ISSN No. 0976-5697

Volume 3, No. 3, May-June 2012

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Frequent Pattern Mining Based On Clustering and Association Rule Algorithm

Kavita M. Gawande*
Department of Computer Engineering
K.J.Somaiya Institute of Engineering and Information Technology, Sion Mumbai, India
kavita.bathe@gmail.com

Mr. Subhash K.Shinde
Department of Computer Engineering
Lokmanya Tilak College of Engineering, Koparkhairane, Navi Mumbai (Maharashtra), India-400 709
subhaskshinde@gmail.com

Mrs .Dipti Patil
Department of Computer Engineering
Pillai Institute of Information technology, New Panvel
Navi Mumbai (Maharashtra), India-400 709
dipti_ajmire@gmail.com

Abstract: Decision making is considered as one of the most difficult tasks in restaurants as food items are perishable. Managers always want to analyze summaries of sales, to get aware of customer preferences, to figure out which items or combinations of items should be put on sale or to simply acquire various kinds of marketing information. To fulfil this need, this paper is aimed to provide customer's buying patterns of food items using data mining techniques. Analysis of sales data shows that some food items are sold frequently while some food products are sold rarely. This paper proposes a method that groups the food items as slow selling, medium-selling and fast selling items using KMedoids clustering algorithm. These clusters intern are given as input for the association rule mining based Apriori algorithm and Most Frequent Pattern Mining algorithm to generate frequent patterns. Experimental results show that the proposed method generates useful patterns which may assist manager in decision making. The algorithm is evaluated by using standard dataset and is compared with the results of other algorithms considering computational time and other parameters as quality measures.

Keywords: Data mining, K-Medoids, Most Frequent pattern mining, clustering, association

I. INTRODUCTION

Decision making is considered as one of the most difficult task in restaurants business. Demand of food vary every day due to variation in interest of consumers.A good restaurant manager understands the psychology of the consumer. Innovation and creation are the keywords. Keeping this in view, a restaurant manager must take an appropriate decision with dual benefits I) innovations and creation for consumers II) increase in the sales proportion. These areas are exclusively governed by decision which is the core of any business survival. Sales data is considered as one of the most valuable assets in every business as it plays vital role in decision making. Technical progress has made it possible to collect and store massive amount of sales data in restaurant database. Sales data typically have information such as date, items in transaction, quantity and price. The data mining process is intended to turn this sales data into information. Data mining is designed to identify relationships, patterns and trends that may be present among data. Traditional data mining methods of market basket analysis can be used to identify transactional data of restaurant. It helps to determine which food products are purchased together and how often and also to examine customer preferences of purchase [1]. But these methods are time consuming as we need to scan the entire database many times. So there is a need to provide an efficient method to generate the frequent patterns. Other data mining techniques like clustering and associations are preferable to find more Meaningful patterns for future predictions. Clustering is used to generate groups of related patterns, while association provides a way to get generalized rules of dependent variables [2].

For selection of particular data mining technique, it is important to analyze sales data by considering parameters such as price, sold quantity, etc. The analysis of sales data revels that all the food items do not have same frequency of selling. Some items are sold rarely, some items are sold frequently or some food items are not sold at all [2]. As food items are perishable, major problem faced by restaurant manager is selling of the food items which are not sold. To deal with this issue, this paper is aims to group the food items depending upon their selling frequency and price. Further application of association rule mining on these clusters gives efficient patterns of single item or multiple items. We have used a method that combines clustering based KMedoids algorithm with Association rule mining based Apriori algorithm and Most Frequent Pattern mining algorithm to generate frequent patterns. These frequent patterns will assist restaurant manager to formulate marketing strategies and maximize profit.

The remainder of this paper is organized as follows. Section II deals with motivation behind this paper followed by Section III which describes the existing system and its drawbacks. Section IV summarizes the clustering based KMedoids algorithm and association Rule mining based frequent pattern mining algorithm and Apriori algorithm.

Section V illustrates experimental setup of the proposed system. Section VI gives performance evaluation with respect to clustering and association. Section VII concludes the paper.

