Recommender system based on customer segmentation (RSCS)

1. Introduction

For internet based business, the importance of appropriate recommendations is growing fast and people are increasing expecting suitable recommendations from those businesses to identify products and services (Carrer-Neto et al., 2012). That is why many companies and websites have initiated the implementation of recommendation systems in recent years to identify customer interests. The recommender system is designed to assist users to identify necessary items (Mazurowski, 2013). Recommender systems are used in different application domains, including tourism, hotels, restaurants, and parks (Yang and Hwang, 2013, Loh et al., 2004), advertisement (Cheung et al., 2003), business (Ghani and Fano, 2002), medical diagnosis (Pérez-Gallardo et al., 2013) and music selection (Bogdanov et al., 2013). Today the use of the recommender system as a strategy is considered vital in the e-commerce domain.

Recent studies in this field aim to solve the marketing strategies based on customer orientation (Huang and Huang, 2009). The system does not only focus on analyzing the customer's needs but also predicting what their future needs will be (Lee, 2010). The customer's high level of satisfaction and the recommender system's high quality, impact and improve customer loyalty (Tsai and Hung, 2012). These systems also have the ability to increase cross-selling potential (Chen and Cheng, 2008).

Recommender systems can be classified in several different ways. One viewpoint divides them in two categories: Systems which collect data on the user's past purchasing behavior and systems which work based on transaction of the current users and their behavior in the purchasing process (Cao and Li, 2007). A different classification divides recommender systems into three main categories: content-based, collaborative filtering-based, and hybrid systems (Chen and Cheng, 2008). The content-based system collect information related to the content and product specifications, then based on existing items and purchasing history of customers, the system presents recommendations to users. Collaborative filtering-based systems perform based on the experience of customer's previous purchases. Employing applied methods in the aforementioned content-based and collaborative filtering-based

systems, the hybrid systems identify user's problems arising from each of these systems and may include additional methods for enhancing the accuracy of the recommender system.

2. Literature Review

The objective of the research on the recommendation systems is to improve the accuracy of the suggestions to the customers. The customer's segmentation goal is to recognize the valuable customers which have higher impacts on the company's profitability and future purchases. If the recommendation systems can be combined with the segmentation methods, the accuracy of the recommendations to each cluster of the customers can be improved as the customers in each cluster have considerably similar attitudes in commodities selections and purchase trends. As a result, it improves the recommendations to each group of the customers which is the final objective of all the recommendation systems. In this research, a customer segmentation method (Rezaeinia et al. 2012) is combined with the recommendation systems which improves the accuracy of the recommendations. It also presents a novel method on recommendation systems.

2.1 Customer Segmentation

The aim of clustering is to classify data so that each data group has the most possible differences from other groups and the data in each group has the most possible similarities with the same data group. Bottcher et al. (2009) believe that customer segmentation is the process of dividing customers into homogeneous groups on the basis of common attributes. Also, Mizuno et al. (2008) explains that segmentation of potentially profitable customers, whom we call "good" customers, becomes significantly important. Many customer segmentation methods exist. Cuadros and Dominguez (2014) used the SOM method for customer segmentation. Casabayó et al. (2015) did the customers' segmentation using the fuzzy method. Coussement et al. (2014) segmented the customers based on the comparison of the RFM (Recency, Frequency, and Monetary) variables, Decision Tree and logistic regression methods. The RFM variables will be discussed further in later sections. Wang

(2009) believes that, “In spite of various types of segmentation variables (demographic, psycho graphic, or purchasing behavior patterns) proposed so far, practical marketers continue to use RFM (recency, frequency and monetary) model since it is easy to use and to be understood by decision makers”. Rezaeinia et al. (2012) combined the use of RFM and Analytic Hierarchy Process (AHP) and the K-Means algorithm to cluster customers of the Banking industry based on the benefit of the customers to the bank.

2.2 Recommendation Systems

The suggestion methodology is the core of a recommender system which directly affects the suggestion results (Lee, 2010). Many recommender systems exists, often addressing different needs; and we summarize several relevant ones below. Xia et al. (2006) has presented a heuristic method to solve the SVM issues in recommendation systems. Zahra et al. (2015) have presented a recommendation system using K-Means clustering which improves the accuracy of the system. Brito et al. (2015) have combined K-Medoids and CN2-SD algorithm to modify the determination process of the customers’ preferences.

As mentioned, the recommender system is generally divided into three general categories: content-based forum (CBF), collaborative filtering (CF) and hybrid systems. Among those, considering the CF characteristics, it is mostly recognized as the most successful recommender system (Tsai and Hung, 2012). These systems receive information about customers, analyze them and finally make their recommendations (Huang and Huang, 2009). There are numerous different systems using CF technique, e.g., the Tapestry which is used to filter out users' emails or the Ringo system which is employed for music recommendations (Carrer-Neto et al., 2012). The MoveiLense website is another example in the movie recommendation field.

Increasing the accuracy is one of the major research topics in recommender systems and CF technique. Recently, many CF systems are combined with other systems to enhance the quality of the results (Lee, 2010). The recommender systems can be combined with clustering methods to increase the accuracy of recommendations (Carrer-Neto et al., 2012).

2.3 The RFM (Recency, Frequency, and Monetary) Method

The segmentation of customers is typically used to identify profitable clients and also develop strategies to target them (Rezaeinia et al., 2012). According to McCarty and Hastak (2007), RFM (Recency, Frequency and Monetary) is a common method used by direct marketers. In the RFM method, the customers are segmented based on three indicators: recency, frequency and monetary (RFM) value. Also, Chan (2008) states that: "To identify customer behavior, the well-known method called recency, frequency and monetary (RFM) model is used to represent customer behavior characteristics. The first dimension is recency, which indicates the length of time since the start of a transaction. Meanwhile, the second dimension is frequency, which indicates how frequently a Customer purchases products during a particular period. Finally, monetary value measures the amount of money that customer spends during a period". The RFM method is an effective attribute for the customers' segmentation. In other words, the RFM is a model that extracts important customers from a large transaction data (Chang and Tsai, 2011). RFM is a behavior-based model. In other words, it is used to analyze the behavior of a customer which is engaged and make predictions based on his behavior (Yeh et al., 2009). According to Coussement et al. (2014), "RFM analysis is a popular approach in database marketing because of its simplicity and reasonable performance". The basic assumption of using the RFM model is that the future patterns of consumer trading resemble past and current patterns (Chan, 2008). Its advantage is in its ability to extract characteristics of customers by using fewer criteria (a three-dimensional) such as cluster attributes which causes reducing the complexity of the customer value analysis models. In this method, RFM variables of each customer are extracted which are defined as follows:

R (Recency) is the amount of recency relative to the last day of the course,
F (Frequency) is the number of each customer's transactions during the course, and
M (Monetary) is the average of the customer's deposits during the course.

