American Journal of Information Science and Technology
2019; 3(3): 72-79
http://www.sciencepublishinggroup.com/j/ajist
doi: 10.11648/j.ajist.20190303.13
ISSN: 2640-057X (Print); ISSN: 2640-0588 (Online)

Topic Analysis of Microblog About "Didi Taxi" Based on K-means Algorithm

Yonghe Lu*, Xin Xiong

School of Information Management, Sun Yat-sen University, Guangzhou, China

Email address:
luyonghe@mail.sysu.edu.cn (Yonghe Lu), xiongx33@mail2.sysu.esu.cn (Xin Xiong)

*Corresponding author

To cite this article:
Yonghe Lu, Xin Xiong. Topic Analysis of Microblog About "Didi Taxi" Based on K-means Algorithm. *American Journal of Information Science and Technology.* Vol. 3, No. 3, 2019, pp. 72-79. doi: 10.11648/j.ajist.20190303.13

Received: July 30, 2019; **Accepted**: August 16, 2019; **Published**: September 2, 2019

Abstract: In the age of information and digitization, most users publish and obtain real-time information by microblog in social networks. Through effective means, we can accurately discover, organize, and utilize the valuable information hidden behind the massive short texts of social networks. Then we can explore hot topics in microblog, which is conducive to public opinion monitoring and marketing development. In today's society, Didi Taxi has become a necessary choice for many users to travel. This paper applied K-means clustering algorithm to topic analysis of Sina microblog short text on Didi Taxi. We crawled 17226 search results of microblog relevant to the topic of Didi Taxi from April 2019 to June 2019. After a series of data cleaning and data preprocessing steps, we used TF-IDF method to represent 15054 pieces of text data after processing. Through the evaluation of silhouette coefficient, we set the dimension of text 300 and the number of clusters 34 with K-means. Next, we extracted 8 topic clusters from 34 clusters, which include the advantages and disadvantages of Didi Taxi and its development status. Finally, we discussed the results by human check in semantic perspective. Through the topic analysis of microblog, we can understand the public's attitude to Didi Taxi and provide the basis for the management of the government or company in the future.

Keywords: K-means Clustering, Topic Analysis, Microblog Text, Didi Taxi

1. Introduction

In today's information age, microblog has become a major place for people to share their lives, express their thoughts and communicate. By the fourth quarter of 2018, the number of monthly active microblog users had increased to 462 million, an increase of more than 70 million for three consecutive years. The number of vertical fields has been expanded to 60, where 32 fields reached over 10 billion of monthly reading. [1] A large number of microblog users provide a large amount of text information on the platform, and the researchers found great value in it. [2] For example, in terms of economics, companies can learn about users' views on their products, so as to improve the functional requirements of products, then recommend products and services more suitable for corresponding users to achieve precision marketing. In terms of politics, for a program or policy issued by the government, it can effectively predict the future development of society through users' opinions on current national policies.

In recent years, the sharing economy has risen rapidly in the global scale, showing a high speed of growth. As one of the most rapidly developing sharing economic models in China, online taxi-hailing service has been full of opportunities and challenges since its emergence. On the one hand, online taxi-hailing service benefits people's daily life. On the other hand, it still exists the problems of the service quality and the regulatory issues. In the domestic taxi-hailing industry, Didi Taxi almost monopolizes the whole market. Since its launch in 2012, Didi Taxi has been expanding its business scope. At present, it is a one-stop travel platform covering taxi, express, private car and substitute driving with a huge user base. In this paper, topic clustering analysis will be conducted on users' published content about the topic of Didi Taxi on microblog platform, so as to understand the development status of Didi Taxi and public feelings and opinions, and further provide decision-making basis for enterprise strategic management and government social

management.

2. Related Work

2.1. Application of Clustering Algorithm

In the study of short text clustering generated by social media, the classical K-means algorithm is mostly applied, and then the hierarchical [3-5] and density-based [6] clustering ideas are introduced. By constantly adjusting relevant parameters in the process of clustering, the dependence and error of clustering results on empirical setting can be reduced. And it effectively alleviates the problem that k-means is sensitive to the initial clustering center resulting in high volatility of clustering results, thus reducing time complexity and improving clustering efficiency. However, these improved methods are highly dependent on hierarchical clustering, and the uncertainty of results increases. In addition to the improvement of k-means clustering algorithm, some scholars also proposed the discovery of hot topics on microblog based on the improved CURE hierarchical clustering algorithm. The better seed of blog article is taken as the representative point, and the method of adjusting contraction factor is adopted to increase the exclusion of different value points, so that the clustering effect is more accurate. [7] Besides, the improvement of single-pass clustering algorithm combines with the characteristics of network public opinion. Multiple structural coefficients or weighting factors of topic characteristics are introduced into the calculation of topic similarity to effectively improve the error rate and time complexity of the algorithm. [8, 9]

