Consumption capability analysis for Micro-blog users based on data mining

Yue Sun

Beijing University of Posts and Telecommunication
Beijing, China
Email: sunmoon5723@gmail.com

ABSTRACT

Data mining is an effective method of discovering useful information in a large amount of data. The capability of understanding the user's consumption is vital for a company. Discovering the significant customers allows the company to focus on the most valuable customers. This paper uses micro-blog users' check-in data and shop information for analysis and cluster method of data mining. We analyze user's spending ability quantitatively based on user's check-in actions. Compared with other clustering method, we choose DBSCAN clustering method of data mining to analyze the shop information and position. Users are divided into different categories according to their spending power. Discovering the users with high consumption level in large amounts of data, which is significant for a firm, can help the firm develop better strategies.

KEYWORDS

DBSCAN clustering method, user consumption action, Micro-blog user

1. INTRODUCTION

It is widely acknowledged that 80% of the revenue of a corporation comes from important clients that take up 20% of the total, while the remaining 20% of the revenue comes from the ordinary clients that take up 80%. As a result, we can conclude that not all clients are of the same value to a corporation, and it is significant to discover and keep this 20% of important clients.

Microblog is a platform for sharing, propagating, and accessing information based on client connection. With the fast development of the Internet in recent years, microblog, as a newly-born pattern of communication, has blossomed into an industry with hundreds of millions of clients around the world.

Based on the data of the microblog users, we are to extract useful information using the technology of data mining. Specifically, we are to evaluate their consumption capability to identify the 20% important clients, making it convenient for corporations to concentrate on the clients with the most values and potentials. Also, we are to learn the characteristics of the consumptions of these users, which provide the opportunity for corporations to offer personalized services to increase user fidelity.

2. RELATED WORK

MicroBlog platform provides a function called check-ins, which means the users can select their current region through mobile or network location system when they tweet a new message. So other users related to them will know their current position together with the new message. The

DOI : 10.5121/ijaia.2013.4404

check-in point is decided by mobile position service or computer IP address, so the users' records including time, position, longitude, latitude are real and reliable.

2.1 User Characteristic Analysis

According to survey data, there are 82.3% MicroBlog users under the 35 years old.[1] So the analysis results we got in the paper can only reflect the consumption characteristics of young MicroBlog users group. In addition, we suppose that all of the user's check-in actions are their general behavior in life.

2.2 Check-in Place Characteristics Analysis

Poi is short for point of interest that is user's check-in place. The scope of Users' check-in places are wide, which include different categories such as restaurant, hotel, park, hospital, subway station, market, school, office building. To analysis users' characteristics in different aspects, we classify all of poi data. In this paper, we take diet service for example.

We select the food and beverage service poi from total poi data that can represent user diet consumption capability. These diet service poi include restaurant, bar, coffee house, drinking bar. Through practical survey and statistics, we have got the poi shops' information including price, position, shop area, category. The price of shop means average per capita cost.

Among the whole shop information we collect, average per capita cost is the suitable attributes of a shop which can reflect its rank. Meanwhile, the shop position decides the different business area the shop belongs to. These business areas are quite different in aspect of prosperous degree that will influence classification of the shop. Therefore we make a shop grade division according to attributes we selected before, namely per capita consumer cost and its geographical location.

3. METHOD SELECTION

We select the clustering method instead of choosing classification method because classification methods typically require a high price collection and a large number of marked training set [1]. Our analysis objects are micro blog users, and we were in trouble when verifying user attributes' authenticity. So it is difficult to get the training sets that are required for classification. We select the clustering method which dividing data set into group based on data similarity and its advantages are less data quantity and easy to adapt changes.

There are many clustering methods used in data dining, such as the most commonly used partition method -- K-means algorithm which is based on distance, but can not find clusters of arbitrary shape. Density-based BVSCAN algorithm can found arbitrary shape. This clustering method will regard cluster as object region that is segmented by low-density regions representing noise in data space and can discover clusters of arbitrary shape in data space including noise.

DBSCAN discover clusters by checking the e neighborhood of each point. If the e neighborhood of P point contains points more than MinPts (namely Minimum points), then create a new cluster which p is a core object. DBSCAN then iteratively gathered the object that is directly density-reachable from the core object.

