Cluster Analysis of the Internet Industry Development in Different Regions of China Based on Improved *K-means* Algorithm

Jian Ma, Runtong Zhang, and Xiaomin Zhu

Abstract Although the rapid development of the network technique and the Internet economy, China's Internet industry has exposed many problems such as imbalanced trend among China's provinces. In this paper, we employ the concept and method of improved *K-means* algorithm to establish a clustering analysis model of Internet development level in various provinces of China based on the imbalance of Internet popularization in China's provinces, and analyze the shortcomings and causes of Internet development in each province in detail. Finally, we recommend the corresponding policy suggestions in view of addressing issues mentioned.

Keywords The Internet industry · *K-means* algorithm · Clustering analysis

1 Introduction

The Internet industry refers to engaging in Internet operation services, application services, information services, network products and a general term for the development and production of network information resources and other industries related to Internet research, education, and service.

This work is supported by the Fundamental Research Funds for the Central Universities with grant No. 2018YJS064, a key project of National Natural Science Foundation of China with grant No. 71532002 and a Major project of the National Social Science Fund with grant No. 18ZDA086.

J. Ma · R. Zhang
School of Economics and Management, Beijing Jiaotong University, Beijing, China
e-mail: 16113142@bjtu.edu.cn

R. Zhang
e-mail: rtzhang@bjtu.edu.cn

X. Zhu (✉)
School of Mechanical, Electronic and Control Engineering, Beijing Jiaotong University, Beijing, China
e-mail: xmzhu@bjtu.edu.cn

M. Li et al. (eds.), *IEIS2019*,
https://doi.org/10.1007/978-981-15-5660-9_3

The development of China's Internet started in the late 1980s. The development process can be roughly divided into four stages: the initial exploration stage, the basic network construction stage, the content active network popularization stage, and the current network prosperity stage.

With the popularity of the Internet and the advancement of technology, various forms of network applications are constantly emerging, and the field of Internet applications is constantly expanding. Internet applications range from early information browsing and e-mail development to online entertainment, information acquisition, communication, business transactions, and government services.

With the development of the economy and the advancement of society, under the guidance of technological innovation, the Internet has penetrated into all aspects of social and economic life and is bringing profound, unimaginable and unexpected changes to human society. As of 2013, global Internet users exceeded 2.2 billion; according to the 33rd Internet report released by China Internet Network Information Center CNNIC, as of the end of 2016, the number of Internet users in China reached 731 million, ranking first in the world; Internet penetration rate reached 53.2%. However, as a developing country with a large poverty gap, the Internet penetration rate in some regions is still not high, and it is necessary to conduct a detailed analysis of the development level of the Internet industry in various provinces in China.

Liu believes that the Internet has a unique advantage for social science popularization. How to make full use of the Internet as a platform to play its role in the popularization of social sciences has become a major issue in adapting to the trend of the times and exploring new channels for popularization of social science [1]. Bao et al. believe that although the Internet penetration rate in China has exceeded the world average, due to low coverage, low tariffs, and information security, there are still gaps with developed countries, and there is still much room for development. Due to the high demand of users on the Internet, the popularity of the Internet in different categories of population in China is not the same [2]. Sun has quantitatively analyzed this in five aspects: gender, marriage, age, education level and industry [3]. Wang et al. judged the stage of Internet penetration in various provinces and cities, and based on the inter-provincial panel data from 2005 to 2016, conducted δ-convergence and β-convergence tests on Internet popularity [4]. Xue et al. use the method of principal component to construct an indicator system for traffic evaluation in various provinces in China [5].

Clustering analysis is an important part of data mining. The *K-means* clustering algorithm is a basic partitioning method in cluster analysis method, and it is also an unsupervised machine learning method. It has the advantages of high efficiency, easy understanding and implementation. At the same time, it can cluster a variety of data types and is widely used in many fields, such as applied mathematics, pattern recognition, image segmentation and bioengineering. In this paper, the machine learning algorithm is applied to the analysis of Internet penetration rate in various provinces of China, and the unsupervised learning method is adopted. Improved *K-means* cluster analysis is carried out on the Internet data of each province in 2016, and corresponding policy recommendations are given.