In the context of big data, the open source distributed storage and the framework of Hadoop provide higher fault tolerance and reliability guarantee for processing massive network public opinion data. [10] The K-means distributed topic clustering method based on MapReduce can effectively solve the problem of low efficiency of traditional K-means by efficiently cooperating and processing with multiple computers, which is of great significance in practical application and future development. [11] It can be concluded that current application of K-means algorithm in the text clustering of social public opinion has been widely recognized.

2.2. Research on Didi Taxi

At present, researches related to Didi Taxi are still limited to the empirical analysis of enterprises' economic and business models. Most research methods are questionnaire survey, case analysis, theoretical research and so on. By comparing Didi Taxi and other cases, Wang Jiabao [12] analyzed the business model of enterprises in the sharing economy in China. Taking Didi Taxi as a typical case, Liu Jiangang [13] summarized and analyzed the factors affecting the business model innovation of Internet enterprises through the Grounded Theory. In addition, He Minghua [14] refer to information interpretation and response model to construct a model of consumers' intention of continuous use through questionnaire survey of consumers who have used Didi Taxi platform. As a typical sharing economy service form, how to evaluate its service quality and improve its service quality management level is an important research direction of online taxi-hailing service. Zuo Wenming [15] chose two representative enterprises ----- Didi Taxi and Uber as an example. And he summarized the service quality problems reflected in network public opinions by coding analysis of news headline related to taxi-hailing services. Then the service quality of taxi-hailing is evaluated from the four dimensions of difference, importance, relevance and satisfaction.

From the above discussion, it can be found that relevant researches of Didi Taxi are relatively new and mainly focus on the research fields of market economy, strategic mode, influencing factors, guidelines and so on. At present, from the perspective of social media platform, it is an innovation to analyze the topic of Didi Taxi by using the technology of data mining and text analysis. This paper will introduce the technical methods used in the experiment process in detail in the third part.

3. Methodology

Microblog information is usually in the form of short texts, which is characterized by its sparsity, irregularity and instantaneity. These characteristics make the widely used ordinary text processing methods unsuitable, requiring higher requirements for researchers to process short text information. This paper used the following technologies of microblog text processing to complete topic analysis of microblog content.

3.1. Data Crawling

At present, there are mainly two methods to collect microblog data: web crawler and API interface. [16] By using the free and open API interface of microblog, a series of structured data such as the number of comments, replies, retweets and user information can be obtained. [17] According to the functions and structures, web crawler methods with a wide range of applications can be classified into general web crawler, focused web crawler, incremental web crawler and deep web crawler. [18] There is some web crawler software for data acquisition of large websites now. It provides comprehensive functions and simple operations, which can reduce much work for users to write crawler code. This paper also applies *GooSeeker*, a web crawler tool to obtain data of Didi Taxi topic.

3.2. Data Cleaning and Preprocessing

Since the data directly obtained from microblog platform contains lots of junk data unrelated to the topic, and these junk data will have a very serious impact on the performance of the following clustering. Therefore, we should clean the collected microblog dataset, including deleting duplicate and incorrect data, deleting irrelevant words and emoticons, eliminating useless and advertising content, converting traditional

Chinese characters into simplified ones and so on.

After obtaining the cleaned microblog data, text preprocessing is generally required to enhance data reliability. The operations of preprocessing for Chinese texts usually include word segmentation and stop words removing.

3.2.1. Word Segmentation

Chinese word segmentation plays an important role in the later text analysis. Its algorithms are mainly divided into rule-based and statistical model-based algorithms, or the combination of the two. (1) Among them, the simplest word segmentation technique is rule-based, also known as dictionary-based matching method, which matches each string of a statement with the words in the word list one by one. Segmentation if found, otherwise no segmentation. According to the way of matching segmentation, there are mainly forward, backward and two-way maximum matching method. The time required for Chinese word segmentation based on rules is usually linear and less consuming. But the disadvantage is the method requires users to provide a dictionary. For Chinese word segmentation in different fields, users are often required to extract and sort out different dictionaries by themselves, and the accuracy of word segmentation is lower than that based on the statistical model. (2) The word segmentation based on the statistical model is applied to the Hidden Markov Model (HMM), Conditional Random Field (CRF) and other methods to calculate the probability of the segmentation results. The main idea of it is to count the number of times that adjacent letters appear in different texts. It The higher the number of times, the more likely these linked letters are to be one word. Therefore, we use the frequency of adjacent letters to reflect the reliability of words. When the frequency of these combinations in the corpus is higher than a threshold, we can consider these letters as a word.