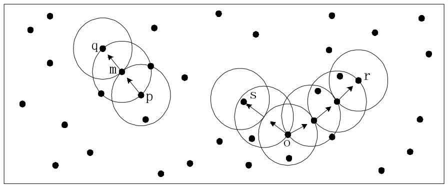


Figure 1. DBSCAN method of discovering clusters

DBSCAN algorithm need choose two parameters artificially, namely MinPts and radius of specified circle. The parameter choice only depends on developer's experiment and different parameter chosen may cause different clustering results. In order to get an optimal clustering result, we attempt range below to choose a suitable parameter. Set parameter ε from 0.01 to 0.3, MinPts from 3 to 6.

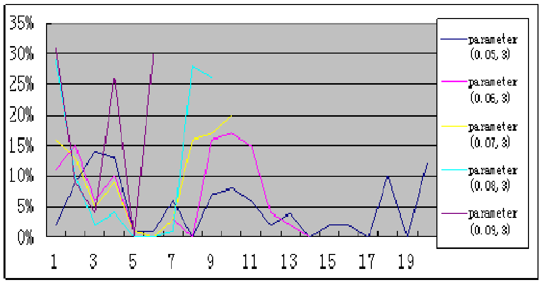


Figure 2. Percentage curves of points in different clusters

The horizontal axis represents cluster number, and the vertical axis represents point number percentage in a cluster.

All data used are 16015 poi records. In the testing, if parameter ε> 0.1, we cannot get multiple clusters but only one cluster. If parameter ε< 0.05, the cluster amount is more than 20 which is too large for the following classification and 25 percent clusters is no meaning for classification because the percentage of the point number in cluster is approximately equal to zero. The points in same cluster represent these shops with similar attributes, so too little points in one cluster is not a suitable. Changing value of MinPts has no apparent effect on clustering result. Selecting parameter ε= 0.8 and MinPts = 3, the clustering result is ten clusters.

Calculate the average price of shops in the same category based on the average per capita cost of different shops. We use normalization processing to obtain a value within the range of 0 to 1 and represent the different categories of shops weights.

Results are as follows：

AV: Average Price
C: Cluster
Unit: Yuan

	C1	C2	C3	C4	C5
AV	52.24	13.54	84.95	217.27	138.60
	C6	C7	C8	C9	C10
AV	556.86	957.79	356.66	52.48	31.44

Table 1. Average price of shops in different categories

Results after normalization processing：

	C1	C2	C3	C4	C5
AV	0.055	0.014	0.089	0.227	0.145
	C6	C7	C8	C9	C10
AV	0.581	1.000	0.372	0.055	0.033

According to user records, we statistic the total number of user records in different consumption levels. The amount of data is different between different users, which will impact on the results. So use normalization processing on the user's check-in data.

Total cost of one user = $\sum_{0}^{9} C_i \times V_i$

V_i = i Cluster record / total record（for a user）

C_i: i Cluster

The total cost represents the normalized user spending which can reflect the level of consumption of each user, but is not the user's actual consumption value.

Select ten users randomly and draw their consumption cost curve of different level.

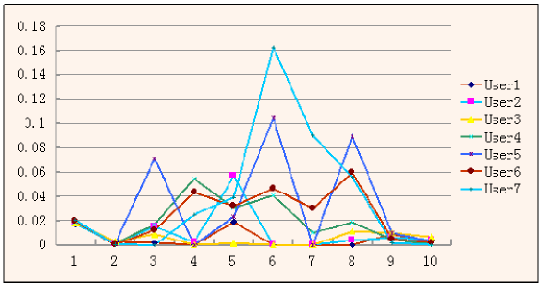


Figure 3. User consumption cost curve

This approach helps us to quantify the user's food and beverage consumption level. It allows us to intuitive understanding of the user's spending power. The maximum value representing consumption power is 10, and the minimum value is 0. Larger value indicates user's stronger spending power, on the contrary, the user spending power is weaker.

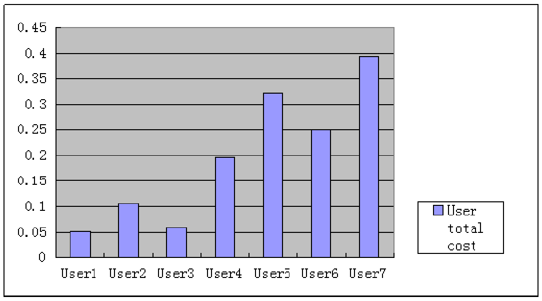


Figure 4. User consumption ability

4. TESTING THE METHOD

In order to verify the rationality of this method, we use MATLAB to fit all user consumption value, and the get the normal distribution graph.