2 Improved *K-means* Clustering Analysis Model

2.1 Selection of Optimal K Value—Contour Coefficient Method

The core indicator of the method is the Contour Coefficient. The Contour coefficients of a sample point x_i are defined as follows:

$$S = \frac{b - a}{\max(a, b)} \tag{1}$$

where a is the average distance between Xi and other samples of the same cluster, called the degree of cohesion, and b is the average distance between x_i and all samples in the nearest cluster, called the degree of separation. And the definition of the most recent cluster is:

$$C_j = \arg\min_{c_k} \frac{1}{n} \sum_{p \in C_k} |p - x_i|^2 \tag{2}$$

where p is a sample in a cluster C_k. In fact, the average distance from all samples of x_i to a cluster is used to measure the distance from the point to the cluster, and then the cluster closest to x_i is selected as the nearest cluster.

After obtaining the contour coefficients of all the samples and then averaging, the contour coefficients are obtained. The average contour coefficient has a value range of $[-1, 1]$, and the closer the distance between the samples in the cluster, the farther the distance between the samples, the larger the average contour coefficient, and the better the clustering effect. Then, naturally, the k with the largest average contour coefficient is the optimal number of clusters.

2.2 K-means Algorithm Model

The *K-means* algorithm takes distance as the standard for the measure of similarity between data objects, usually using the Euclidean distance to calculate the distance between data objects.

Euclidean distance calculation formula:

$$dist(x_i, x_j) = \sqrt{\sum_{d=1}^{D} (x_{i,d} - x_{j,d})^2} \tag{3}$$

Using Sum of the Squared Error (SSE) as the objective function of clustering, defined as J. Where D represents the number of attributes of the data object. In the

clustering process of *K-means* algorithm, each iteration, the corresponding cluster center needs to be recalculated (updated): in the corresponding cluster, the mean of all data objects is the cluster center of the cluster after the update. The center of the cluster that defines the kth cluster is Center k, and the cluster center is updated as follows:

$$Center_k = \frac{1}{C_k} \sum_{x_i \in C_k} x_i \tag{4}$$

where K represents the number of clusters. When the difference between two iterations J is less than a certain threshold, that is, $\Delta J < \delta$, the iteration is terminated, and the obtained cluster is the final clustering result.

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} dist(x_i, Center_k) \tag{5}$$

$$\begin{aligned} \frac{\partial}{\partial c_k} J &= \frac{\partial}{\partial c_k} \sum_{i=1}^{K} \sum_{x \in C_i} (c_i - x)^2 \\ &= \sum_{i=1}^{K} \sum_{x \in C_i} \frac{\partial}{\partial c_k} (c_i - x)^2 \\ &= \sum_{x \in C_i} 2(c_i - x) = 0 \end{aligned} \tag{6}$$

$$\sum_{x \in C_k} 2(c_k - x_k) = 0 \Rightarrow m_k c_k = \sum_{x \in C_k} x_k \Rightarrow c_k = \frac{1}{m_k} \sum_{x \in C_k} x_k \tag{7}$$

The *K-means* algorithm can be described as:

- *Initialize K class cluster centers.*
- *Calculate the distance of each data object to the cluster center, and divide the data object into the clusters of the cluster center closest to it.*
- *Update the cluster center based on the resulting cluster.*
- *Continue to calculate the distance of each data object to the cluster center, and divide the data object into clusters in the cluster center closest to it.*
- *Continue to update the cluster center based on the resulting cluster.*
- *Iterating until the maximum number of iterations T is reached, or the difference between two iterations J is less than a certain threshold, the iteration is terminated, and the final clustering result is obtained.*

The detailed process of the algorithm is described as Fig. 1.

