ISBN: 978-1-948012-02-7

Contents List available at VOLKSON PRESS

Economics & Management Innovations(EMI)

DOI : http://doi.org/10.26480/icemi.01.2017.09.11

Research on railway freight customer segmentation based on RFM model

ZHANG Bin*, PENG Qi-yuan, LIU Fan-xiao

School of Transportation & Logistics, Southwest Jiaotong University, Chengdu ,China.
*Email: zbin0470@163.com

ARTICLE DETAILS

Article History:

Received 02 october 2017
Accepted 06 october 2017
Available online 11 october 2017

Keywords:
RFM; customer segmentation; K-means; Hadoop; Spark.

ABSTRACT

Because of various aspects influence, China railway freight transportation has been subjected to greater impact, and lost plenty of sources of good and freight customers. One way to solve this problem is to improve the CRM and market management. This paper proposes RFM mode of freight customer segmentation. Finally, this paper makes the computer simulation based on Hadoop using Spark to deal with the complexity of K-means algorithm. The simulation proves the freight customer segmentation based on RFM model is efficient, and the K-means algorithm is high efficiency based on Hadoop using Spark, and some suggestions are put forward according to the results.

1. Introduction

With a sharp drop in demand of the coal, iron and steel and other bulk cargo transport, railway freight business is facing greater difficulties because of freight decline. Because the continuous development of aviation, road, sea and other modes of transport, there are many deficiencies in railway freight CRM and marketing, and railway freight faces the situation of losing customers. Railway freight transportation needs to establish an efficient system of CRM to manage railway freight customers. With the development of railway information, the railway freight has accumulated a large amount of freight data, how to carry on the depth analysis to these data, get the behavior characteristic of the customer, excavate the potential of the customer and realize the purpose of the accurate marketing, are the important problems that the railway freight transportation needs to face [1]. Some papers have discussed the role of CRM in railway freight marketing and management from different aspects [2-3]. Customer segmentation, as an important means of precision marketing, is one of the core concepts of CRM. Clustering analysis as an effective method of customer segmentation. The K-means method is widely used in many fields because of its advantages such as easy description, high time efficiency and large scale data processing [4]. ZHONG Yan et al used K-means clustering method to analyze the historical data of freight transportation, and the data mining technology is applied to the segmentation of railway freight customers [5]. DENG Cheng et al applied the improved K-means algorithm to the railway customer segmentation [6]. However, the traditional clustering methods can not meet the demand of high efficiency mining and analysis for railway freight customers in the face of massive data.

According to the actual situation of railway freight transportation, this paper puts forward the customer segmentation model based on RFM model and uses big data technology to build Hadoop clusters and Spark computing framework. Then, the analytic hierarchy process (AHP), combined with expert advice is used to set the weights of the parameters, which is expected to be more objective to get the railway freight customer value segmentation. Finally, the K-means algorithm is used in the big data processing platform built on the Hadoop Spark, and the segmentation results are obtained according to the cluster analysis of railway freight customer data.

2. Railway Freight Customer Segmentation Model and algorithm design

2.1.1 Freight customer segmentation mode

Based on the selected observation window, this paper constructs RFM model of railway freight customer segmentation from the three aspects of information as the standard of railway freight customer segmentation, which are customer recent delivery behavior, customer delivery frequency, freight revenue contribution ability.

2.1.1 RFM model

RFM model as a quantitative analysis model in the field of customer relationship management was proposed by Hughes in 1994. Hughes believed that there were three main elements constitute the best indicators of customer segmentation, namely R (Recently), M (Monetary), F (Frequency) three variables [7]. Among them, R is the number of days between the nearest purchase time and observation points. M is the total purchase amount for customer. F represents the purchase number of customer in the observation window.

RFM model describes the customer's value through these three indexes. In the railway freight customer segmentation model, it is shown that R is the number of days between the nearest delivery time and the observation point; M is the total delivery payment of freight customer in the observation window, which reflects the customer's contribution to railway freight revenue; F is the delivery number of freight customer in the observation window. The RFM model reflects the overall profile of the customer, which provides a solid foundation for the personalized customization model.

