ATLANTIS PRESS
Advances in Social Science, Education and Humanities Research (ASSEHR), volume 156
2nd International Seminar on Education Innovation and Economic Management (SEIEM 2017)

Customer Segmentation of Third Party Review Website Based on Cluster Analysis

Shan Gao
Wuhan University of Technology
Wuhan, China
gaooshann@163.com

Changbin Jiang
Wuhan University of Technology
Wuhan, China
175521@qq.com

***Abstract*—This paper made segmentation of customers on the website of Dianping Holdings and identified the characteristics of different types of customers, so as to give recommendations to different clients. The paper used locomotive collector to grasp the customer information on the website of Dianping Holdings. After processing the original data, 38791 pieces of final data were got. Then, SPSS software was used in cluster analysis and discriminant analysis. In the end, the paper subdivides the customers on the Dianping Holdings website into six categories: hyperactive customers, active customers, moderately active customers, low-value customers, potentially high-value customers and high-value customers. It also finds that the number of female customers is more than male customers in six categories.**

Keywords—cluster analysis; discriminant analysis; customer segmentation;Dianping Holdings website

I. INTRODUCTION

Third party review websites are platforms on which people can comment about the goods or the experience. Their entire value lies in the comment information[1]. Third party review websites develop from e-commerce and build communication bridge between customers and merchants. The scarce resources can be contributed, created and shared by people through the websites among which the most representative website is Dianping Holdings website. Dianping Holdings website was established in 2003 and it was the first established third party review website in China. Dianping Holdings website provides users with a platform on which customers can communicate with each other [2].

With the fierce competition among enterprises, more and more enterprises realize the importance of customers. They begin to focus on customers and the core work is made based on the demand of customers. The traditional method of marketing is that customers are unified and enterprises make the same strategy. However, with the development of society, there are various requests among customers, and each customer has different characteristics. The key to the development of contemporary enterprises lies in customer segmentation. So enterprises make differentiated marketing strategies and provide various products and services to meet the demand of different customers.

II. LITERATURE REVIEW

In the current study, researchers used the method of cluster analysis to segment different kinds of research objects. Du Wei[3] used the improved clustering algorithm to subdivide the tourism customers and verified the feasibility and efficiency of the improved algorithm. Zhao Ming[4] used K-means algorithm of cluster analysis to classify the customers of commercial banks and effectively identified the customers of commercial banks. Foreign scholar SMS Hosseini[5] proposed a new method based on extended RFM model which included K-means algorithm to classify the loyalty of customers. Yi Famin[6] established an evaluation index system of agricultural e-commerce website to do cluster analysis on the agricultural websites. He Shengyu[7] created an e-commerce development level measurement model based on the e-commerce development index system which was released by the CII and used the factor analysis and cluster analysis method to analyze the level of development of electronic commerce in China. Zhai Lili[8] proposed a collaborative filtering algorithm based on context clustering optimization, using K-means algorithm to subdivide the mobile e-commerce users and she combined with 'firefly' algorithm to improve the initial point on the basis of collaboration Filter to improve the accuracy of the recommended results. Li Xinxin[9] used the improved K-means algorithm based on semi-supervised neighbor propagation to subdivide customers of cosmetic website.

After sorting out the literature, it can be found that there is little literature about customer segmentation of e-commerce, and there is little research on customer segmentation of third party review websites. The traditional research object is about financial customers, travel customers, telecommunication customers and so on. This paper takes Dianping Holdings website as an example and collects customer information to make an analysis of customers. Third party review websites are based on the Web2.0 environment; customers can freely communicate with each other on the websites or express their views without any form of restraint. Therefore, taking Dianping Holdings website as the research object is the innovation of this paper.

III. DATA ANALYSIS

A. Data Preparation

This paper took Dianping Holdings website as an example and used the locomotive collector to grasp the customer information. The locomotive collector is the software that can be used to crawl, process, analyze and mine data. It can flexibly capture the scattered data in the web page and extract

the required data accurately through a series of analysis and processing. It has functions of website acquisition, content collection, content publishing, POST URL acquisition and content filtering. Locomotive collector is the most popular data acquisition software[10].

Customer information of Dianping Holdings website included: name, gender, contribution value, community level, registration time, comments, number of favorites, attendance, and number of published pictures, lists, posts, followers, fans, interactive value and user URL. The contribution value showed the contribution of customers made to the website. The higher the contribution value, the more likely the customers were to get the trust and respect of other customers. The website had a function which needed the contribution value to reach a certain value and then customers could use the function. Customers with high contribution values would have higher priority in some activities. Community grade was determined by community credit, and community credit needed to be got in the community of Dianping Holdings website. The community level was divided into wild child, preschool, the first grade, the second grade and so on, which had a total of 26 levels. The lists indicated that the customer himself or herself published and shared a list of products or experience that other customers could read and comment on. The posts indicated the number of original posts written by the customers in the community of Dianping Holdings website.