The existing Chinese word segmentation technology and tools have been quite mature, such as the ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) segmentation tools developed by Zhang Huaping et al. in Chinese Academy of Sciences, Institute. [19] The word segmentation tool used in this paper is *jieba*, which combines rules and statistical model based with good effect.

3.2.2. Stop Words Removing

The main idea of removing stop words is deleting the words that are difficult to reflect their original meaning or word length of 1, and closely related to the topic or have high frequency. Now there are many stop lists on the network, this paper used the stop list of Harbin Institute of Technology.

3.3. Text Representation

3.3.1. Text Representation Model

Text representation is the process of converting the actual text into a machine-readable data structure. At present, the common models mainly include Boolean model, probability model and vector space model. Among them, VSM and its extended model are most widely used. In the field of social network research, the text representation models can be divided into vector space model, latent semantic analysis model and implicit topic analysis model. Based on the shortcomings of traditional VSM and the characteristics of forum data, Zhang Haidong [20] proposed a multi-vector dimension representation method including 4 sub-vectors of time, place, person and event, which is more accurate than traditional VSM model. Mi Wenli [21] used probabilistic potential semantic analysis (PLSA) method for topic modeling of microblog data to find hot topics, which effectively solved the problem that K means algorithm was sensitive to the clustering center. Xu [22] proposed a topic recognition method for network sensitive information based on the weighted Latent Dirichlet Allocation (LDA) model. Experiments show that this method can effectively improve the quantity and quality of topic recognition of sensitive information.

Considering the characteristics of microblog and non-supervision of clustering analysis, this paper chooses the vector space model for text representation. The main idea of this model is that each document can be mapped to a point in the vector space of a set of normal orthogonal vectors. [23]

3.3.2. Feature Selection Model

After preprocessing and representation, the text information still belongs to the vector matrix with high dimensions and high sparsity, which increases the burden on the computing, learning and training process of the computer. In order to further reduce dimension, we need to select text features. The research on feature selection started in the 1960s. [24] The key of it is to find the optimal feature subset in the solution space containing all feature subsets and select the most representative feature combination under the premise of the minimum time cost. The commonly used feature selection algorithms include Document Frequency (DF), Term Frequency-Inverse Document Frequency (TF-IDF), Information Gain (IG) and CHI, etc.

(1) Document Frequency refers to the number of documents containing a certain feature item in the training document, which is the simplest evaluation function. This method has the advantages of simple algorithm, low complexity, good effect in practical application, and can be applied to large-scale data sets. The disadvantage is that the words below the threshold are removed from the original space vector. Although it can effectively reduce the dimension of feature space, it will also filter out some feature words (such as special words) with low document frequency, which may contain important information, thus affecting the classification judgment.

(2) Term Frequency-Inverse Document Frequency considers the frequency TF and inverse frequency IDF of the document fully. It obtained good results in the calculation of feature weight function. The application of feature weight calculation to feature extraction is a common feature extraction method, which has been widely used in the field

of text classification. The main idea is that a word or phrase that appears frequently in one article but rarely in other classes is considered important to this article. [25] TF refers to the frequency of a word or phrase appearing in an article; IDF represents the reciprocal of the number of documents containing a word or phrase. The formula is as follows.

$$W_i = TF_i * IDF_i = \frac{N_{i,j}}{\sum_k N_{k,j}} * log\frac{D}{card(\{j|i \in d_j\})} \quad (1)$$

Among it, $N_{i,j}$ represents the number of times the word i appears in document j, $\sum$ is the total number of terms that appear in document j, D represents the total number of documents in the corpus, $card(\{j|i \in d_j\})$ represents the number of documents containing the term j.

(3) Information Gain is an assessment method based on information entropy, which defines the difference of information entropy between the occurrence and non-occurrence of feature words in one document. The larger the information gain value of a feature word, the more important the information is. [26]

(4) In CHI-squared test, the characteristic words and text categories are subject to χ^2 distribution. χ^2 can measure the independence between features and classes. If feature T and classes are independent of each other, the χ^2 value of feature T is 0. The higher its value is, the less independent it is between them.

Comparing the above feature selection methods, this paper chooses the TF-IDF method, which is the most common and has the best weight calculation effect, as the feature extraction algorithm for microblog topic analysis.