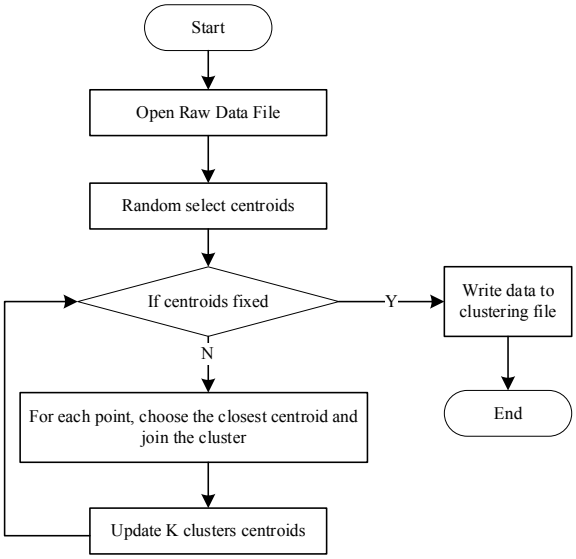

Fig. 1 *K-means* clustering algorithm flow chart

2.3 K-means++ Algorithm Model

- *K-means++ algorithm is an improvement of k means algorithm. The basic idea of K-means++ algorithm to select initial seeds is that the distance between the initial cluster centers should be as far as possible.*
- *Randomly select a point from the set of input data points as the first cluster center.*
- *For each point x in the dataset, calculate its distance D(x) from the nearest cluster center (referring to the selected cluster center).*
- *Select a new data point as the new cluster center. The principle of selection is: the point where D(x) is larger, the probability of being selected as the cluster center is larger.*
- *Repeat 2 and 3 until k cluster centers are selected.*

$$Random = \text{sum}(\text{D}(\text{x}))^{*} \tag{8}$$

Compared with the traditional algorithm, the improved algorithm has better stability, higher computational efficiency and lower time complexity.

3 Analysis Process

3.1 Constructing the Indicator System of Internet Development Level in Each Province

It is very important to select the evaluation index in the evaluation of the Internet development level in a region. Based on the principles of system, comprehensiveness, effectiveness and data availability, we select indicators to build an index system of Internet development level analysis, a total of 5 indicators are selected. Those are *V1* Number of Internet users (10,000 people), *V2* Number of domain names (10,000), *V3* Internet dial-up users (10,000 households), *V4* Urban broadband access users (10,000 households) and *V5* Rural broadband access users (10,000 households).

3.2 Sample Data Source

According to the evaluation index system of Internet development level constructed above, the samples of 31 provinces and regions in the mainland of China except Hongkong, Taiwan and Macao are selected. The data in this article are mainly from the Chinese Statistical Yearbook 2017.

3.3 Results

In order to remove the influence of the population base, the four variables outside *V2* are replaced by the ratio of the indicator value to the total population of the region. The correlation coefficient matrix of each variable is shown in Fig. 2.

There is no correlation between the four factors selected in the clustering process, which can eliminate the phenomenon that the factors are correlated and the result is distorted.

Then, the contour coefficient method is used to select the optimal K value. When the K value is selected, the data is first normalized to the four factors, and then the K is evaluated. Since the general K is not too large, the traversal K is 3 to 8. Since *K-means* has a certain randomness, it does not converge to the global minimum every time. Therefore, for each k value, it is repeated 30 times, and the contour coefficients are calculated and finally averaged as the final evaluation criterion. The optimal K value is shown in the Fig. 3.

The results show that when K is 3, the average contour coefficient is much larger than other values.

Table 1 shows the center vectors for each category. And the clustering results are shown in Table 2 and Fig. 4.