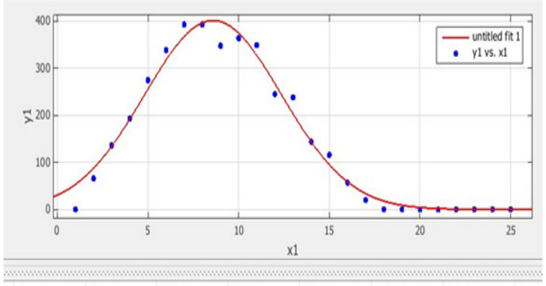
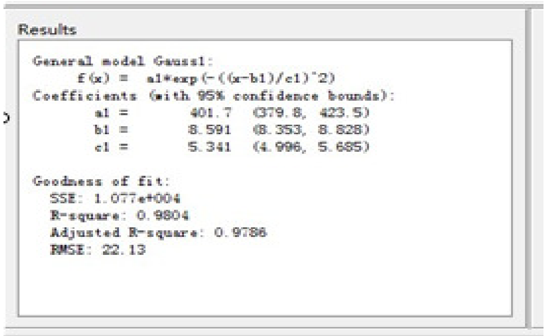


Figure 5. Matlab testing curve



```
Results
General model Gauss1:
     f(x) =  a1*exp(-((x-b1)/c1)^2)
Coefficients (with 95% confidence bounds):
       a1 =       401.7  (379.8, 423.5)
       b1 =       8.591  (8.353, 8.828)
       c1 =       5.341  (4.996, 5.685)

Goodness of fit:
  SSE: 1.077e+004
  R-square: 0.9804
  Adjusted R-square: 0.9786
  RMSE: 22.13
```

Figure 6. Matlab testing parameter

Goodness of fit is 0.9805, and adjusted goodness of fit is 0.9786. μ equals to 8.591. In order to facilitate the mapping, the horizontal axis x1 that represents the user consumption value has been enlarged 50 times. The actual value of μ is 0.17182. It means the average consumption value of the micro-blog user groups is 0.17182. As we have learned, the spending power of an ordinary consumer groups should be consistent with the normal distribution. We verify the feasibility of this approach of analyzing micro-blog user's spending power.

It can be learned from the results of our analysis, the consumption cost of users of whom the spending power is more than 0.2354 is accounted for 80% of the total. And the number of these users is accounted for 30% of the total. This also proved to us that 20% of customers would bring the company 80% of the profits.

5. CONCLUSION

From above discussion, we get an analytical method of the spending power of the user, which is dividing all poi locations into different levels by DBSCAN clustering method and quantify the user's consumption ability by adding different weight. In order to get the appropriate DBSCAN parameter, we compared the different parameter values on the clustering results. Through the distribution of user's theoretical consumption cost, we obtain the top 20% of users who can bring huge profits for the firm. It also show us that the distribution of consumer values of all users is normal distribution.

6. FURTHER WORK

According to the top 20% users of significant influence on the firm we have discovered, we can further analyze the behavioral characteristics of the consumer groups to develop them to become loyal users. To assess the user's consumption level, we have another method of statistics user's consumer value before clustering.

REFERENCE

[1] J. Han, M. Kamber, Data Mining: Concepts and Techniques, p. cm. London: Academic Press, 2001

[2] Margaret H. Dunham. Data Mining Tutorial [M]. Tsinghua University Press, 2005 33-46.

[3] M. Hu and B. Liu, "Mining and summarizing customer reviews," in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, Seattle, WA, USA, August 2004, pp. 168–177.

[4] A. Balahur and A. Montoyo, "A feature dependent method for opinion mining and classification," in Proceeding of International Conference on Natural Language Processing and Knowledge Engineering, Beijing, China, Octorber 2008, pp. 1–7.

[5] Girish Punj and David W. Stewart. Cluster analysis in marketing research: Review and suggestions for application. Journal of Marketing Research, 20(2):134–148, 1983.