II. MOTIVATION

Typical business decisions by the management of a retail store includes what to put on sale, how to design coupons and how to place merchandise on shelves in order to maximise the profit[1,4]. Analysis of past transaction data is a commonly used approach in order to improve the quality of such decisions. Data mining researchers often try to find most feasible and efficient methods for extraction of useful patterns from database. Most of the research regarding stock data mining uses the history of transactions as it likely to persist in future[2]. A typical usage scenario for searching frequent patterns is commonly called as "market basket analysis" that involves analysing the transactional data of a supermarket or retail store in order to determine which products are purchased together and how often and also examine customer purchase preferences[3,4,5]. The Apriori algorithm introduced by Agrawal et al. in 1994 is an efficient technique to generate all significant association rules among items in a database [1, 6].

It has been claimed that the discovery of association rules is well suited for the applications of market basket analysis to reveal regularities in the purchase behavior of the customers. Apriori algorithm does have some limitations. It generates many rules which are difficult to understand. Secondly database has to be scanned many times so it takes much processing time [5]. Retail market segmentation is necessary and often critical to the development of effective marketing strategies in today's competitive marketplace [7,8]. Wang Jun, OuYang Zheng-Zheng proposes a method that combines clustering and association [9, 10, 11]. Aurangzeb Khan, Khairullah Khan, Baharum B. Baharudin in their paper entitled Frequent Patterns Mining of Stock Data using Hybrid Clustering Association Algorithm, have described that the sales data can be divided in groups based on their sold quantities using KMeans Clustering algorithm and then to generate the frequent patterns based on Most Frequent pattern Mining algorithm[2].K Means algorithm has some limitations. K-Means is sensitive to outliers and slow. Sometimes this may not give accurate clusters. Improved KMeans algorithm removes problems with KMeans [12]. A. P. Reynolds, G. Richards and V. J. Rayward have applied of K Medoids to the clustering of rules [13, 14]. S.K.Shinde, U.V.Kulkarni proposed a new clustering algorithm as fast KMedoids algorithm [16, 19]. Different data mining algorithm helps to cluster sales data as well as consumers based on various criteria [21, 22, 23].

III. EXISTING SYSTEM

Sales data plays vital role in organization. All products are not sold equally. Some products are sold frequently while some products are sold rarely. By considering this factor Aurangzeb Khan, Khairullah khan, Baharum B. Baharudin[2] have proposed a method that divides sales data in three clusters as Dead stock, Slow moving and Fast moving data using KMeans clustering algorithm. These clusters intern are given as input to frequent pattern mining algorithm for identifying the patterns. We have observed certain drawbacks of existing system as

a. K Means Algorithm is sensitive to the selection of initial cluster center.
b. It is sensitive to the outliers.
c. There is no applicable evidence for the decision of the value of K and sensitive to initial value, for different initial value, there may be different clusters generated.
d. Most Frequent pattern algorithm gives frequent pattern of single item. But in business many times we have to do analysis of which particular combination is sold together. It is not possible to do such type of analysis using Most Frequent Pattern Mining algorithm.

In order to overcome these drawbacks we are proposing a method for clustering of sales data and identifying the frequent patterns.

IV. PROPOSED SYSTEM

In this proposed system (Kavita M. Gawande, Subhash K. Shinde, Dipti Patil 2012, Aurangzeb Khan, Khairullah khan, Baharum B. Baharudin 2009) clustering and association are combined together to overcome the limitations of existing system.Phase1 is based on clustering and phase2 is based on association rule mining. Fig 1. depicts the proposed Architecture.

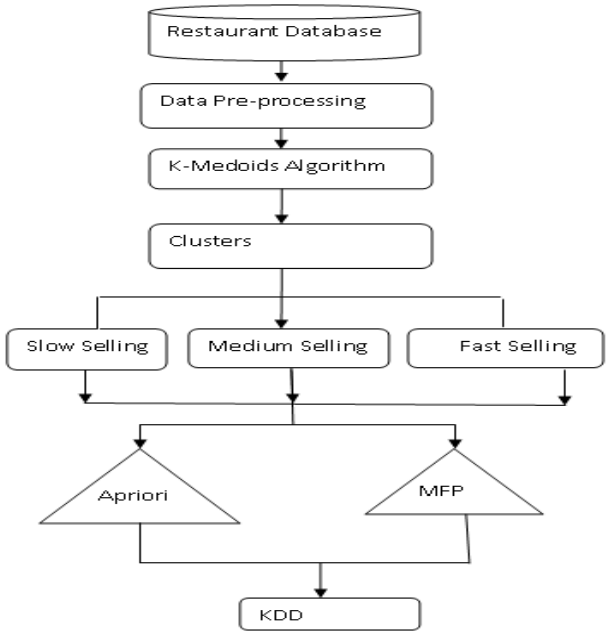

Figure.1 Proposed Architecture

A. Phase1:

Phase 1 deal with clustering. This phase use K-Medoids clustering algorithm to form the clusters of food items as slow selling data, medium selling data and fast selling data.

a. K-Medoids Algorithm:

K-Medoids is also a partitioning technique of clustering that clusters the data set of *n* objects into *k* clusters

with *k* known a priori. A medoid, finite dataset is a data point from this set, whose average dissimilarity to all the data points is minimal i.e. it is the most centrally located point in the set.