Cheng and Chen (2009) also pointed out that the RFM model is one of the well-known customer value analysis methods. The calculated RFM values are summarised to clarify customer behaviour patterns.

3. Research Framework

In Fig. 1, The research framework of this study is shown together with the comparison methods. We provide an overview below of each phase, and then expand the details thereof in the subsections that follow. First, in the Data Collection and Cleanup phase, the data is acquired and the cleaned by removing transactions which have missing values. Then, in the RFM Variables Extraction phase, RFM variables are extracted according the methodology outlined in section 2.3. In the RFM Variable Weight Calculation phase, the weight of each of the RFM variables are calculated as described in section 3.4. In the Customer Segmentation phase, the customers are segmented by AHP and EM algorithm. Next, the collaborative filtering is calculated based on the nearest distinctive neighbor for each cluster of customers. The proposed method are evaluated and compared with the Conventional Method that uses collaborative filtering without segmentation. In addition, our proposed method is evaluated and compared with Multi-class SVM method, which is described in section 4. Then, the collaborative filtering (CF) is calculated for the Multi-Class SVM method and our proposed method. The final step is to evaluate the acquired results, select the optimum method, and then apply an appropriate strategy for the above-mentioned customers.

The goal of the section is to compare the presented method based on the segmentation of the customers with the two aforementioned methods, such that CF is calculated separately for each cluster based on the KNN. At the end of this section, the results of the presented method are compared against the conventional and Multi-Class SVM methods.

Fig. 1 The Research Framework of the Proposed Method versus Comparison Methods

3.1 Dataset Description

Our dataset contains the sales data of a 10-month period of a wholesale center in Tehran. This center distributes the commodities between various sellers in Tehran. This dataset contains 5 main fields which are CustomerID, SellerID, Requestdate, ComID (commodities ID), and Box fields. The Box field demonstrates how much commodities have been received by each seller from the center. The data includes 500,220 sales records for selling 179

different commodities to 12,429 customers. The noise in this research is defined as those records which have some unfilled fields. As an example, the SellerID or RequestDate or Box are not reported in some records. These records are not considered in further calculations.

3.2 Data Collection and Cleanup

First, sale transactions of the above-mentioned center over a period of 10 months are taken into account which includes 500,220 records. These transactions contain the sale data of 179 different types of items to 12,429 customers in the city. Considering the facts that some of the items have been sold in a short period of time and also some customers have bought only once and some others have only bought in a short period of time, the initial data had considerable noise and as a result, the cleanup of transactions are carried out with high precision. Consequently, the number of transactions is reduced to 274,716. These transactions are in fact related to the products and customers during the under-study period. The active clients and types of goods are then declined to 4472 clients and 117 types, respectively. Fig. 2 shows the details of customers' data before the cleanup step.

Fig. 2 Sample of Customer Transactions before Data Cleanup

3.3 The RFM Variables Extraction

Bose and Chen (2009) believe that data on customers behaviour include customers transaction records, feedbacks from customers, and web browsing records. Coussement et al. (2014) believe that "the impact of data accuracy issues on the performance of RFM analysis is evaluated". The commonly used RFM variables are often extracted from transaction records of customers. In this step, the variables *R*, *F* and *M* of each customer are calculated.

3.4 Calculating the RFM Variables' Weight

Today, the RFM is combined with other methods such as data mining in order to achieve more precise and better results. Data mining includes various techniques and methods in clustering and classification. The reported research in literature shows that some methods are more efficient than others in customers' segmentation. However, the data mining has some shortcomings, i.e., neural network gets long training times and genetic algorithm is a brute computing method (Cheng and Chen, 2009). In decision trees, too many instances lead to a large decision tree which may also decrease the rate of segmentation accuracy. In the artificial neural networks, because of the number of hidden neurons, various layers and training parameters have to be determined. As a result, it takes lengthy training times, especially in a large dataset. The aforementioned genetic algorithms have also drawbacks such as slow convergence, brute computing methods, long computation times and low stabilities. In association rules, major drawback is the number of generated rules which is huge and may be redundant (Cheng and Chen, 2009).

On the other hand, AHP is a hierarchical process to solve the complex problems involving multiple attributes by constructing the problem into goal, attribute and alternative for decision-maker (Chen, 2009). Chen and Wang (2010) believe that AHP as a qualitative and also quantitative method is a useful approach for evaluating the alternatives of complex multiple criteria methods involving subject judgement.

3.4.1 SVM (Support Vector Machine) and Multi-Class SVM

The SVM method is a classic method in customers' segmentation, which we compare here and show results in the Experimental section below. For this application, Mizuno et al. (2008) reported that the logistic regression method (LRM) is more efficient than SVM in customers' segmentation. Huang et al. (2007) presented a new method called SVC to improve the SVM method in customers' clustering. Tu and Yang (2013) claimed that the SVM method is efficient for unbalanced data, however it is not recommended as it improves the performance of the minority classes by decreasing the performance of the main and bigger classes. Ha (2010) has shown that the ANN method has higher accuracy than the SVM method in customers clustering. Lee et al. (2006) reported that the CART and MARS methods are considerably more efficient than SVM method. Xia et al. (2006) reported that the performance of the SVM method is not acceptable in recommendation systems because

of the user-item matrix sparsity. They presented a heuristic method to solve the issues of the recommendation systems.

3.5 Customer Segmentation

SVM is an efficient method for classification. However, it is also used for clustering. As the number of clusters are not clear in the customer segmentation process, the clustering method has to be employed. The SVM method is combined with other methods to improve it in the clustering process.

The studies which emphasise the benefit of customer retention indicate that a 1% improvement in the customer retention rate improves firm value by 5%. Similarly, Reichheld and Sasser showed that a 5% increase in customer retention increases a firm's profits at a range between 25% and 85%. There are different clustering algorithms (Hidalgo et al., 2008).