3.4. Text Clustering

3.4.1. Clustering Algorithm

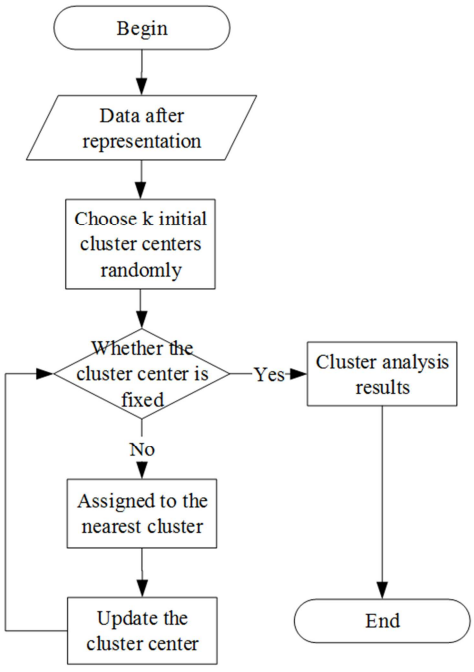

Figure 1. *Process of K-means clustering analysis.*

The core of microblog topic analysis is the process of text clustering. Since text clustering does not require manual identification of documents in advance, it is an unsupervised machine learning method. According to the widely accepted clustering hypothesis, different types of documents have low similarity, while similar documents have high similarity. Different clustering algorithms have their specific effectiveness and limitations. Traditional text clustering methods include partition-based clustering algorithms, such as K-means; hierarchical-based clustering algorithm, such as CURE, BIRCH; grid-based clustering algorithm, such as BANG, CLQUE; density-based clustering algorithm, such as DBSCAN, DENCLUE.

This paper applied the simple and efficient K-means algorithm, which has a low time complexity and can operate on large-scale short texts. The specific flow chart of the algorithm is as figure 1.

3.4.2. Clustering Evaluation Index

Cluster evaluation is to estimate the feasibility of clustering on data sets and the quality of the clustering results, which mainly consist of estimating clustering trend, determining the number of clusters in data sets and determining clustering quality. For K-means, it is of great significance to select the appropriate cluster number for subsequent text clustering results. At the same time, it is necessary to evaluate the effect and quality of clustering properly after using the clustering method to the data set. The Silhouette Coefficient used in this paper combines two factors of cohesion and separation to evaluate the impact of different running modes on the clustering results, so as to select the optimal solution. The calculation formula is as follows.

$$S_i = \frac{b_i - a_i}{max\{a_i, b_i\}} \quad (2)$$

Among it, the value of a_i reflect the degree of cohesion, which is the average distance from the object i to all other objects in the cluster. The value of b_i reflect the degree of separation, which is the object i and any cluster that doesn't include the object. S_i is the silhouette coefficient of object i, whose value is between negative 1 and 1. The higher the value is, the better the cohesion and separation is.

3.4.3. Manual Check

Since text clustering is an unsupervised training process, in order to check the clustering effect, we check the cluster results manually. That is to say, we selected a number of microblog data from each cluster randomly. Then we determined whether this microblog is similar to the topic of the cluster by manual semantic analysis, especially those don't contain the key words of this cluster.

4. Experiment

4.1. Data Collection and Cleaning

In order to dig out the latest public opinion orientation of Chinese people on the topic of Didi Taxi, we conducted topic

analysis of the recent relevant microblog content according to the methodology and process proposed in chapter 3. In this paper, we used the crawler software *GooSeeker* in the advanced search page of Sina microblog. The keywords were set "Didi, Taxi"; the microblog type was set "original" which can directly reflect the personal feelings of users; the time restriction was three months from April 1, 2019 to June 30, 2019; the published location was no limitation. Finally, we obtained 17,226 pieces of microblog data in Excel format. The data content mainly includes user name, microblog content, publication time, etc.

Having collected microblog data on Didi Taxi, we carried out data cleaning on microblog content. As the data collected by the web crawler may be misplaced, it is necessary to screen and check whether the microblog information is within the limited range and whether the data format conforms to standards, and then delete and modify the incorrect data information. Users usually publish original posts and share information with some emoticons and words in specific formats, but these have nothing to do with the content. It will increase the vector dimension later and thus affect the accuracy of the analysis results. Therefore, words and emoticons unrelated to the content of microblog should be deleted. The main irrelevant formats for microblogs are shown in table 1 below. And the irrelevant information is cleared through regular expression and the replace and delete function of Excel. In addition, some data which contain search keywords but have advertising words such as "red packet", "dididi" and "registration and activation", should also be deleted. After the above data cleaning steps, some data with empty should be deleted to facilitate the following data preprocessing.