***Algorithm*:**

a. Initialize: Randomly select k of the *n* data points as the medoids.
b. Assignment step: Associate each data point to the closest medoid.
c. Update step: For each medoid m and each data point *o* associated to *m* swap *m* and *o* and compute the total cost of the configuration. Select the medoid *o* with the lowest cost of the configuration.
d. Repeat alternating steps 2 and 3 until there is no change in the assignments.

B. Phase 2:

Phase2 deal with association rule mining. Association rule mining is one of the most important and well defines technique for extract correlations, frequent patterns, associations or causal structures among sets of items in the transaction databases or other repositories. Association rules are data mining techniques that describe events that tend to occur together. It searches for interesting relationships among items in a given dataset. By examining transactions, we can find which products are commonly purchased together. This knowledge can be used in advertising or in food items placement on menu card. The two association rule mining based algorithms such as Apriori and Most Frequent pattern Mining algorithm are applied here to generate patterns.

a. Apriori:

Apriori algorithm, developed by Agrawal and Srikant 1994, is innovative way to find association rules on large scale.Given a set of transactions D, association rule mining generates all association rules that have support and confidence greater than user defined support and confidence value. Apriori algorithm considers the terms such as frequent item set, association rules, Apriori property, minimum support, confidence, precision, Recall.

a) Item set: An Item set is a set of single items from the database of transactions.
b) Frequent Item set: Items that occur together can be associated with each other such items are called frequent Item set.
c) Association rules: Conclusions based on the frequent item set form association rules.
d) Apriori Property: Any subset of frequent item set must be frequent.
e) Support: Support of an item set expresses how often the item set appears in a single transaction in the database.i.e The support of an item (or set of item)is the percentage of transactions in which that item or items occurs

If A=>B

$$\text{Support } (A=>B) = \frac{\text{Tuples\_containing\_both\_A\_and\_B}}{\text{Total\_No\_of\_tuples}}$$

f) Confidence: Confidence is the measure of certainty associated with each discovered pattern. Confidence of association rule A=>B is the ratio of the number of transaction that contain A U B to the number of transaction that contain A.

$$\text{Confidence } (A=>B) = \frac{\text{Tuples\_containing\_both\_A\_and\_B}}{\text{Tuples\_containing\_ A}}$$

g) Strong association rules: Rules that satisfy both a minimum support threshold and minimum confidence threshold are called strong association rules.
h) Precision: Precision is defined as ratio of number of relevant item selected to number of items selected.
i) Recall: Recall is Number of relevant items selected to Number of Reverent items.

Algorithm:

a) Input: Database D of transactions
b) Output: L, frequent item set in D
c) Join step: To find Lk ,a set of candidate k-item sets is generated by joining Lk-1 with itself. This set of candidate is denoted by Ck.
d) Prune step: Ck is a superset of Lk.It's members may or may not be frequent but all of the frequent k-item sets are included in Ck

Apriori Algorithm in Pseudo Code

Ck: Candidate item set of size k
k: frequent item set of size k
L_1= {frequent items};
for (k= 1; L_k !=∅; k++) do begin
C_{k+1}= candidates generated from L_k;
for each transaction t in database do
Increment the count of all candidates in C_{k+1} that are contained in t
L_{k+1} = candidates in C_{k+1} with min_support
end
return $\cup_k L_k$;

b. Most frequent pattern mining algorithm [2]:

Let set X of N food items in a Dataset have set Y of attributes. This algorithm counts maximum of each attribute values for each item in the dataset.MFP considers property matrix here to find frequent pattern.MFP considers X as food items from any cluster of slow selling, medium selling or fast selling formed using KMedoids algorithm and Y as attributes related to food items.