3.5.1 EM Algorithm

In this paper, the EM algorithm is used for customers' clustering. Bilmes (1998) stated that the EM algorithm is an efficient technique in this issue. This algorithm estimates the maximum-likelihood of the parameters in a given data set based on an underlying distribution in a case that the data is not complete. It is called 'E-step' as it is not necessary to explicitly form a probability distribution over completions, and it just needs to compute the 'expected' sufficient statistics over these completions. Likewise, the 'M-step' means that the model re-estimation is in fact 'maximization' of the expected log-likelihood of the data (Do and Batzoglu, 2008). These two methods can also be compared in the following way. The E-step is in fact constructing a local lower-bound for the subsequent distribution. However, the M-step improves the estimations of the unknowns by optimizing the bound. The following example illustrates these differences more clearly. Dampster et al. (1977) has presented detailed discussions about the EM algorithm.

This clustering is achieved based on weighted RFM variables and hence, as highlighted in Fig. 3, customers are classified into 5 categories.

Fig. 3 Clusters of Segmented Customers

The customers' segmentation is based on the method presented before by Rezaeinia et al. (2012). First, some of the sellers and managers were interviewed and their opinions about the values of the R, F, and M parameters were asked. Then, the weights of the R, F, and M parameters which are called W_R, W_F, and W_M, respectively, are calculated using the AHP method. The calculated weights were multiplied by the corresponding parameters and then, the clustering was done using EM algorithm.

Table 1 shows the characteristics of these clusters. It includes members in each cluster and the total number of RFM variables. As it is shown in the table, Cluster 1 and 2 have the lowest and highest numbers of customers, respectively.

Table 1 Characteristics of Customers' Clusters

3.6 Calculating the Value of Each Customers' Category

Knowing the value ratio of each one of the customers' cluster is crucial for all salesmen because attracting new customers is usually much more expensive than the existing customers and loyal and valued customers should be maintained as a competitive asset. To calculate the loyalty ratio of each one of the customers' cluster, first, the normalized average ratio (C_R^j), the normalized average rotation (C_F^j) and similarly the normalized average ratio of currency value (C_M^j) of each group of customers have to be calculated independently. According to Liu and Shih (2005) and Rezaeinia et al (2012), to measure the loyalty of each customer group, we must calculate the amounts of normalised averages of recency (C_R^j), frequency(C_F^j) and monetary value (C_M^j), separately.

The $C_I^j = W_R \times C_R^j + W_F \times C_F^j + W_M \times C_M^j$ formula is finally applied to calculate the ratio of customers' loyalty where C_I^j stands for the ratio of each cluster of customers. It is also called integrated rating.

Table 2 Calculation of the Value of Each Customer's Cluster

According to Table 2, the most valuable customers are in Cluster 5 which includes 17 key customers. These customers have the highest ratio of purchasing power and hence are the most loyal customers which are called the golden customers in the marketing term. They have the highest profitability for sales center. As it is evident from the table, the value of these customers are almost 10 times more than the customer group in the second category and this is very important for the sales center. Next valuable customers are in Cluster 3 which includes 1517 customers and are called silver clients. The customers have the ability to become a golden customer but it depends on the marketing and sale of the center. Customers of Cluster 1 are at a short distance from customers of Cluster 3 and are prone to be transformed into silver category. Cluster 2 has the least valuable customers.

3.7 Recommendations to Customers based on Proposed Methods

One of the methods widely prevalent in the recommender systems is collaborative filtering (CF) method which is the best method for this research considering existing data. The proposed method in this paper is the combining of CF method with the K-nearest neighbors. In other words, the K-nearest neighbors is used to identify the goods which are more favorable to customers. The recommendation systems designed based on the combination of CF and KNN are popular in business and academic field (Bobadilla et al., 2013). Similarity ranking based on the KNN is a classic approach, but it is still used in CF. Therefore, the combination of CF and KNN is a high priority research (Luo et al., 2013). As an example, Park et al. (2015) have combined CF and KNN and presented a fast algorithm. CF is a well-known method in recommendation systems which can be categorized as neighborhood-based and model-based methods (Zhu et al., 2015). One of the preferred methods in the CF recommendation systems is using KNN classifiers which are based on the similarity level or measuring the distance (Bagchi, 2015).

In this step, first, the customer-good matrix is established considering the customers purchase transactions for each cluster. This matrix shows the goods purchased by the customers. However, the customer-goods matrix in the proposed method of this research also shows the quantity of goods sold. Using this method, the quantity of customers' purchases is also determined in each cluster. This matrix is shown in Fig. 4.

Fig. 4 A View of the Customer-good Matrix of one of the Clusters

The next step uses Pearson coefficient to compute the purchasing similarity of customers. The Pearson coefficient equation is used as a well-known metric for calculating the similarity in the literature (e.g., Chen et al. 2015, Ekstrand et al. 2011, Candillier et al. 2007, and Eckhardt et al. 2012).The similarity between users *a* and *b* is obtained from the following formula:

$$Sim(a\ ,b) = \frac{\sum_{p\in P}\left(r_{a,b} - \bar{r_a}\right)\left(r_{b,p} - \bar{r_b}\right)}{\sqrt{\sum_{p\in P}\left(r_{a,p} - \bar{r_a}\right)^2}\sqrt{\sum_{p\in P}\left(r_{b,p} - \bar{r_b}\right)^2}} \qquad (1)$$

Which $\bar{r_a}$ and $\bar{r_{\mathrm{b}}}$ represent the average number of products bought by customers *a* and *b*, respectively. Likewise, $r_{\mathrm{a,p}}$ and $r_{\mathrm{b,p}}$ indicate that customers *a* and *b* have purchased the item *p*.

After calculating CF, the customers are sorted based on their resemblance to the proposed customer. Then, the nearest K neighbors to the customer are selected and the process is repeated for all customers which are shown below.

Fig. 5 The CF Calculation of one of Clusters

After determining K customers similar to the customer of interest, the number of goods bought by K customers is calculated and N goods which have maximum purchase are suggested as proposed goods.