Table 1. *Irrelevant formats that microblogs mainly contain.*

I shared the @<username> post	published a blog post
http://<page link>	[link address]
(share from <website name>)	(shared from @<username>)
@<username>	#<topic>#
second shot video	[<emotion>]
published the headline post	|<link title>...
... full text c	read the full post

4.2. Text Preprocessing and Representation

In order to prevent Chinese garble and improve data reliability, the cleaned Excel data table needs to be converted into txt text encoded in UTF-8. The data preprocessing usually includes two steps: word segmentation and stop words removing. [27] This paper calls the *jiaba* function module of Python, imports the user dictionary added manually about Didi Taxi to improve the accuracy of segmentation. After word segmentation, stop words were removed by using the stop list of Harbin Institute of Technology. We deleted the words that are difficult to reflect their original meaning or word length of 1, and closely related to the topic or have high frequency. Finally, 15054 pieces of pre-processed data were obtained.

After data preprocessing, the text needs to be expressed in a form that can be understood by the computer. In this paper, the Vector Space Model (VSM) is adopted as the text representation model and the document frequency method TF-IDF is adopted to select feature words. Finally, each microblog data is represented as a vector collection related to extracted feature words. And multiple topic clusters are formed by calculating the distance similarity of different microblog text vectors in the space.

4.3. Text Clustering Based on K-means

For K-means clustering of data sets after representation, we need to set appropriate text dimensions and the number of clusters. We used Silhouette Coefficient to determine them in this experiment. We set the range of clusters 2-48 and text dimensions 300-500, and calculated the corresponding coefficient value in figure 2. It can be seen from part 3 that the larger the silhouette coefficient value is, the better the clustering effect will be. So we choose the text dimension of 300 and the number of clusters of 34 due to the relatively high coefficient value and the clustering effect is relatively good. Then the K-means clustering results obtained after the experiment are shown in the following table 2.

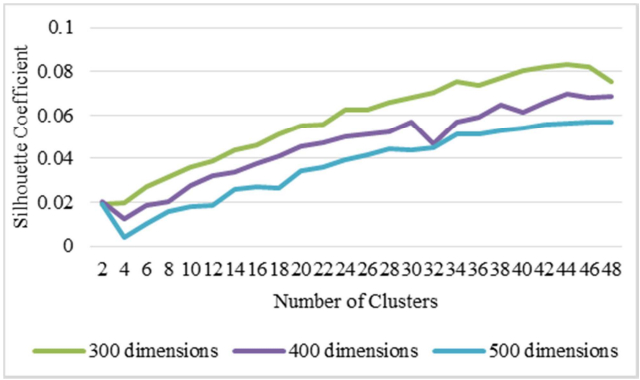

Figure 2. *Silhouette Coefficient of different dimensions and clusters.*

Table 2. *K-means clustering results.*

Cluster	Amount of data	Cluster	Amount of data
0	97 (1%)	17	113 (1%)
1	291 (2%)	18	692 (5%)
2	163 (1%)	19	140 (1%)
3	615 (4%)	20	140 (1%)
4	173 (1%)	21	208 (1%)
5	434 (3%)	22	293 (2%)
6	275 (2%)	23	237 (2%)
7	306 (2%)	24	226 (2%)
8	204 (1%)	25	226 (2%)
9	231 (2%)	26	309 (2%)
10	323 (2%)	27	402 (3%)
11	251 (2%)	28	234 (2%)
12	283 (2%)	29	354 (2%)
13	56 (0%)	30	485 (3%)
14	294 (2%)	31	5515 (37%)
15	257 (2%)	32	195 (1%)
16	421 (3%)	33	435 (3%)

4.4. Topic Analysis

We selected 8 clusters with certain significance from the 34 clusters of K-means clustering results, and listed the top 20 keywords in these 8 clusters, then extracted the topic meaning of the cluster from the keywords.