For ex. Food items(X) could be Veg Manchurian, Veg spring roll, etc and related attributes could be high calories or low calories, preferred by male or female, vegetarian or non vegetarian, count of items.

Algorithm:

Input: Datasets (DS)
Output: Matrix Most Frequent Pattern (MFP):
MFP (DS)
Begin
For each item Xi in Data Set
a. for each attribute
a) count occurrences for Xi

C=Count (Xi)

b) Find attribute name of C having
Maximum count
Mi=Attribute (Ci)
Next [End of inner loop]

b. Find Most Frequent Pattern
MFP=Combine(Mi)
Next [End of outer loop]

V. IMPLEMENTATION

We have conducted a set of experiments on restaurant database to examine the effectiveness of our proposed system in terms of accuracy of clusters formed and accuracy in terms of precision and recall for association.

Step 1: In database initially we assumed 110 records. These records are preprocessed in preprocessing stage to remove irrelevant or duplicate data. Data preprocessing is done manually. After preprocessing stage we got a total of 100 records in database.

Step 2: The preprocessed data in step1 is given as input to KMedoids clustering algorithms to form the clusters of slow selling data, Medium selling data and Fast selling data.

i. The Fast Selling Cluster formed with KMedoids clustering algorithm:

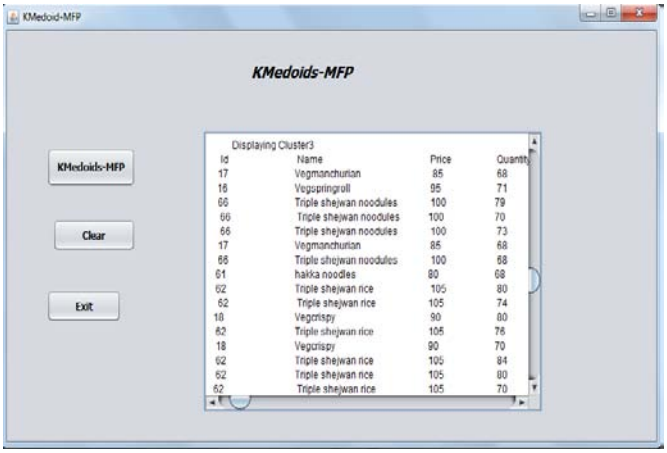

Figure.2 Fast Selling Cluster with KMedoids

ii. Medium Selling Cluster formed with K-Medoids clustering algorithm:

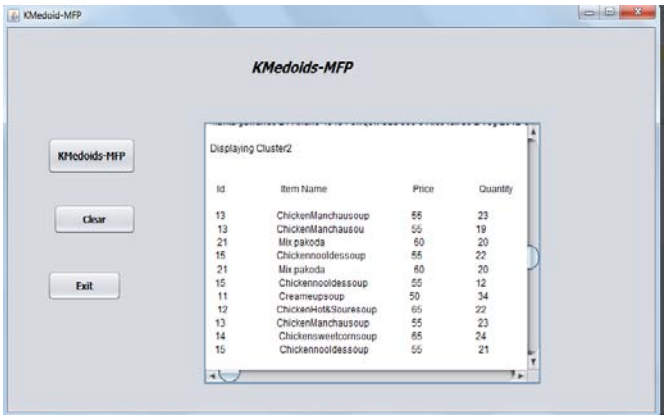

Figure.3 Medium Selling Cluster with K-Medoids

iii. Slow Selling Cluster formed with KMedoids

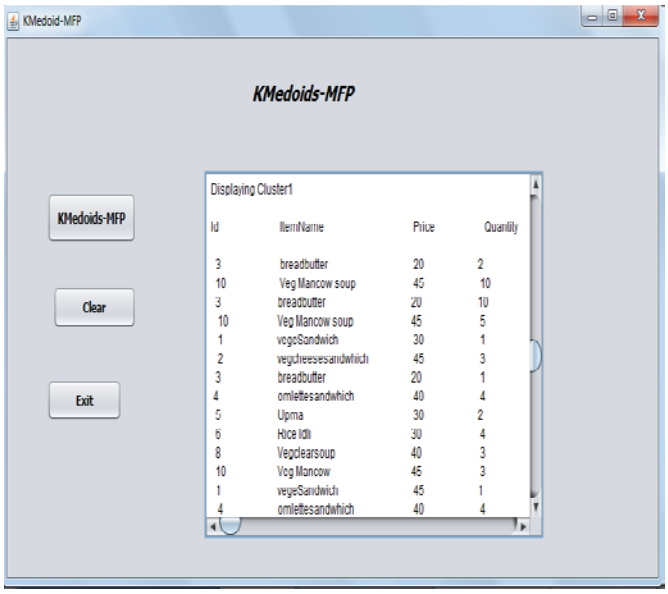

Figure 4 .Slow Selling Cluster with KMedoids

Step3: The Fast selling cluster formed in previous step is given as input to MFP and Apriori algorithm respectively.