3.8 Recommendations to Customers based on Conventional Method

Here, in order to compare the proposed method with the conventional method, collaborative filtering (CF) is applied on the existing data regardless of customer clustering and results are examined. Like the previous step, customers of one cluster are not just compared with another cluster but all customers are compared together. Then, the results are studied based on customers of each cluster in order to compare the results with the proposed method.

4. Experimental Results

To evaluate the acquired results, customers' data are divided into two parts of training and test. In this way, 75% and 25% of data are taken into account for training and test, respectively. The training set consists of goods purchased by customers in a particular period (approximately 7 months). The acquired results are studied in the test section.

With respect to our dataset (described in Section 3.1), True Positive (TP) means that an product was recommended to a customer and the customer bought it. False Positive (FP) means that an product was recommended to a customer and the customer did not buy it. False Negative (FN) means that an product was not recommended to a customer and the customer bought it. True Negative (TN) means that an product was not recommended to a customer and the customer did not buy it.

Recall and Precision are popular scales for measuring the quality in recommender and information retrieval systems whose results are obtained on F1 scale. Precision-Recall is considered as the most prevalent assessment criterion (Cao and Li, 2007). The formulas used for the Precision and Recall are defined as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \qquad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

The increase in recommendation causes decrease in Precision and increase in Recall. The formula F1 leads to a balance in Precision and Recall.

$$F1 = (2 * \text{Recall} * \text{Precision})/(\text{Recall} + \text{Precision}) \quad (4)$$

In each of the above scales, each of the customers' clusters is calculated separately and then the mean value of each cluster is calculated so that the proposed method could be assessed.

Table 3 Assessment of the Results of our Proposed Recommender System

As shown in the Table 3, the cluster with the highest value which includes 17 customers has the highest application compared to other clusters. Recommendation quality in this cluster is also the highest one. Considering the obtained results, the higher values of customers in one cluster is equivalent to the higher accuracy of the recommendations of that cluster. Conclusively, it can be said that customers' value ratio has direct relationship with accuracy of recommender system. Similarly, results of collaborative filtering method are obtained as follows.

Table 4 Comparison of the Proposed and Conventional Methods

As it can be seen in Table.4 the accuracy of conventional approach in clusters are approximately equal. However, the accuracy of the proposed scheme is higher than the conventional approach in the clusters with higher values. In this research, the SVM method is also used for customers clustering and the results are compared against the presented method. As it can be seen in the table 5, the customers are categorized in five clusters. The cluster's value for each cluster was calculated like before and the results were compared against CF Common Method.

Table 5 Comparison of the Multiclass SVM and Conventional Methods

Customers segmented into five clusters using SVM method. As it can be seen in table 5, the accuracy of the SVM method is higher than the CF common method. However, comparing the results of our proposed method with SVM, it can be concluded that our proposed method has higher accuracy.

5. Conclusion and Future Works

Today, the CF recommender system has successful uses. In the current research, a new method was presented to improve the accuracy of CF. In this method, customers' clustering was combined with CF and its nearest K neighbors to check the transactions of one of the sales centers in Tehran, Iran. The results indicated that the proposed method has higher accuracy compared to the conventional CF method. Likewise, the clusters which have higher values were received more accurate recommendations. This is very important for businesses and trade centers as more than 80% of their profits come from valued customers and hence, recommendations with higher accuracy to these valued customers lead to more profits to sales centers. Another point was that the proposed method was faster on obtaining the results than the conventional method as the recommendations were performed with respect to the customers of the same cluster, while all clients were assessed in the convectional method and as a result, the calculation speed is reduced as the number of customers increases in this method. The other advantage of this approach was the identification of valuable customers as some businesses can capitalize on their valued customers. Since the valued customers were calculated in the proposed method and the value of each customer was distinguished for sales representatives, the accomplished recommendations can be coordinated with sales' strategies to make it more targeted.

For future work, to improve the accuracy of the recommendation systems, this research can be extended using other customers' clustering methods and the clustering accuracy can be improved. Do the recommendations become more accurate by increasing the number of

clusters? The authors are planned to use new machine learning methods such as Deep learning in customers' clustering and study the recommendations accuracy. Likewise, profitable commodities can also be combined with low income products and increase their sales which are all related to the company's sales and marketing strategies.

References:

Bagchi, S. (2015), "Performance and Quality Assessment of Similarity Measures in Collaborative Filtering Using Mahout", *Procedia Computer Science*, Vol. 50, pp. 229 - 234.

Bilmes, J.A. (1998), "A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models", *International Computer Science Institute, Berkeley, California.*

Bobadilla, J., Ortega, F., Hernando, A. and Glez-de-Rivera G. (2013), "A similarity metric designed to speed up, using hardware, the recommender systems k-nearest neighbors algorithm", *Knowledge-Based Systems*, Vol. 51, pp. 27-34.

Bogdanov, D., Haro, M., Fuhrmann, F., Xambo, A., Gomez, E. and Herrera, P. (2013), "Semantic audio content-based music recommendation and visualization based on user preference examples", *Information Processing & Management,* Vol. 49**,** pp. 13-33.

Bose, I. and Chen, X. (2009), "Quantitative models for direct marketing: A review from systems perspective", *European Journal of Operational Research,* Vol. 195**,** pp. 1-16.

Bottcher, M., Spott, M., Nauck, D. and Kruse, R. (2009) 'Mining changing customer segments in dynamic markets', Expert systems with Applications, Vol. 36, No. 1, pp.155–164.

Brito, P.Q., Soares, C., Almeida, S., Monte, A. and Byvoet M. (2015), " Customer segmentation in a large database of an online customized fashion business", *Robotics and Computer-Integrated Manufacturing,* Vol. 36, pp. 93-100.

Candillier, L., Meyer, F. and Boullé M. (2007), "Comparing state-of-the-art collaborative filtering systems", *Lecture Notes in Computer Sciece*, Vol. 4571, pp. 548–562.

Cao, Y. and Li, Y. (2007), "An intelligent fuzzy-based recommendation system for consumer electronic products", *Expert Systems with Applications,* Vol. 33**,** pp. 230-240.

Carrer-Neto, W., Hernandez-Alcaraz, M. L., Valencia-Garcia, R. and Garcia-Sanchez, F. (2012), "Social knowledge-based recommender system. Application to the movies domain", *Expert Systems with Applications,* Vol. 39**,** pp. 10990-11000.