Table 3. *Top 20 keywords of 8 clusters.*

Cluster	Top 20 keywords
Cluster 31	drivers, in the car, instructors, actually, not access to, get on, feel, taxi, like, in the morning, go out, find, take, bus, in the evening, get off, minutes, one car, place, on the road
Cluster 16	bus, public transportation, not access to, actually, minutes, go out, subway, be late, drivers, take a taxi, take, go back, miss, half an hour, taxi, want, in the morning, find, choose
Cluster 30	Airport, hotel, train station, train, illegal vehicles, taxi, plane, high-speed railway, drivers, Shanghai, reach, coach, in the evening, bus, eventually, instructors, not access to, actually, take a taxi
Cluster 27	take a taxi, taxi, substitute driving, drivers, fare, not access to, software, actually, cause, cancel, order, every time, instructors, nicely, girls, hitchhiking, in the evening, travel, in the morning, feel
Cluster 1	make a phone call, customer service, drivers, cancel, phone, order, make an appointment, instructors, actually, receive a order, make a complaint, place, take a car, yesterday, time, not access to, taxi, get on, find, get off
Cluster 3	online taxi-hailing service, platform, travel, service, cooperation, car, business, Japan, market, driving, China, publish, drivers, taxi, officially, industry, launch, provide, enterprise, city
Cluster 18	hitchhiking, launch, return back, car, business, hope, offline, person in charge, rectify and reform, DiDa Taxi, platform, market, travel, actually, software, car owner, substitute driving, drivers, Amap, event
Cluster 15	grey level, test, open up, hitchhiking, offline, rectify and reform, product, users, respond to, technology, result in, deny, cause, price, make an appointment, in the car, information, go for a ride, recently, launch

Cluster 31: People use Didi Taxi in their daily life, which is often reflected in commuting. In this process, people enjoy the great convenience brought by the taxi-hailing service, and also complain about some problems at the same time.

Cluster 16: Didi Taxi is often seen as an alternative to public transportation such as buses and subways. Didi Taxi saves time to some extent, but it also reduces user satisfaction in poor road conditions. People may still encounter unexpected situations such as missing or being lat.

Cluster 30: People usually choose Didi Taxi app when they are traveling to airports, railway stations, high-speed railway stations, bus stations, hotels and other destinations, which can not only save the time spent on public transportation, but also make appointments in advance through mobile devices at designated time, so as to achieve fast and convenient transportation.

Cluster 27: In the process of taking a taxi, users are likely to pay high fares due to the driver's detour, unreasonable driving and other behaviors. Besides, Didi's substitute driving service can easily produce poor user experience and even cancellations of order.

Cluster 1: People need to call the driver to confirm the itinerary when taking Didi Taxi. In case of emergency, users will complain to customer service to express their dissatisfaction. For example, when users make an appointment of a taxi on Didi, the drivers may accept orders but refuse to pick up passengers due to the distance or other reasons; the pre-estimated price before getting on the car is different from the actual price when getting off the car and even some drivers raise the prices.

Cluster 3: Didi Taxi enterprise have strengthened platform cooperation and entered the Japanese taxi-hailing industry. Liu Qing, President of Didi Taxi, said: "Didi believes that innovation in AI technology can bring new development to the taxi industry and public transport industry. We look forward to working extensively with our transport partners to contribute to the development of smart cities in Japan and Asia." In July 2018, Didi Taxi announced a joint venture with Softbank to launch Didi taxi-hailing service in Tokyo, Japan. Didi in Japan will introduce China's advanced data platform to help local taxi companies improve operating efficiency, improve user satisfaction and expand their user base. While visiting Japan, many Chinese tourists have also experienced the convenient online taxi-hailing service.

Cluster 18: Since Didi hitchhiking was frequently involved in vicious incidents before, it has been removed from Didi in August 2018. Internet users were divided on the issue. The proponent believes that the abolition of hitchhiking service can effectively rectify the current standard and safety of online taxi-hailing services, while the opponent believes that hitchhiking service is still needed in the market. Hitchhiking is a more economical way to travel for people who find it too expensive to take a taxi and inconvenient to use public transportation. Dida Taxi has been around for years with fewer drivers and better billing, so it is uncompetitive with Didi. But due to the withdrawal of Didi hitchhiking, some users who still have the demand began to choose Dida hitchhiking to realize economic travel.

Cluster 15: Didi Taxi denied that hitchhiking was open to grayscale testing. Some netizens reported that Didi hitchhiking recently opened its grayscale test in a small range, which triggered a wave of public opinion discussion upsurge. However, Zhang Rui, the head of Didi hitchhiking's business unit, issued an open letter to deny the rumor. He said that he would continue to make rectification of Didi hitchhiking but denied that it would go online soon.