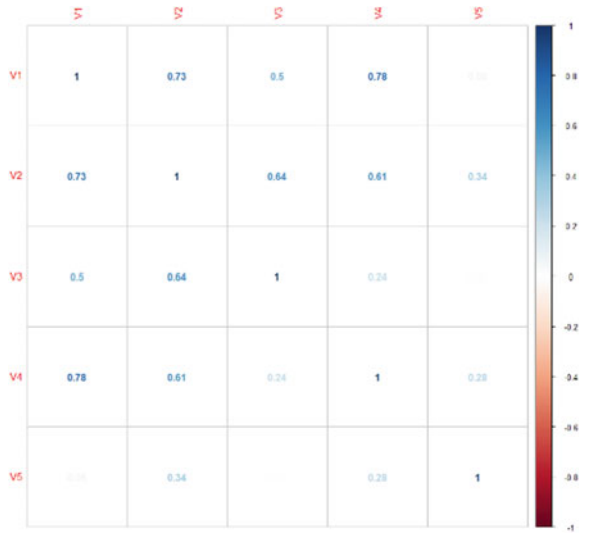

Fig. 2 Correlation coefficient matrix

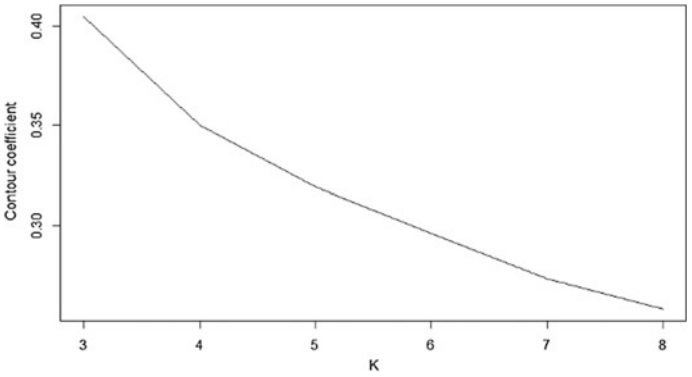

Fig. 3 Contour coefficient and *k* relationship diagram

Table 1 Clustering mean

Category	Internet penetration rate	Number of domain names	Internet dialup user ratio	Urban broadband access users	Rural broadband access users	Quantity
1	0.795	570.633	0.0101	0.2123	0.0648	3
2	0.558	191.101	0.0016	0.1979	0.0652	7
3	0.526	33.664	0.0016	0.1477	0.0378	21

Table 2 Provincial cluster

Area	Category	Area	Category	Area	Category
Beijing	1	Hebei	3	Chongqing	3
Fujian	1	Shanxi	3	Guizhou	3
Guangdong	1	Inner Mongolia	3	Yunnan	3
Shanghai	2	Liaoning	3	Tibet	3
Jiangsu	2	Jilin	3	Shaanxi	3
Zhejiang	2	Heilongjiang	3	Gansu	3
Shandong	2	Anhui	3	Qinghai	3
Henan	2	Jiangxi	3	Ningxia	3
Hunan	2	Hubei	3	Xinjiang	3
Sichuan	2	Guangxi	3		
Tianjin	3	Hainan	3		

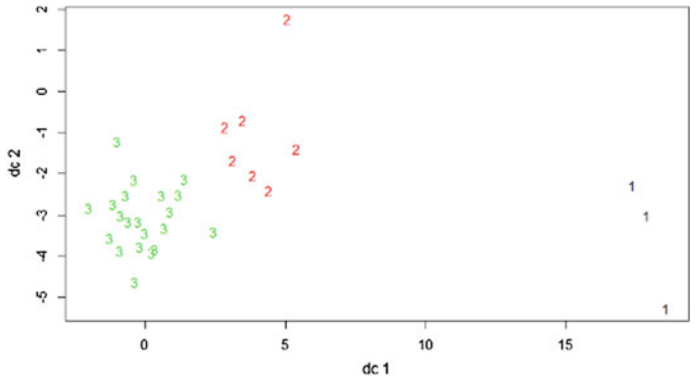

Fig. 4 Two-dimensional clustering map

Figure 4 shows the cluster analysis of each category in a two-dimensional space. It can be seen that the categories 2 and 3 have few differences, but they are very different from category 1.

It can be seen from Table 1 that the Internet penetration rate and domain name number of Category 1 are significantly higher than the other two categories. It can be seen that the overall development level of the Internet in the three regions of Category 1 is a leading level in China. Category 2 represents the level of China's medium Internet development, except that the above two indicators lag behind Category 1. The broadband penetration rate of cities and villages is not behind. It can be seen that among the seven cities in category 2, the popularity of the Internet has reached a leading level, but there is still room for improvement in the commercial application of the Internet. There are 21 provinces in category 3, accounting for two-thirds of the provinces in China. The overall penetration rate of these regions and the popularity

of urban broadband users have reached a medium level. However, the network access in the countryside is very backward, and the commercial application of the Internet is far below the medium level.

According to the specific province categories in Table 2, we find that the provinces in Category 3 contain almost all the western regions, while in Category 1, Beijing, Guangdong and Fujian are the most developed regions in China's domestic economy. It can be seen that the level of development of the Internet also reflects the overall economic level. In addition, Shanghai is the top two cities in China, but in the cluster, it is classified as the second file. The detailed data found that the Internet penetration rate in Shanghai is slightly lower than that in Beijing, but the number of domain names is much lower than Beijing. It can be seen that companies in Shanghai do not focus on the application of the Internet.