i. Patterns generated as outcome of Most Frequent pattern mining algorithm on Fast Selling cluster**.**

Most Frequent pattern mining algorithm tries to generate frequent pattern by considering maximum count of each attribute.

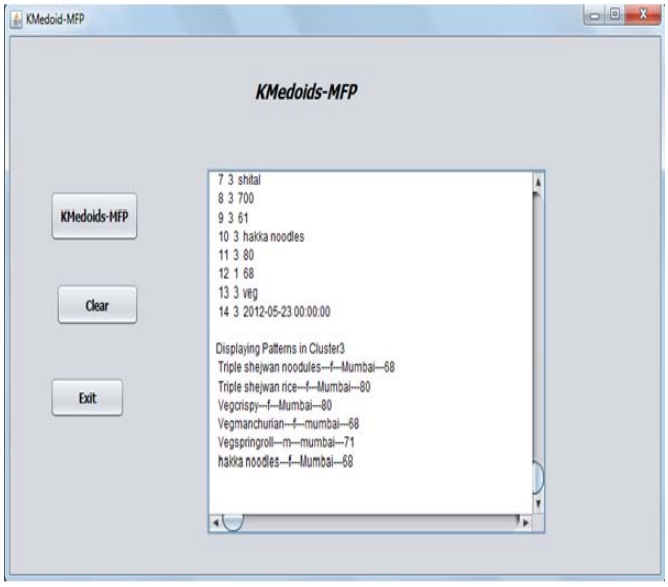

Figure 5. Frequent pattern with KMedoids and MFP

ii. Patterns generated as outcome of Apriori algorithm.

Apriori algorithm is used to generate association rules from large dataset. We have considered fast selling cluster and tried to generate association rules by considering user defined input parameter of minimum support and confidence.

Fig 6. shows frequent patterns generated as outcome of Apriori on fast selling cluster.

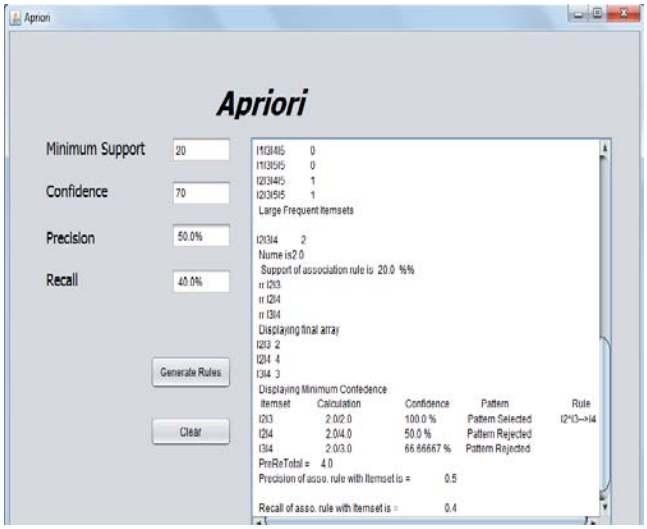

Figure 6. Frequent pattern with KMedoids and Apriori

VI. EXPERIMENTAL RESULTS

A set of experiments were conducted to check effectiveness of our system for its clustering ability with algorithm of existing system and also to check effectiveness of association rules generated. The proposed system is implemented in Net beans 6.9.1. The experiments are conducted on a 2.0 GHz, Intel Pentium-IV PC with 512 MB memory, running Microsoft Windows 7 home basic.

A. Cluster Analysis:

In order to check performance of the proposed algorithm, the algorithm was applied to real dataset, 'Iris data' whose true classes are already known. The Iris dataset includes 150 objects. There are 50 objects in each classes- 'Setosa','Versicolor' ,and 'Virginica' with four attributes such as 'Sepal length',' Sepal width', 'Petal length', 'Petal width'. The performance is measured by accuracy and computational time.