2.1.2 Value of railway freight customer based on RFM model

In the RFM model, there is no set of weights for the three parameters, that is, the role of the three parameters in the customer segmentation is the same. However, there is a big defect in the method of non discrimination. Miglautsch believes that the parameters of the RFM model should be set corresponding weights [8]. Similarly, this paper gives the weight of the parameters of the RFM model as [. $_R$,ω_F,ω_M]=[0.20,0.38.0.42], by using the analytic hierarchy process (AHP), combining with the expert consultation method, and considering the actual situation of the railway freight transportation. Thus, we obtain the value of the freight customer based on the RFM model, as Formula (1)

$$C \quad {}^{j}_{RFM} = \omega_R \times (1 \qquad {}^{j}_{R}) + \omega_F \times C_F^j + \omega_M \times C_M \qquad (1)$$

Where C_{RFM}^{j} is the comprehensive RFM score of the jth customer, ω_R、ω_F、ω_M are the weights of the parameter of R、M、F, and C_R^j、C_F^j、C_M are the standardized *R, F, M* values of the jth customer.

Considering the maximum and minimum values of *R, F, M*, the min-max normalization method is used so that the normalized values can be mapped to the [0,1] interval, as Formula (2)

$$C_i^{j\prime} = \frac{C_i^j - \min\limits_{1\le\varepsilon\le n}\{C_i^\varepsilon\}}{\max\limits_{1\le\varepsilon\le n}\{C_i^\varepsilon\} - \min\limits_{1\le\varepsilon\le n}\{C_i^\varepsilon\}} \quad i\in R,F,M\} \tag{2}$$

$C_i^{j\prime}$ is the standardization value of the ith parameter of the jth customer.

2.2 Customer segmentation algorithm design

In this paper, K-means algorithm is used to cluster the RFM model of freight customers. K-means algorithm is one of the classical algorithms to solve the clustering problem, and it is also the most commonly used algorithm in customer segmentation. *X* is a data set contained n elements with t dimension, $X = \{ x_1, x_2, x_3, \dots, x_n \}$, and $x_i = (x_{i1}, x_{i2}, x_{i3}, \dots, x_{it})$, *X* is divided into k clusters, $C=\{C_i, = 1,2, \dots, k\}$. K-means selects the center of each cluster randomly defined as δ_i, and divided the remaining objects into the nearest cluster by calculating the Euclidean distance between the remaining objects and the center *X* of each cluster. Then, the K-means calculates the average distance between the objects in each cluster to get the new cluster center. The algorithm repeats this process until the average error criterion function is stable at the minimum or less than the prescribed threshold.

Average error criterion function defined as Formula (3)

$$I(C) = \sum_{i=1}^{k}\sum_{j=1}^{m_i} f_{kj} \| x_j - \delta_i \|^2$$
$$f_{kj} = \begin{cases} 1, \text{ if } x_j \in C_i \\ 0, \text{if } x_j \notin C_i \end{cases} \tag{3}$$

Where $m_.$ is the number of the ith cluster, δ_i is the center of the ith cluster.

3. Simulation and Analysis

The K-means algorithm needs to calculate the Euclidean distance between any two points in the early stage, and the time complexity is $O(N^2)$, and K-means requires a large number of iterative computations, and the time complexity of the algorithm is bounded by $O(N \times K \times C)$, where N is the number of cluster members, K is the number of clusters, C is the number of iterations. In the face of a large number of railway freight customers, the algorithm is faced with the problem of large computation cost. In order to solve the above problems, this paper created a big data platform based on Hadoop cluster, and used the Spark computing framework to calculate the distributed data. Because of the large number of railway freight customers, and K-means algorithm needs a lot of iterative calculation, we chose Spark as the computing framework of big data platform.

3.1 Simulation data

The simulation data is sourced from Railway Freight Customer Waybill Information ranged from January 2016 to December 2016. We chose 20185160 data randomly, which included railway freight customer delivery time, delivery frequency, transportation costs and other information, meeting the parameters of RFM model.

3.2 Simulation environment

The simulation platform used 5 PC as the server node in LAN, each PC equipped with virtual machine and Linux operating system, and equipped with 4G memory and 500G hard disk storage. Each PC installed Java Linux development package based on JDK, and installed the Hadoop version 2.6.3; Spark version 1.6.2; Sacla version of 2.11.8. Hadoop wrote simulation code by java, Spark wrote simulation code by Sacla. Simulation platform used Hadoop YARN mode, with 1 PC as master, and the other 4 PC as slave, the Spark job submitted to the Hadoop cluster which was used for job allocation of resources.