Chan, C.C.H. (2008), "Intelligent value-based customer segmentation method for campaign management: A case study of automobile retailer", *Expert Systems with Applications,* Vol. 34**,** pp. 2754-2762.

Chang, H.C. and Tsai, H.P. (2011), "Group RFM analysis as a novel framework to discover better customer consumption behavior", *Expert Systems with Applications,* Vol. 38**,** pp. 14499-14513.

Chen, M.H., Teng C.H. and Chang P.C. (2015), " Applying artificial immune systems to collaborative filtering for movie recommendation", *Advanced Engineering Informatics,* In Press

Chen, M.K. and Wang, S.C. (2010), "The critical factors of success for information service industry in developing international market: Using analytic hierarchy process (AHP) approach", *Expert Systems with Applications,* Vol. 37**,** pp. 694-704.

Chen, R.Y. (2009), "RFM-based eco-efficiency analysis using Takagi–Sugeno fuzzy and AHP approach", *Environmental Impact Assessment Review,* Vol. 29**,** pp. 157-164.

Chen, Y.L. and Cheng, L.C. (2008), "A novel collaborative filtering approach for recommending ranked items". *Expert Systems with Applications,* Vol. 34**,** pp. 2396-2405.

Cheng, C.H. and Chen, Y.S. (2009), "Classifying the segmentation of customer value via RFM model and RS theory", *Expert Systems with Applications,* Vol. 36**,** pp. 4176-4184.
Cheung, K.W., Kwok, J.T., Law, M.H. and Tsui, K.C. (2003), "Mining customer product ratings for personalized marketing", *Decision Support Systems,* Vol. 35**,** pp. 231-243.
Coussement, K., Van Den Bossche, F.A.M. and De Bock, K.W. (2014), "Data accuracy's impact on segmentation performance: Benchmarking RFM analysis, logistic regression, and decision trees", *Journal of Business Research,* Vol. 67**,** pp. 2751-2758.
Cuadros, A.J. and Domínguez V.E. (2014), " Customer segmentation model based on value generation for marketing strategies formulation", *ESTUDIOS GERENCIALES*, pp. 25-30.
Dempster, A.P., Laird, N.M. and Rubin D.B. (1977), " Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society*, Vol. 39, pp. 1-38.
Do, C.B. and Batzoglou S. (2008), " What is the expectation maximization algorithm?", *Nature biotechnology*, Vol. 26, pp. 897-899.
Eckhardt, A. (2012), " Similarity of users' (content-based) preference models for Collaborative filtering in few ratings scenario", *Expert Systems with Applications*, Vol. 39, pp. 11511–11516.
Ekstrand, M.D., Riedl, J.T. and Konstan, J.A. (2011), "Collaborative filtering recommender systems, Found.", *Trends Human–Comput*. Vol. 4, pp. 81-173.
Ghani, R. and Fano, A. (2002), "Building Recommender Systems using a Knowledge Base of Product Semantics", *In 2nd International Conference on Adaptive Hypermedia and Adaptive Web Based Systems.* Malaga.
Ha S.H. (2010), " Behavioral assessment of recoverable credit of retailer's customers", *Information Sciences,* Vol. 180, pp. 3703–3717.
Hidalgo, P., Manzur, E., Olavarrieta, S. and Farias, P. (2008), "Customer retention and price matching: The AFPs case", *Journal of Business Research,* Vol. 61**,** pp. 691-696.
Huang, C.L. and Huang, W.L. (2009), "Handling sequential pattern decay: Developing a two-stage collaborative recommender system", *Electronic Commerce Research and Applications,* Vol. 8**,** pp. 117-129.
Huang, J.J., Tzeng, G.H. and Ong C.S. (2007), " Marketing segmentation using support vector clustering", *Expert Systems with Applications,* Vol. 32, pp. 313–317.
Lee, S.L. (2010), "Commodity recommendations of retail business based on decisiontree induction", *Expert Systems with Applications,* Vol. 37**,** pp. 3685-3694.
Lee, T.S., Chiu, C.C., Chou, Y.C. and Lu C.J. (2006), " Mining the customer credit using classification and regression tree and multivariate adaptive regression splines", *Computational Statistics & Data Analysis,* Vol. 50, pp. 1113 – 1130.
Liu, D.R. and Shih, Y.Y. (2005) "Hybrid approaches to product recommendation based on customer lifetime value and purchase preferences", *The Journal of Systems and Software*¸Vol. 77, pp.181–191.
Loh, S., Lorenzi, F., Saldana, R. and Licthnow, D. (2004), "A tourism recommender system based on collaboration and text analysis", *Information Technology and Tourism,* Vol. 6.
Luo, X., Xia, Y., Zhu, Q. and Li. Y. (2013), " Boosting the K-Nearest-Neighborhood based incremental collaborative filtering", *Knowledge-Based Systems*, Vol. 53, pp. 90-99.
McCarty, J.A. and Hastak, H. (2007) "Segmentation approaches in data-mining: a comparison of RFM, CHAID, and logistic regression", *Journal of Business Research*, Vol. 60, pp.656–662.
Mazurowski, M.A. (2013), "Estimating confidence of individual rating predictions in collaborative filtering recommender systems", *Expert Systems with Applications,* 40**,** pp. 3847-3857.
Mizuno, M., Saji, A., Sumita, U. and Suzuki, H. (2008) 'Optimal threshold analysis of segmentation methods for identifying target customers', *European Journal of OperationaResearch*, Vol. 186, ,pp.358–379.

Park, Y., Park, S., Jung, W. and Lee S.G. (2015), " Reversed CF: A fast collaborative filtering algorithm using a k-nearest neighbor graph", *Expert Systems with Applications*, Vol. 42, pp. 4022–4028.

Perez-Gallardo, Y., Alor-Hernandez, G., Cortes-Robles, G. and Rodriguez-Gonzalez, A. (2013), "Collective intelligence as mechanism of medical diagnosis: The iPixel approach", *Expert Systems with Applications,* Vol. 40**,** pp. 2726-2737.

Rezaeinia, S.M., Keramati, A. and Albadvi, A. (2012), "An integrated AHP–RFM method to banking customer segmentation", *International Journal of Electronic Customer Relationship Management,* Vol. 6**,** pp. 153-168.