Table 1 Cluster Results Of Iris Dataset

Algorithms	Setosa	Versicolor	Virginica	Computational Time
K-Means	38	46	66	10.97 Seconds
K-Medoids	48	41	61	9.93 Seconds

The Table 1 shows that K-Medoids clustering algorithm works superior than the traditional K-Means clustering algorithms with respect to accuracy and computational time.

a. Graphical Comparison of clusters:

Following graph depicts comparison of Computational time of KMeans and KMedoids for IRIS dataset which has 150 data objects. X axis in graph denotes algorithm and y axis denotes Computational time in seconds. Computational time is measured in terms of seconds. The graph shows that KMedoids clustering algorithm takes less time.

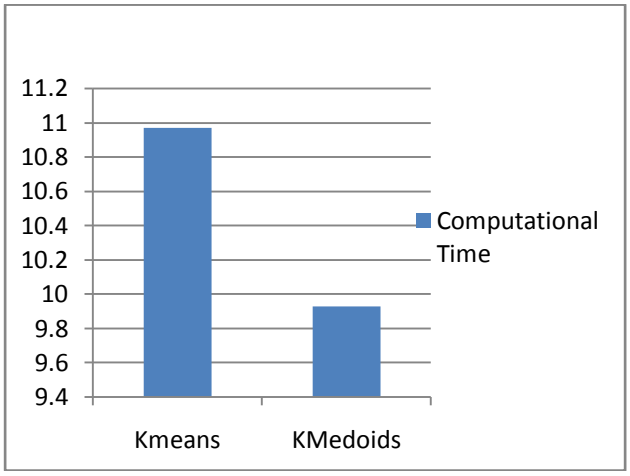

Figure 7.Comparision of KMeans and KMedoids

B. Performance Evaluation of Association:

In order to check performance of the Association algorithms, the algorithm we applied to cafeteria dataset that contains 35 objects. Comparison of different algorithms is done by considering three parameters as Computational Time, Precision and Recall.

- ***a. Precision:*** It is defined as ratio of number of relevant item selected to number of items selected.
- ***b.*** **Recall:** It is s number of relevant items selected to a number of Reverent items.

Here the precision and recall percentage is calculated for food items in cafeteria database with respect to Apriori and MFP algorithm.

Table 2: Comparison Of Algorithms

Algorithm	Computational Time	Precision	Recall
K-Means-MFP	13 seconds	50%	60%
K-Means-Apriori	15 Seconds	50%	40%
K-Medoids-MFP	11 Seconds	62%	40%
K-Medoids-Apriori	14 Seconds	50%	20%

VII. CONCLUSION

Decision making in business sector is considered as one of the critical tasks .Hybrid clustering and association mining approach is used to classify Restaurant data and find compact form of associated patterns of sale. From the experimental results it is clear that proposed approach is very efficient for mining patterns of huge data and predicting the factors affecting the sale of products. K-Medoids algorithm gives better result as compared to K-Means clustering algorithm. MFP and Apriori give frequent patterns. So we conclude that the Application of K-Medoids with Apriori and MFP will give useful patterns of multiple items or single item. These frequent patterns may assist restaurant manager to formulate marketing strategies and maximize profit.

VIII. REFERENCES

[1]. Pramod Prasad, Dr. Latesh Malik," Using association rule mining for extracting product sales patterns in retail store transactions" International Journal on computer Science and Engineering (IJCSE) ,Vol. 3 No. 5 May 2011,pp. 2177-2182.

[2]. Aurangzeb Khan, Khairullah khan, Baharum B. Baharudin, " Frequent patterns mining of stock data using hybrid clustering association algorithm" proceedings of International Conference on Information Management and Engineering (ICIME)Malaysia,Vol.1,April2009,pp.667–671.DOI 10.1109/ICIME.2009.129

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

[4]. R. Agrawal, T. Imielinski and A. Swami, "Mining Association Rules between Sets of Items in Large Databases", Proceedings of ACM SIGMOD Conference, Washington DC, USA, May 1993.

[5]. V.Umarani, Dr.M.Punithavalli,"A study on effective mining of association rules from huge database", International Journal of Computer Science and Research, vol. 1 issue 1, 2010.pp.30-34.

[6]. K. Shyamala and S. P. Rajagopalan, "Mining Essential and Interesting Rules for Efficient Prediction", Asian Journal of Information Technology (AJIT), vol. 6 issue 11 , 2009.pp. 1192-1195.doi:ajit.2007.1192.1195.