3.3 Simulation flow

We cleaned 20185160 data of railway waybill, by excluding missing values, secret value, independent value of abnormal value, obtained 19586317 data. The data was sent to the Hadoop HDFS system in the form of text to obtain the final results of customer segmentation through the Spark computing framework for distributed data calculation. The simulation steps are as follows:

Step1: HDFS sent the simulation data obtained from the Spark framework to 4 slave PC, then stored the data in the corresponding PC memory.

Step2: Spark preprocessed the data; firstly we clean the simulation data, by removing the abnormal value, and then got 19586317 data. Secondly, we integrate customers' data and got 975839 data totally.

Step3: Using Spark for distributed computing of RFM model and got 27942 data by converting the customer data with customer RFM information.

Step4: Using the measures of AHP and expert information to determine the cluster number of the K-means algorithm is 5 clusters, namely, gold customers, big customers, high-value customers, maintaining customers, low value customers.

Step5: Carrying on the K-means cluster analysis to the simulation data using Spark distributed learning library MLLIB, and return the result to HDFS.

3.4 Simulation results

We get the simulation results of the freight customer segmentation on the Hadoop platform by using Spark computing framework and using the K-means algorithm to carry out cluster analysis and simulation of the RFM model of railway freight customers. The results are shown in Table 1.

Customer segmentation	**Ratio(%)**	**number**	**R center**	**F center**	**M center**	**C center**
gold customers	5.68%	1587	0.11	0.43	0.66	0.4626
big customers	14.73%	4115	0.23	0.21	0.37	0.2812
high-value customers,	41.28%	11534	0.31	0.091	0.09	0.1344
maintaining customers	12.54%	3504	0.34	0.029	0.12	0.1294
low value customers	25.77%	7202	0.47	0.007	0.02	0.1051
average		6659	0.29	0.15	0.25	0.2225

Table1 Simulation result of freight customer segmentation

3.5 Analysis of simulation result

From the simulation results in Table 1, Gold customers account for less, accounting for only 5.68%, and they have the lowest R and highest F and highest M, which meant they have a highest short-term delivery capacity and highest delivery frequency, and the greatest contribution to railway freight revenue; Big customers accounted for 14.73%, such customers' frequency of delivery and revenue contribution are high in all freight customers relatively; High value customers account for the highest proportion of all freight customers, reaching 41.28%; Maintain customers account 12.54%, they show the recent delivery capacity and delivery frequency is low, but its contribution to railway freight revenue is higher; Low value customers, accounting for up to 25.77%, however, their indicators in RFM are very low, which means their value and potential are low, and the loss of them is strong. Railway freight transportation needs to formulate corresponding marketing strategy according to different freight customers.

4 Conclusions

This paper proposed a RFM model for customer segmentation based on customer characteristics of railway freight transportation, and used K-means clustering algorithm to calculate this model. Then we established the big data Hadoop cluster and used Spark computing framework to simulate the

railway freight data based on RFM. We obtained the segmentation results, and provided the customer marketing advice of various types of customers on the basis of the analysis of the results.

References:

[1] YU Xiao-bing, CAO Jie, GONG Zai-wu. Review on customer churn issue[J]. Computer Integrated Manufacturing Systems, 2012, 18(10):2253-2263.

[2] YANG Jin-hua, LI Ming-bo, JI Ling. The application of customer relationship management theory in rail freight Marketing [J]. Railway Transport and Economy, 2008, (2)30:87-90.

[3] WANG Qi-dong, ZENG Wei-dong. Thoughts and Discussion on the construction of railway freight transportation customer relationship management system[J]. Railway Transport and Economy, 2011, 33(1):35-38.

[4] WU Su-hui, CHENG Ying, ZHENG Yan-ning, PAN Yun-tao. Survey on K-means algorithm [J]. New Technology of Library and Information Service, 2011, 27(5):28-35.

[5] ZHONG Yan,GUO Yu-song. Research of applying data mining in segmentation of railway freight customer [J]. Journal of Beijing Jiaotong University, 2008,32(3):25-36.

[6] DENG Cheng, YANG Zhuang-ying, GU Jun-jie, CAI Zhi, LI Yue. Improved K-means Algorithm and its application in railway customer segmentation [J]. Railway Computer Application, 2014, 23(6):45-48.

[7] Hughes A M. Strategic Database Marketing [M]. Chicago: Probus Publishing Company, 1994.

[8] Miglautsch J R. Thoughts on RFM Scoring [J]. Database Mark, 1995, 8(1).