There were 15 customer information attributes, and the customer information attributes were divided into 3 categories which were the basic attributes, behavioral attributes and value attributes. The basic attributes of customers reflected customers' basic information, such as name, age, level of education and so on, and through the basic attributes customers could generally be understood. But if only depended on the basic attributes of customers, the accuracy rate would not be high. Therefore, with the basic attributes of customers combining with behavioral attributes and value attributes, comprehensive understandings of the customers could be got and the efficiency and accuracy of customer segmentation could be enhanced. Behavioral attributes referred to the behavior patterns which were displayed by customers in the process of using the website. The value attributes referred to the attributes which could reflect the value of customers to the website. The basic attributes of customer information included: name, gender and registration time. Behavior attributes included: comments, number of favorites, registration time, number of published pictures, lists, posts and followers. And Value attributes included: contribution value, community level, fans and interactive value.

Firstly, the raw data was processed by excluding invalid data, empty data and repeated data and 38791 valid customer data was remained. Then customer data was encoded. Male was coded as 1 and female was coded as 2. In community level encoding, the wild and preschool customers were coded as 1, and grade 1 to 6 customers were coded as 2, grade 7 to 9 customers were coded as 3, and so on. After removing name and user URL, 38791 valid data was got which contained gender, contribution value, community level, registration time, comments, number of favorites, attendance, and number of published pictures, lists, posts, followers and fans, as well as interactive value.

Then the data was processed to remove outliers. The traditional way of dealing with outliers is to remove the observation of the outliers. However, the method used in this paper was to replace the outliers with specific values. The method of processing was that the value was replaced by the value of 99% quantile when the value is greater than the value of the 99% quantile. After the exception value processing, data conversion was needed. The purpose of data conversion was to maximize the extraction and use of information. Before data conversion, the distribution of data needed to be explored. Take contribution value and comments as examples.

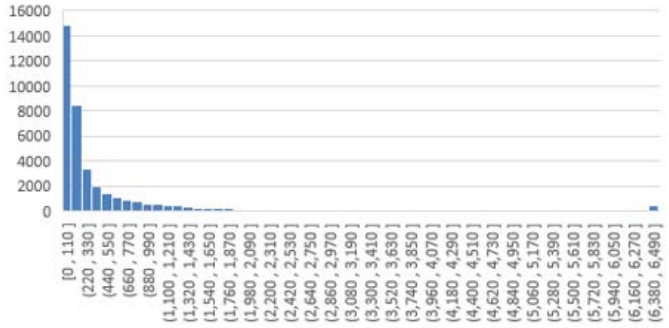

Fig. 1. Distribution map of contribution value

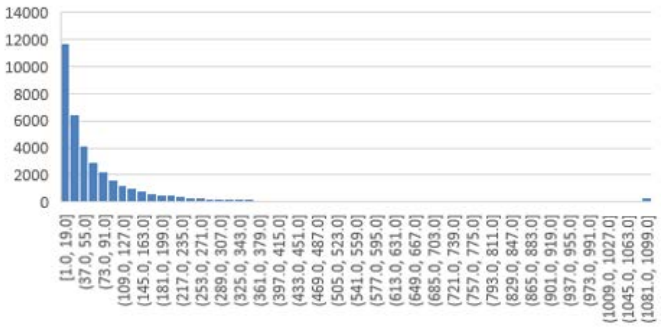

Fig. 2. Distribution map of comments

Fig. 1 and Fig. 2 were distribution maps of customer contribution and comments. From the two distribution maps, it could be seen that data distribution was extreme skewed and data conversion was needed. The data conversion method used in this paper was converted by the formula $\log_{10}(var+1)$, and the distribution after data conversion was shown in Fig. 3 and Fig. 4. The degree of skewness of the data distribution after logarithmic transformation was obviously reduced which was helpful for the later cluster analysis and the interpretation of the results.