Tsai, C.F. and Hung, C. (2012), "Cluster ensembles in collaborative filtering recommendation", *Applied Soft Computing,* Vol. 12**,** pp. 1417-1425.

Tu, Y. and Yang, Z. (2013), " An enhanced Customer Relationship Management classification framework with Partial Focus Feature Reduction", *Expert Systems with Applications,*Vol. 40, pp. 2137–2146.

Wang, C.H. (2009) 'Outlier identification and market segmentation using kernel-based clustering techniques', Expert Systems with Applications, Vol. 36, No. 2, pp.3744–3750.

Xia, Z., Dong, Y. and Xing G (2006), " Support Vector Machines For Collaborative Filtering", ACM SE'06, pp. 169-174.

Yang, W.S. and Hwang, S.Y. (2013), "iTravel: A recommender system in mobile peer-to-peer environment", *Journal of Systems and Software,* Vol. 86**,** pp. 12-20.

Yeh, I.C., Yang, K.J. and Ting, T.M. (2009), "Knowledge discovery on RFM model using Bernoulli sequence", *Expert Systems with Applications,* Vol. 36**,** pp. 5866-5871.

Zahra, S., Ghazanfar, M.A, Khalid, A, Azam, M.A., Naeem, O. and Prugel-Bennett A. (2015), "Novel Centroid Selection Approaches for KMeans-Clustering Based Recommender Systems", *Information Sciences,* Vol. 320, pp. 156-189.

Zhu, T., Ren, Y., Zhou, W., Rong, J. and Xiong, P. (2014), "An effective privacy preserving algorithm for neighborhood-based collaborative filtering", *Future Generation Computer Systems*, Vol. 36, pp. 142-155.

Fig. 1 The Research Framework of the Proposed Method versus Comparison Methods

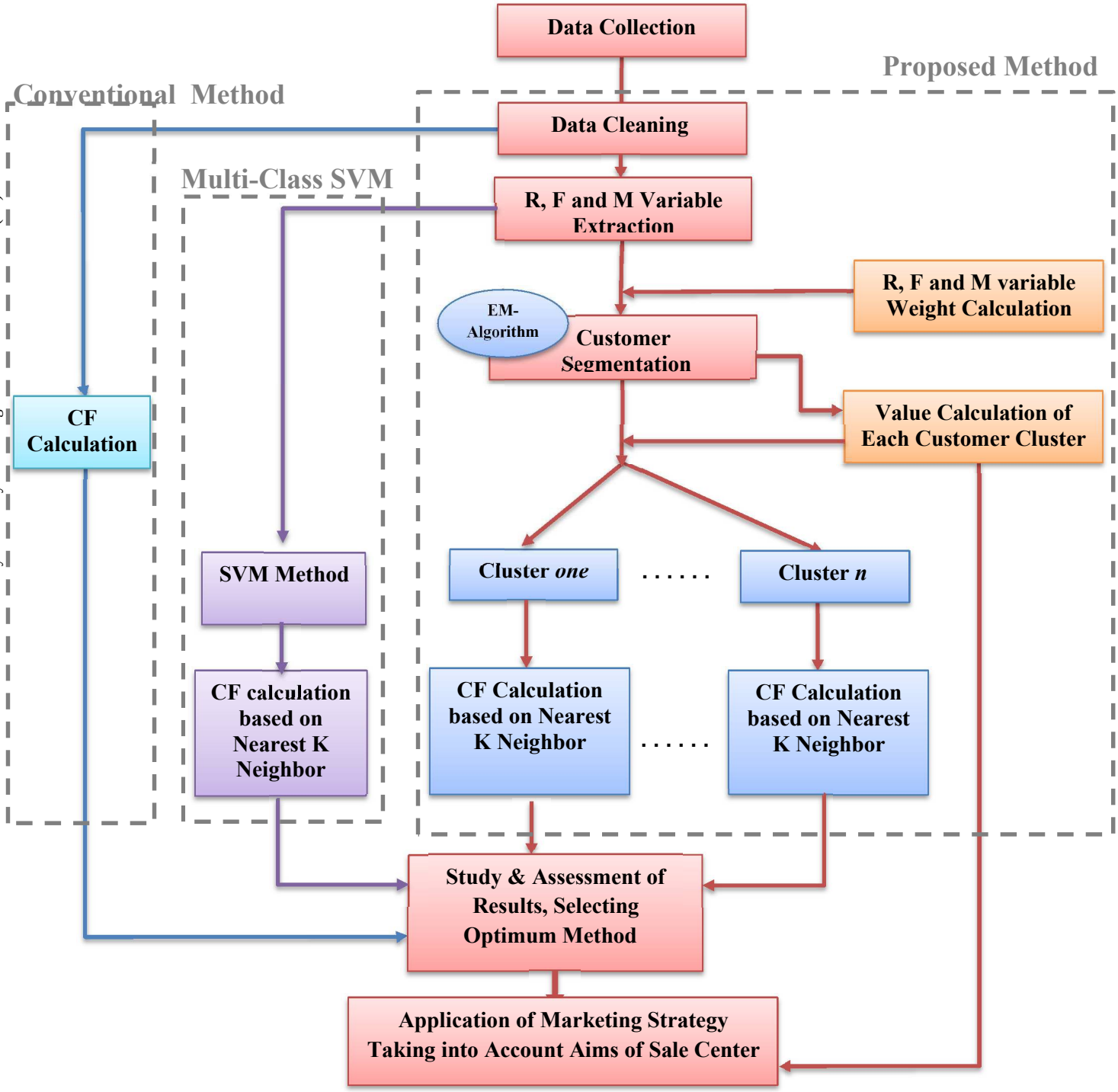

Fig. 2 Sample of Customer Transactions before Data Cleanup

	CustomerID	SellerID	RequestDate	ComID	Box
1	400284	3420	910129	PYD_1301041	0.5
2	105030	3424	910129	PYD_1501011	0.333333333
3	20698	3420	910129	PYD_1301041	0.5
4	201068	3283	910129	PYD_1301041	0.5
5	181170	3480	910129	PYD_1301041	1
6	160277	3261	910129	PYD_1501011	0.333333333
7	100144	3424	910129	PYD_1001012	1
8	601111	5051	910115	PYD_1001012	2
9	601116	5086	910114	PYD_2501101	2
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
500218	501712	6200	910729	PYD_2501001	70
500219	400473	1215	910710	PYD_1801013	3
500220	210446	3498	910710	PYD_1801013	1