4.5. Manual Check

In order to test the effect of k-means clustering, we should manually check the clusters obtained from the above clustering. We selected several pieces of data from each cluster randomly, read the microblog content in semantic perspectives, and judge whether it contains high-frequency keywords of this cluster and whether it is close to the cluster

topic. In this paper, 50 pieces of data were randomly selected from 8 topic clusters for manual semantic analysis. Then we counted the number of microblogs that contain and don't contain the corresponding cluster keywords, and those close to and deviated from the corresponding cluster topics. The results are shown in the following table 4, where valid data refers to whatever whether it contains cluster keywords, the text semantics are close to the cluster topic. Noise data refers to whatever whether it contains cluster keywords, the text semantics are deviated from the cluster topic.

Table 4. *Manual check result of 8 clusters.*

Cluster	keyword close to topic	no keyword close to topic	keyword deviated from topic	no keyword deviated from topic	Valid data	Noise data
31	64%	24%	6%	6%	88%	12%
16	96%	0%	4%	0%	96%	4%
30	88%	0%	12%	0%	88%	12%
27	56%	0%	44%	0%	56%	44%
1	80%	0%	18%	2%	80%	20%
3	62%	8%	28%	2%	70%	30%
18	84%	0%	16%	0%	84%	16%

Through the manual sample test of semantic analysis, it is found that except cluster 31, the proportion of data containing keywords and close to the topic is much larger than that without keywords and close to the topic, which indicates that the keywords have a strong ability to describe the meaning of the cluster topic. On the whole, most of the data in the 8 clusters are close to the topic of the cluster. Except for the valid data rate of 56% of cluster 27, the rate of all the other clusters is over 70%. And the highest rate of cluster 16 and 15 is 96%, which indicates a good clustering effect.

5. Conclusion

5.1. Research Conclusions

Didi Taxi, an Internet & transportation mode, has been accepted by most people and has brought great changes to people's life. Didi Corporation has also begun to focus on cooperation with the outside world. It has established a joint venture with Japan's Softbank to launch Didi Taxi service in Tokyo. While enjoying the convenience Didi brings to life, some prominent problems have been found. For example, drivers take detour or drive irrationally, which results in price increase; drivers accept orders but refuse to pick up passengers due to personal reasons; some unregistered cars in Didi Taxi may exist security threat. Didi Taxi should strengthen customer service management to timely solve the problems of users. As for the price, the profit of Didi is to draw a percentage of the customers' order after completing, so it needs to improve the transparency to improve the customer satisfaction. The government can also strengthen the audit and management of taxis and express cars. It can help enterprises avoid users' moral risks. crack down on inaction, promote civilized online taxi-hailing atmosphere, and shoulder their own responsibilities.

For Didi hitchhiking that has been removed from the market, it is still one of the problems that the enterprise needs to solve in the future development. At present, Didi denied the public opinion that hitchhiking is open to grayscale test and launch. The head of hitchhiking department said that he would continue to rectify and reform hitchhiking business. The government should also strengthen the supervision and management of taxi-hailing, crack down on illegal behaviors and effectively improve the environment of this industry.

5.2. Limitations and Prospects

The research limitation of this paper mainly lies in that there is no targeted algorithm improvement on the problems such as sparse features of short articles and reliability of data results. And the data set of the experiment has certain limitations, so it is not possible to conduct topic analysis of Didi Taxi from a longer time dimension.

In the future research, we can consider to supplement the experiments of other clustering algorithms in topic analysis of microblog, and explore the optimal solution of short text clustering. Besides, the improvement of K-means clustering algorithm for different documents is applied to extending current enterprise strategy analysis. Traditional clustering algorithms are mostly based on spatial distance calculation, which can be combined with semantic web or deep learning to realize distance calculation based on semantic.

Acknowledgements

This research was supported by National Natural Science Foundation of China (Grant No. 71373291). This work also was supported by Science and Technology Planning Project of Guangdong Province, China (Grant No. 2015A030401037).

References

[1] Sina microblog data center (2019). "2018 microblog user development report," https://data.weibo.com/report/reportDetail?id=433.

[2] YUAN Bo. "Microblog topic mining based on relation network," in Harbin Institute of Technology, 2014, pp. 1-3.

[3] LU Rong, XIANG Liang, LIU Mingrong, YANG Qing (2012). Discovering News Topics from Microblogs Based on Hidden Topics Analysis and Text Clustering. Pattern Recognition and Artificial Intelligence, 25 (3): 382-387.

[4] MA Wenwen, WEI Wenhan, DENG Yigui (2014). Micro-blog topic detection method based on Latent Semantic Analysis. Computer Engineering and Applications, 50 (1): 96-100.