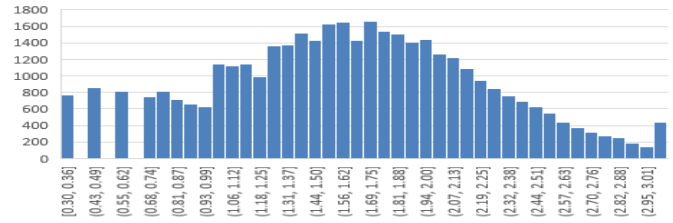

Fig. 3. Distribution map of contribution value after conversion

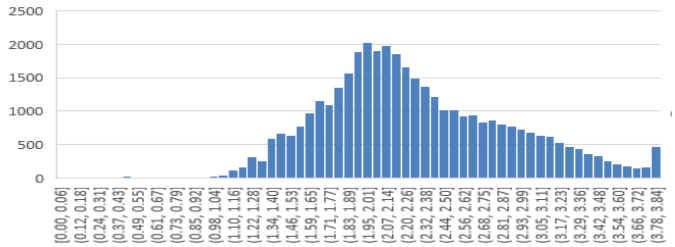

Fig. 4. Distribution map of comments after conversion

After data conversion, data needed to be standardized. The average value of the data was 0 and the standard deviation was 1. This process provided the basic base for the later data cluster analysis.

B. Data Processing

1) Cluster Analysis

In this paper, the cluster analysis method was K-means and the number of clustering K was 6. SPSS software was used in this analysis. The results were shown in Table 1.

TABLE I. ANOVA TABLE

	K=6	
	F	***Sig.***
contribution value	14599.42	0.000
community level	22129.31	0.000
registration time	729.93	0.000
comments	10399.52	0.000
number of favorites	2324.95	0.000
attendance	6142.39	0.000
number of published pictures	12815.36	0.000
lists	76642.10	0.000
posts	17468.46	0.000
followers	3706.85	0.000
fans	6033.91	0.000
interactive value	8826.75	0.000

The ANOVA table contained the F value and the associated probability Sig. The larger the F value was, the better the clustering effect was. As it could be seen from Table 1, the maximum F value was the number of lists, the results showed that all the Sig values were 0 which were less than Sig = 0.01 significant level. This indicated that the number of clusters is good because all variables are significant.

2) Discriminant Analysis

In this paper, SPSS software was used in discriminant analysis and the method was Fisher. 38791 customer data was put into the model including 19396 data which was training data and 19395 data for unknown classification data. The final results were shown in Table 2.

TABLE II. DISCRIMINANT RESULT

		Discriminant group						**Total**
		1	***2***	***3***	***4***	***5***	***6***	
Original group	1	127	11	0	0	0	0	138
	2	2	200	0	0	4	0	206
	3	0	5	10244	181	0	12	10442
	4	0	0	221	7250	0	34	7505
	5	0	0	0	0	48	0	48
	6	0	0	0	108	0	948	1056
%	1	92.0	8.0	0.0	0.0	0.0	0.0	100.0
	2	1.0	97.1	0.0	0.0	1.9	0.0	100.0
	3	0.0	0.0	98.1	1.7	0.0	1.0	100.0
	4	0.0	0.0	2.9	96.6	0.0	0.5	100.0
	5	0.0	0.0	0.0	0.0	100.0	0.0	100.0
	6	0.0	0.0	0.0	10.2	0.0	89.8	100.0
A: 97% accuracy								

As it could be seen from Table 2, the results of the success rate were 97%. The results of the cluster analysis and discriminant analysis are great which showed that the cluster analysis was reasonable to subdivide the customers into 6 categories.

IV. CLUSTERING RESULTS

Customers of Dianping Holdings website are divided into 6 categories which are C1, C2, C3, C4, C5 and C6, respectively.

TABLE III. CLUSTERING RESULT

	C1	**C2**	**C3**	**C4**	**C5**	**C6**
total number of people	371	508	21662	13743	132	2375
female	82%	73%	69%	80%	67%	84%
male	18%	27%	31%	20%	33%	16%
contribution value	4826	1566	100	695	9721	3035
community level	4.5	1.6	1	1	4	3.6
registration time	7.2	6.8	4.3	5.4	7.8	6.2
comments	822	324	26	138	1390	498
number of favorites	153	145	24	111	310	127
attendance	700	245	2.5	61	2006	349
number of published pictures	2826	765	12.8	324	6406	1756
lists	1.7	1.5	0	0	13	0
posts	155	2	0.06	0.24	322	52
followers	358	111	14	80	227	160
fans	288	100	12	66	549	146
interactive value	1355	241	3	81	6695	136

C1 are active customers. They account for 0.95% of the total number of customers. The contribution value is very high and the average registration time is more than 7 years. C1 customers are loyal customers who have long registration time, are willing to share pictures, publish posts and actively participate in community activities. C2 are moderately active customers. They account for 1.3% of the total number of customers. The contribution value is higher than C1 customers

and the average registration time is more than 6 years, but the value of community level is low. Therefore, C2 customers are mainly active in publishing comments on the Dianping Holdings website and publish pictures, but they are not actively participating in community activities.

