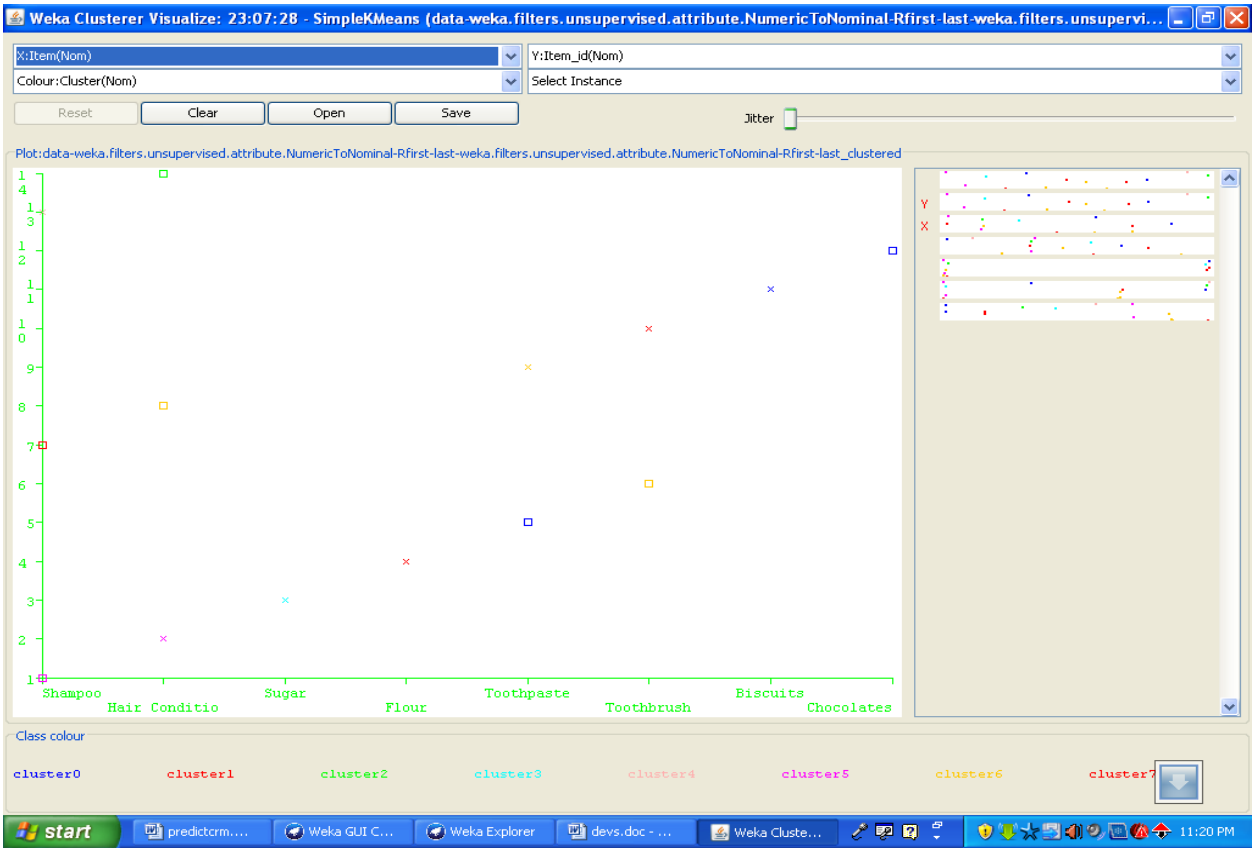

Fig 3: Cluster for Item Attribute

4.2 Apriori Algorithm

All data that are recorded in the transaction database is fed as input for the Apriori algorithm, which generates rules based on the support and confidence measures.Association rules are adopted to discover the interesting relationship and knowledge in a large dataset [14]. Apriori is designed to operate on databases containing transactions [7].

=== Run information ===

Scheme: weka.associations.Apriori -N 18 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

Relation: data-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last

Instances: 14 Attributes: 5

1. Item_id
2. Item
3. Quantity
4. Gender
5. Date

=== Associator model (full training set) ===

Apriori

=======

- Minimum support: 0.1 (1 instances)
- Minimum metric <confidence>: 0.9
- Number of cycles performed: 18
- Generated sets of large itemsets:
- Size of set of large itemsets L(1): 38
- Size of set of large itemsets L(2): 123
- Size of set of large itemsets L(3): 138
- Size of set of large itemsets L(4): 70
- Size of set of large itemsets L(5): 14

Table 5 contains best rules.