[7]. Madhav N. Segal and Ralph W. Giacobbe,"Market Segmentation and Competitive Analysis for Supermarket Retailing" International Journal of retail and distribution management ,vol. 22, issue 1, 1994,pp. 38-48.doi: 10.1108 /09590559410051395.

[8]. L. Kaufman and P. J. Rousseeuw, "Finding Groups in Data: An Introduction to Cluster Analysis," Wiley, New York, December 1990.

[9]. Wang Jun, OuYang Zheng-Zheng," The Research of K-means Clustering Algorithm Based on Association Rules " International Conference on Challenges in Environmental Science and Computer Engineering, vol. 1,March 2010,pp.285-286.

[10]. Noppol Thangsupachai, Phichayasini Kitwatthanathawon, Supachanun Wanapu, and Nittaya Kerdprasop,"Clustering large datasets with Apriori-based algorithm and concurrent processing", proceedings of the International Multiconference of Engineers and computer scienntists,vol.1,Hongkong,March 2011.

[11]. Dr.O. Nagaraju, B.Kotaiah, Dr. R.A. Khan, M.RamiReddy, N.S.Kalyan Chakravarthy," Implementing and compiling clustering using Mac Queens alias K-means apriori algorithm", International Journal of Database Management Systems (IJDMS) vol.4, no.2, April 2012.pp.69-83.doi:10.5121/ijdms.2012.4205 69.

[12]. Anwiti Jain, Anand Rajavat, Rupali Bhartiya," An efficient modified K-Means algorithm to cluster large data-set in data mining", International Journal of Advanced Research in Computer Science and Electronics Engineering Volume 1, Issue 3, May 2012.pp.92-96.

[13]. A. P. Reynolds, G. Richards and V. J. Rayward, "The application of k-medoids and PAM to the clustering of rules," Intelligent Data Engineering and Automated Learning, Lecture Notes in Computer Science, , 2004,Springer,pp. 173–178.

[14]. T.Velmurugan and T.Santhanam "A Comparative Analysis Between K-Medoids and Fuzzy C-Means Clustering Algorithms for statistically Distributed Data Points", Journal of Theoretical and Applied Information technology vol. 27 No.1, 2005-2011.pp.19-30.

[15]. Hae-Sang Park, Chi-Hyuck Jun," A simple and fast algorithm for K-medoids clustering". Expert Systems with Applications, vol. 36, issue 2, part 2, March 2009, Pages 3336-3341

[16]. S. K. Shinde and U. V. Kulkarni, "Hybrid Personalized Recommender System Using Fast K-medoids Clustering Algorithm," Journal of Advances in Information technology Vol.2, No. 3 August 2011,pp.152-158,doi: 10.4304/ jait.2.3.152-158

[17]. Jiawei Han, Micheline Kamber ," Data mining: concepts and techniques", second Edition, Morgan Kaufmann publisher, 2006, pp.401-407.

[18]. Huebner, Richard A." Diversity based interestingngness measure for association rule mining" Proceedings of ASBBS,vol.16 No.1,February 2009.

[19]. Raghuvira Pratap, K Suvarna Vani A, " An Efficient Density based Improved K- Medoids Clustering algorithm", International Journal of Advanced Computer Science and Applications(IJACSA), vol. 2, no. 6, 2011.pp.49-54.

[20]. Agrawal,R and Shrishant,R. "Fast algorithms for mining association rules",Proceedings of the 20'th International conference on very large databases,Santiago,Chile,Sept1994

[21]. Charles V. Trappey, Amy J.C. Trappey, Ai-Che Chang & Ashley Y.L. Huang," The analysis of customer service choices and promotion preferences using hierarchical clustering" Journal of the Chinese Institute of Industrial Engineers vol. 26 ,issue 5,January 2009,pp. 367-376. doi:10.1080/10170660909509151.

[22]. Robert S. Yeh, Robert D. Plante & Deepak Agrawal," Consumer Data Analysis and Its Managerial Application for the Grocery Industry" Journal of Promotion Management, vol. 17, issue 1, 2011.pp.96-113. doi: 10.1080/ 10496491. 2011.553790.

[23]. Dr. Sankar Rajagopal,"Customer data clustering using data mining technique", International Journal of Database Management systems"vol.3, no. 4, November 2011. doi: 10.5121/ijdms.2011.3401.