Fig. 3 Clusters of Segmented Customers

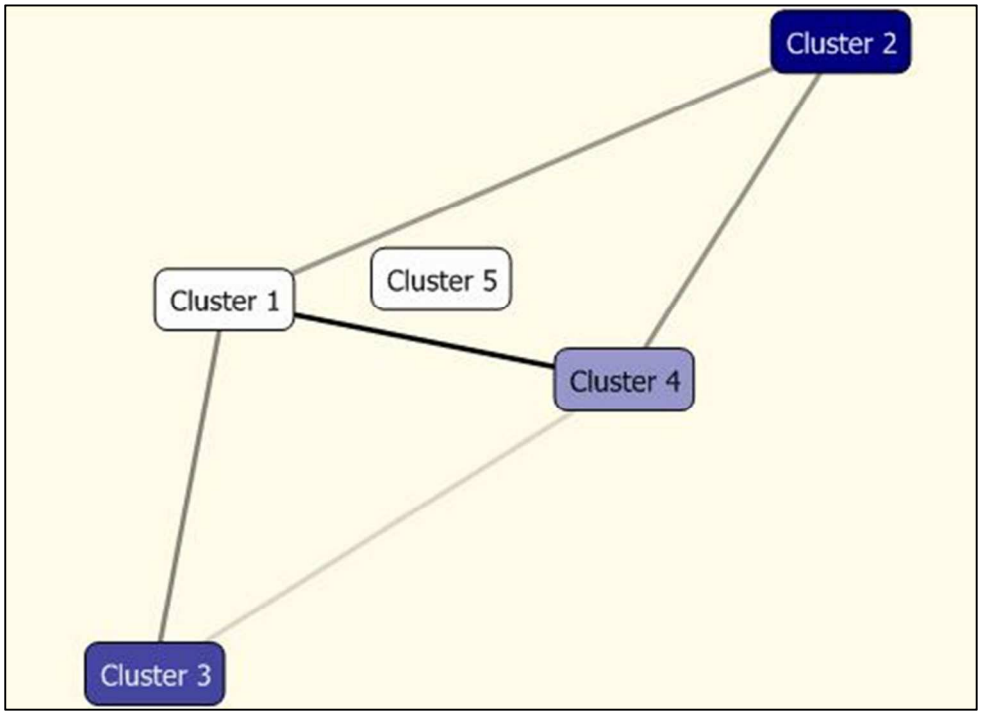

Fig. 4 A View of the Customer-good Matrix of one of the Clusters

	CustomerID	PYD_1001013	PYD_1801071	PYD_1001003	PYD_1301062	PYD_1501026	PYD_5010201	PYD_1501023	PYD_1001002	PYD_1801034	PYD_1801021	PYD_1501018	PYD_1801013	PYD_130110
1	403976	150	NULL	NULL	150	369	317	167	30	NULL	NULL	60	410	NULL
2	501213	809	100	1350	576	NULL	36	NULL	1190	100	184	246	328	80
3	501511	378	156	890	767	NULL	131	NULL	950	101	72	190	482	10
4	501512	130	NULL	800	NULL	NULL	NULL	NULL	808	150	110	5	30	NULL
5	501712	560	220	650	280	NULL	30	NULL	713	70	178	278	533	35
6	502311	144	30	155	140	NULL	193	NULL	255	50	72	90	150	55
7	502512	30	1042	60	425	40	66	15	65	345	297	592	360	130
8	503111	492	960	120	926	180	126	NULL	1180	452	300	2019	1872	NULL
9	503612	35	300	NULL	30	30	100	20	NULL	130	25	160	323	70
10	503811	158	296	170	246	NULL	65	NULL	170	NULL	55	168	225	127
11	505412	140	20	225	245	30	227	NULL	275	NULL	154	170	NULL	185
12	506117	260	1306	840	1840	NULL	56	NULL	1020	210	220	1044	1606	160
13	506612	NULL	100	NULL	150	NULL	94	NULL	NULL	25	20	130	96	40
14	507114	726	1754	970	1555	45	1214	NULL	1700	190	222	966	2338	366
15	507611	700	320	840	828	NULL	508	NULL	710	NULL	444	276	606	NULL
16	507712	195	111	355	611	NULL	NULL	NULL	205	70	66	102	450	30
17	508117	60	98	80	136	NULL	NULL	NULL	100	75	59	70	235	45
18	508312	80	232	290	130	NULL	162	NULL	220	55	30	110	443	15
19	508313	115	160	240	100	10	NULL	5	240	60	35	100	300	85
20	508613	196	50	445	330	NULL	93	NULL	445	70	NULL	90	348	142
21	508712	173	680	984	166	NULL	31	NULL	916	95	NULL	164	284	NULL
22	509001	30	120	110	76	10	NULL	10	210	30	10	60	192	82
23	600601	846	1257	534	158	957	54	583	520	189	313	1512	944	125
24	600602	236	566	100	27	84	15	57	107	27	32	121	398	23