Table 5. Best Rules

Best Rules
1. Item=Hair Conditioner 3 ==> Gender=F 3 conf:(1)
2. Item=Toothbrush 2 ==> Date=30/12/2011 2 conf:(1)
3. Quantity=12 2 ==> Date=28/12/2011 2 conf:(1)
4. Item=Hair Conditioner Quantity=5 2 ==> Gender=F 2 conf:(1)
5. Quantity=5 Date=28/12/2011 2 ==> Gender=F 2 conf:(1)
6. Item_id=1 1 ==> Item=Shampoo 1 conf:(1)
7. Item_id=1 1 ==> Quantity=5 1 conf: (1)
8. Item_id=1 1 ==> Gender=F 1 conf:(1)
9. Item_id=1 1 ==> Date=28/12/2011 1 conf:(1)
10. Item_id=2 1 ==> Item=Hair Conditioner 1 conf:(1)
11. Item_id=2 1 ==> Quantity=5 1 conf:(1)
12. Item_id=2 1 ==> Gender=F 1 conf:(1)
13. Item_id=2 1 ==> Date=28/12/2011 1 conf:(1)
14. Item=Sugar 1 ==> Item_id=3 1 conf:(1)
15. Item_id=3 1 ==> Item=Sugar 1 conf:(1)
16. Item_id=3 1 ==> Quantity=12 1 conf:(1)
17. Item_id=3 1 ==> Gender=M 1 conf:(1)
18. Item_id=3 1 ==> Date=28/12/2011 1 conf:(1)

5. CONCLUSIONS

Analysis of customer behavior enables companies to improve support of their customer oriented business processes, which aims to improve the overall performance of the enterprise. This paper also focuses on getting more customer satisfaction.

Data mining methodology has a tremendous contribution for researchers to extract the hidden knowledge and information [12]. The research described in this paper also identified significant product association rules within each segment. Product association rules can be used to motivate customers to increase their purchases and keep loyal to the company. The behavior of customers can be easily identified.

Apriori algorithm is very helpful to find the best association rules. The most frequent itemsets can be easily found out from the database. In future, other data mining techniques can also be applicable when the database is larger.

6. REFERNCES

[1] Jawed Siddiqi, Babak Akhgar & Tim Wise, Sheffield Hallam, July-September 2002, A framework for the implementation of a Customer Relationship Management strategy in retail sector, Department of Applied Computer Science, Sheffield Hallam University (UK).

[2] Jaideep Srivastava, Jau-Hwang Wang, Ee-Peng Lim, and San-Yih Hwang, 2004, A Case for Analytical Customer Relationship Management" National Sun-Yat Sen University.

[3] Joseph L. Breeden,, January 1999, GA-Optimal Fitness Functions, Center for Adaptive Systems Applications, Inc. 1302 Osage, Suite A Santa Fe, NM 87505.

[4] Tibshirani, R., Walther, G., and Hastie, T, 2001, Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society. Series B, Statistical methodology, 63:411–423.

[5] Han, J., Pei, J., and Yin, Y, 2000, Mining frequent patterns without candidate generation. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of data, pages 1–12, Dallas, Texas, United States.

[6] S. Kotsiantis, Kanellopoulos, 2006, Association Rules Mining: A Recent Overview, GESTS International Transactions on Computer Science and Engineering, Vol.32(1), pp 71- 82.

[7] Rajesh V.Argiddi, S.S.Apte, 2012, Future trend prediction of stock market using Association Rule Mining of transaction data, International Journal of Computer Applications, Vol.39, pp-30-34.

[8] V.Thanuja, B.Venkateswarlu and G.S.G.N.Anjaneyulu, 2011, Applications of Data Mining in Customer Relationship Management, Journal of Computer and Mathematical Sciences, Vol. 2, Issue 3, pp 423-433.

[9] Dawn E. Holmes, Jeffrey Tweedale and Lakhmi C. Jain, 2012, Data Mining Techniques in Clustering, Association and Classification, Data Mining: Foundations and Intelligent Paradigms, Intelligent Systems Reference Library, Springer, Volume 23, 1-6.

[10] Erkan Bayraktar, Ekrem Tatoglu, Ali Turkyilmaz, Dursun Delen, 2011, Measuring the eficiency of customer satisfaction and loyalty for mobile phone brands with DEA, Expert Systems with Applications, Vol 39, Issue 1, pp. 99-106.

[11] E.W.T. Ngai , Li Xiu and D.C.K. Chau, 2009, Application of data mining techniques in customer relationship management: A literature review and classification, Expert Systems with Applications, Vol 36,Issue 2, Part 2 , pp 2592-2602.

[12] Seyed Mohammad Seyed Hosseini ,Anahita Maleki, Mohammad Reza Gholamian, 2010, Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty, Expert Systems with Applications, Vol 37, Issue 7, pp. 5259-5264.

[13] Shim, Beom-Soo and Suh, Yong-Moo, 2010, CRM Strategies for A Small-Sized Online Shopping Mall Based on Association Rules and Sequential Patterns, *PACIS 2010 Proceedings.*

[14] Ismail, R., Othman, Z. and Bakar, A.A,2010, Associative prediction model and clustering for product forecast data, Intelligent Systems Design and Applications(ISDA), 10th International Conference.