C3 are low-value customers. They account for 55.8% of the total number of customers. The customers' contribution and interaction values are the lowest among all customers, and the customers' behavioral attributes and value attributes are at a low level. These customers do not participate in community activities and rarely share resources with other customers. But the length of registration has been more than four years which indicate they use Dianping Holdings website to read other customers' comments for reference. C4 are potentially high-value customers. They account for 35.4% of the total number of customers. The contribution value and interaction value of C4 customers are generally low, but the average registration time is longer than C3 customers. It shows that contribution value and interactive value of C4 customers have improved significantly than C3 customers.

C5 are hyperactive customers. They account for 0.34% of the total number of customers. C5 customers are the most active customers with the highest value of interaction and the longest average registration time. They are willing to comment on the products or experience and are also happy to share photos to help other customers to judge. C6 are high-value customers. They account for 6.1% of the total number of customers. The contribution value is high, the community level is high and the average length of registration time is 6 years. This kind of customers actively participates in community activities and share pictures or other resources with other customers.

V. Suggestions and Prospects

Based on the above findings, the following suggestions are made for different categories of customers.

Firstly, for C1, C2 and C5 customers, Dianping Holdings website needs to promote the customers' willingness to continue using the website and improve customer loyalty. For example, website can push the latest information to these customers and provide them with the opportunity to try free products, so that customers will continue to participate in and share their experiences. C1 are active customers, C2 are moderately active customers and C5 are hyperactive customers. They have high contribution and interaction value, especially for C5 customers. Therefore, these customers are contributors of the Dianping Holdings website and constantly release information which can attract new users.

Secondly, for C3 customers, Dianping Holdings website needs to enhance the continuous use intention of customers by optimizing the layout of the web interface, doing customer survey to understand what customers concerned about and then displaying the content of information with high quality page which will improve customer satisfaction and enhance the customer continuance intention.

Thirdly, for C4 customers, Dianping Holdings website needs to enhance the continuous use intention of this kind of customers and promote them to share information. For example, the website can encourage customers to share information through making hot topics or sharing vouchers.

Fourthly, for C6 customers, the website should focus on the viewpoints of these customers about the community and enhance the content layout of the community of Dianping Holdings website.

Fifthly, the number of female customers is more than male customers in six categories. For female customers, the website can launch some activities and make hot topics targeting about women. For example, the topic is about 'Top Ten hot afternoon tea dessert places' to promote female customers to read and participate in the topic or the website can give customers who are female and their comments reach a certain value vouchers thatt can be used on afternoon tea or manicure.

This paper grasped the customer information from Dianping Holdings website and used the SPSS software in cluster analysis and discriminant analysis. Finally the results of the customer segmentation were got. However, in this paper, the research limitation is that the discussions about how to attract new customers and retain old customers are not enough. It will be strengthened in the future.

References

[1] Lv Xiuying. Analysis of the development status of China's third party review websites under Web2.0 environment -- taking public comment network and watercress network as an example, [J]. Journal of Southeast University (PHILOSOPHY AND SOCIAL SCIENCES EDITION), 2011, 13 (S1): 87-92. (In Chinese)

[2] You Jianxin, Meng Yinwei. Web data mining site knowledge acquisition and application of the dianping.com as an example based on [J]. Journal of Shanghai Univer (NATURAL SCIENCE EDITION), 2014, 20 (3): 261-273. (In Chinese)

[3] Du Wei, Zhao Chunrong, Huang Weijian. Application of improved k-means clustering algorithm in customer segmentation [J]. Journal of Hebei University of Economics and Business, 2014, 35 (1): 118-121. (In Chinese)

[4] Zhao Ming, Li Xue, Li Xiuting, et al. Classification of fund clients in commercial banks based on cluster analysis [J]. management review, 2013, 25 (7): 38-44. (In Chinese)

[5] Hosseini S M S, Maleki A, Gholamian M R. Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty[J]. Expert Systems with Applications, 2010, 37(7):5259-5264.

[6] Yi Famin, Rong kyumin. Guangdong Province agricultural e-commerce website clustering analysis [J]. Chinese Technology Forum, 2010 (3): 129-133. (In Chinese)

[7] He Shengyu, Ma Huijie, Teng Xihua. Factor analysis and cluster analysis on China's level of development of [J]. e-commerce based on the reform of economic system, 2017 (2): 196-200. (In Chinese)

[8] Zhai Lili, Xing Hailong, Zhang Shuchen. Collaborative filtering recommendation for mobile e-commerce based on scenario clustering optimization [J]. information theory and practice, 2016, 39 (8): 106-110. (In Chinese)

[9] Li Xinxin. Application of clustering algorithm in e-commerce customer segmentation [D]. Ocean University of China, 2012. (In Chinese)

[10] Locomotive collector [EB/OL]. 20160531 http://www.locoy.com/.