Fig. 5 The CF Calculation of one of Clusters

Results | Messages

	CF1	CF2	CF3	CF4	CF5	CF6	CF7	CF8	CF9	CF10
1	1	0.313952177848116	0.480475457163579	0.211500205704461	0.396104443273335	0.402190110089379	0.336029099270518	0.543166924632495	0.598559834379907	0.375499296361251
2	0.313952177848116	1	0.780698984379024	0.791547100746891	0.765000163843768	0.68329635966107	0.447463513105479	0.472410356415178	0.378349848788809	0.671250894667292
3	0.480475457163579	0.780698984379024	1	0.688065784128369	0.820705937339365	0.759566684457537	0.530908467997874	0.671836378203934	0.599852317735669	0.745241151041794
4	0.211500205704461	0.791547100746891	0.688065784128369	1	0.746895738863199	0.578562465587923	0.452303833961257	0.425130028072166	0.293278681863166	0.483457901036189
5	0.396104443273335	0.765000163843768	0.820705937339365	0.746895738863199	1	0.671197662211906	0.594477618748154	0.721378571748607	0.569766845155163	0.741530638277975
6	0.402190110089379	0.68329635966107	0.759566684457537	0.578562465587923	0.671197662211906	1	0.538520062217666	0.528342309049363	0.560549015949272	0.721073157841182
7	0.336029099270518	0.447463513105479	0.530908467997874	0.452303833961257	0.594477618748154	0.538520062217666	1	0.637379965725918	0.80339027675953	0.691244935887163
8	0.543166924632495	0.472410356415178	0.671836378203934	0.425130028072166	0.721378571748607	0.528342309049363	0.637379965725918	1	0.764768200020027	0.696186324718704
9	0.598559834379907	0.378349848788809	0.599852317735669	0.293278681863166	0.569766845155163	0.560549015949272	0.80339027675953	0.764768200020027	1	0.665450849438291
10	0.375499296361251	0.671250894667292	0.745241151041794	0.483457901036189	0.741530638277975	0.721073157841182	0.691244935887163	0.696186324718704	0.665450849438291	1
11	0.19320530118398	0.649682908804961	0.551862630990087	0.576377339792293	0.522615809031843	0.654767719080848	0.390017337241402	0.317629624861717	0.297079703274355	0.476951096162612
12	0.528929173070126	0.557576128595977	0.709384267694704	0.363068857792513	0.643614527353364	0.588319665927108	0.730811356983213	0.741828348463981	0.845467142078583	0.736800667043588
13	0.240958653564186	0.243435106304158	0.469769702959305	0.104182718398887	0.363304512423433	0.502034774410402	0.329503058764352	0.449238588665772	0.426608599580052	0.553765930959067
14	0.679432289678775	0.567703666767442	0.7698662016583	0.449298142191256	0.765863944882311	0.679037685865502	0.575423615242501	0.787524763922749	0.738635012788951	0.774255436338651
15	0.389272024455121	0.704227824743331	0.678886918484124	0.684465454947963	0.767133228376515	0.677348228292673	0.46425019646428	0.503090264397053	0.357804732660974	0.595268485561589
16	0.514343238424273	0.675642669620939	0.75409651884151	0.528113751376197	0.750139959810192	0.634933879487336	0.57277419624696	0.628181542954123	0.580944539455564	0.722962340897493
17	0.339273111789719	0.659857312993321	0.723179233301406	0.549348499605082	0.719478027982714	0.64684854041999	0.685683212608751	0.623834366107782	0.67318629624008	0.751459729606637
18	0.578121134525066	0.440585315504061	0.672227902857282	0.287469715796578	0.561383202223531	0.484762442430048	0.48495298343894	0.645061404564758	0.604148934473268	0.604036950204101
19	0.401502994888668	0.568776474729341	0.748000292809997	0.450652107730604	0.721806251164784	0.517348733824854	0.488874523890703	0.72206670214669	0.592552797938693	0.732139781566235
20	0.331542388810062	0.671539542893306	0.828174201527818	0.608917913590159	0.779870059508674	0.764237498676252	0.641716503035924	0.62560337174641	0.601202790002265	0.765198948611356
21	0.21740902834317	0.643502049684818	0.691155476295136	0.522127244488145	0.653909187990542	0.532144673313444	0.409818578534107	0.492991869761113	0.3992959490645	0.663569859845139
22	0.296001583008797	0.623530689134702	0.623279303700319	0.674318327762974	0.785724639702456	0.605356647736595	0.668204723789392	0.516706734967386	0.517073046822044	0.641607634362718
23	0.362298650768721	0.157208720850089	0.244988601204253	0.0536735848205...	0.273607015306199	0.133667506598335	0.387177317896063	0.46190292789648	0.470596573726495	0.332417267347479
24	0.403613536586834	0.179195714294866	0.292982850158855	0.0642628382897...	0.32099223542872	0.123930374867406	0.362732741753489	0.393879636176402	0.49155475649173	0.418220602402644
25	-0.0501236144423	0.0887881310099	0.0414732559367	-0.0067721480486	-0.0008841974938	-0.0173043046114	0.0283878782385	0.106047065019612	-0.0642243365091	0.0637936918437

Table 1 Characteristics of Customers' Clusters

Clusters	No. Customers	Sum R	Sum F	Sum M
Cluster1	13	0.7571	0.2141	0.0275
Cluster2	2066	116.1016	14.266	0.529
Cluster3	1517	85.9587	42.063	1.4568
Cluster4	859	48.4032	12.8275	0.2874
Cluster5	17	1.1359	1.5732	2.4813

Table 2 Calculation of Value of Each Customer's Cluster

Clusters	No. Customers	C_R^j	C_F^j	C_M^j	C_I^j	Customer's Value
Cluster1	13	0.0042	0.0046	0.0013	0.0101	**3**
Cluster2	2066	0.0041	0.0019	0.0001	0.0061	**5**
Cluster3	1517	0.0041	0.0078	0.0005	0.0124	**2**
Cluster4	859	0.0041	0.0042	0.0001	0.0084	**4**
Cluster5	17	0.0049	0.0261	0.0938	0.1248	**1**

Table 3 Assessment of the Results of our Proposed Recommender System

Clusters	No. Customers	Recall	Precision	F1-Metric
Cluster1	13	0.766	0.733	0.749
Cluster2	2066	0.689	0.652	0.67
Cluster3	1517	0.769	0.731	0.749
Cluster4	859	0.697	0.67	0.683
Cluster5	17	0.798	0.756	0.777

Table 4 Comparison of the Proposed and Conventional Methods

Clusters	Cluster's Value	No. Customers	F1 (CF Common Method)	F1 (Our Method)
Cluster1	**3**	13	0.544	0.749
Cluster2	**5**	2066	0.484	0.67
Cluster3	**2**	1517	0.503	0.749
Cluster4	**4**	859	0.499	0.683
Cluster5	**1**	17	0.554	0.777

Table 5 Comparison of the Multiclass SVM and Conventional Methods

Clusters	Cluster's Value	No. Customers	F1 (CF Common Method)	F1 (SVM)
Cluster1	**4**	980	0.42	0.593
Cluster2	**1**	246	0.521	0.717
Cluster3	**2**	610	0.504	0.698
Cluster4	**5**	1567	0.402	0.55
Cluster5	**3**	1069	0.475	0.644