4 Discussion

4.1 Lag Development of the Internet in Western China

The backwardness of the economic level in the western region has led to the lag of the development of the Internet. The central cities such as Chongqing have advantages such as strong traditional industrial base, rich population resources and superior geographical position, but they also generally face narrow development space and lack of natural resources, to realize the development of small space, developing the Internet industry is an inevitable choice for the development strategy of the central city of the western city. It is possible to support these cities with better economic bases to fully popularize the Internet and play an overall leading role in the western region. In addition, a major reason for the low rural penetration rate is that there are fewer new types of professional farmers in the western region who have culture, management, and technology. For this reason, cultural cultivation is needed. Only with a large number of new professional farmers who have culture, management and technology, the "Internet +" agriculture will be implemented.

The specific countermeasures are as follows [6]:

- Scientific planning and rational layout
- Well-equipped and powerful science and Technology Park
- School-enterprise cooperation and innovative services
- Policy and system support, necessary environmental and financial security.

4.2 The Internet Industry Development Gap Between Shanghai and Beijing

In response to this problem, we analyze the gap in Internet development between Shanghai and Beijing. In the history of China's Internet development, Shanghai once took the leading position. Before 2010, Shanghai was the capital of online games in China. From the perspective of the current scale of Internet companies, the giants in Beijing are everywhere, whereas, in Shanghai, it is smaller market scale and performed sophisticated. Behind the scale difference is the gap between the Internet talent pools in the two places. The stabilizing and balanced aspect of Shanghai culture has limited more top talents from the top 500 office buildings to the basement startups. There are more idealistic colors in Beijing's urban culture. There are a lot of "Northern drifters" who do not have Beijing registered residence, and they are longing for "Poetry and Distance". Every brightly lit office building in Zhongguancun, Wangjing, and Shangdi is a soulful, imaginative and ambitious soul. Secondly, the Shanghai government's response may act too slow, and the support for the Internet companies is not sufficient in policies. Shanghai has not existed as a large platform for the Internet companies for a long time, so it is difficult to obtain the talent agglomeration effect. Consequently, no talents, no active atmosphere, and no development in this area.

In 2015, the State Council passed the "Internet +" Action Guidance Opinion, which clarified the promotion of "Internet +", promoted entrepreneurial innovation, collaborative manufacturing, modern agriculture, smart energy, inclusive finance, public services, efficient logistics, e-commerce, Convenient transportation, green ecology, artificial intelligence and other key areas of development that can form a new industry model, and identify relevant support measures. Shanghai should keep up with the situation and use policy support to make up for the shortcomings in the Internet industry [7].

The specific countermeasures could be concluded and listed as follow:

- To adopt a top-down design and system implementation related support policies;
- To focus on key areas and cultivate leading enterprises in the Internet economy;
- To develop industrial internet with the help of industrial capabilities and traditional industry advantages;
- To focus on the construction of science and technology innovation center, vigorously implementing industrial basic technology and model innovation;
- Attracting and cultivating industry talents and strengthen industrial intelligence construction.

5 Conclusion

In this paper, the improved *K-means* algorithm is used to cluster the Internet index data of various provinces in China, and all provinces are divided into three categories. It also analyzes the characteristics of each category, which leads to the shortage of China's Internet development, and analyzes the problems of Internet penetration rate in the western rural areas and the underdeveloped Internet industry in Shanghai. Finally, some policy recommendations are given. Further research can provide more detailed data for in-depth analysis of the Internet penetration in rural areas in western China.

References

1. Liu, C. (2016). Analysis of the advantages and countermeasures of building the Internet into an important platform for social science popularization. *Journal of the Party School of Jinan Municipal Committee of the Communist Party of China, 6,* 96–99.
2. Bao, L., & Zhu, X. (2013). The current status of Internet popularization in China and its development countermeasures. *Electronic Production, 5,* 161–162.
3. Sun, Y. (2006). Analysis of the popularity of the Internet in China's population and its imbalance. *Journal of Statistics and Information, 1,* 89–92.
4. Wang, M., Wang, W., & Wan, B. (2018). Analysis of spatial differences and influencing factors of Internet popularization in China. *Statistics & Decision, 34,* 101–104.
5. Xue, G., & Liu, S. (2018). Evaluation of traffic development in China's provinces based on principal component analysis. *Journal of Discrete Mathematical Sciences and Cryptography, 21,* 969–978.
6. Zhao, Y. (2011). Analysis of the status quo and countermeasures of the Internet industry in the central city of western cities-taking Yuzhong District of Chongqing as an example. *Laws and Society, 17,* 179–180.
7. Wu, Z. (2016). Research on improving Internet economic competitiveness in Shanghai. *Science Development, 5,* 60–69.