[5] WANG Xuren, LI Na, HE Famei, WANG Yanli, SONG Bei (2014). Research and Implementation of Desktop Search Engine Based on Tika and Lucene. Journal of The China Society for Scientific and Technical Information, 33 (5): 530-537.

[6] DING Ruoyao. "Research on Internet topic detection and tracking based on blog," in Beijing Jiaotong University, 2011, pp. 27-30.

[7] YANG Changchun, ZHOU Meng, YE Shiren, XU Xiaosong (2013). An Improved Hot Topic Detection Method for Microblog Based On CURE Algorithm. Computer Simulation, 30 (11): 383-387.

[8] GESANG Duoji, QIAO Shaojie, HAN Nan, ZHANG Xiaosong, YANG Yan, YUAN Changan, et al. (2015). An Internet Public Opinion Hotspot Detection Algorithm Based on Single-Pass. Journal of University of Electronic Science and Technology of China, 4 (44): 599-600.

[9] FANG Xingxing, LV Yongqiang (2014). Discovering the Topic of Network Public Opinion Based on Improved Single-pass Algorithm. Computer and Digital Engineering, (7): 1233-1237.

[10] ZHANG Meng. "Research on LDA short text classification algorithm based on Hadoop platform," in Tianjing University of Finance and Economics, 2016, pp. 24-27.

[11] Wang, F., Liu II, P., & Zhu III, Z. (2018). Hadoop-based analysis model of network public opinion and its implementation. In Third International Workshop on Pattern Recognition. (Vol. 10828, p. 108281H). International Society for Optics and Photonics.

[12] WANG Jia-bao, XUE Man, DUN Shuai (2017). Business Model of Sharing Economy in Chinese Context based on Comparative Study of Multiple Cases. Commercial Research, (09): 21-27.

[13] Liu Jiangang, Zhang Meijuan, Chen Changjie, Zhao Lingling (2017). Influence Factors on Business Model Innovation of Internet Platform Enterprise——A Case Study of DiDi Based on Ground Theory. Forum on Science and Technology in China, (06): 185-192.

[14] HE Minghua, LIANG Xiaobei (2018). Effect of Platform's and Service Providers' Reputation on Consumers' Continuous Intention to Use a Sharing Option in the Context of Sharing Economy——An Empirical Study Based on Didi Chuxing Platform. Reform of the Economic System, (02): 85-92.

[15] ZUO Wenming, ZHU Wenfeng (2018). Research on Service Quality Management of Online Car-hailing Based on SERVQUAL in Sharing Economy: Case Study of Didichuxing and Uber. Journal of Management Case Studies, 11 (04): 349-367.

[16] Guo, K., Shi, L., Ye, W., & Li, X. (2014). A survey of internet public opinion mining. In 2014 IEEE International Conference on Progress in Informatics and Computing (pp. 173-179). IEEE.

[17] ZHANG Hua. "The research of micro-blog public opinion predictive model based on optimized BP neural network," in Central China Normal University, 2014, pp. 12-14.

[18] TANG Luyang. "Research on web data acquisition and management for online public opinion analysis," in University of Electronic Science and Technology of China, 2017, pp. 19-23.

[19] Zhang, H. P., Yu, H. K., Xiong, D. Y., & Liu, Q. (2003). HHMM-based Chinese lexical analyzer ICTCLAS. In Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17 (pp. 184-187). Association for Computational Linguistics.

[20] ZHANG Haidong. "Research on hot topic identification and trend prediction based on BBS" in Shanghai Normal University, 2014, pp. 19-27.

[21] MI Wenli, SUN Yuexin (2014). Microblog Hot Topics Discovery Method Based on Probabilistic Topic Model. Computer Systems & Applications, (8): 163-167.

[22] Xu, G., Wu, X., Yao, H., Li, F., & Yu, Z. (2019). Research on Topic Recognition of Network Sensitive Information Based on SW-LDA Model. IEEE Access, 7, 21527-21538.

[23] SU Yu, ZHENG Cheng, MA Zhongjie (2011). The Improvement of VSM Model Based on Semantics. Computer Applications and Software, 28 (08): 158-161.

[24] Lewis, P. (1962). The characteristic selection problem in recognition systems. IRE Transactions on information theory, 8 (2), 171-178.

[25] Zhang, Y. (2013). Overview of keyword extraction in single document. Scientific Journal of Information Engineering, 3 (1).

[26] Hu, Y., & Loizou, P. C. (2004). Speech enhancement based on wavelet thresholding the multitaper spectrum. IEEE Transactions on Speech and Audio Processing, 12 (1), 59-67.

[27] Han, J. (2005). Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc.