The Analysis of China's Integrity Situation Based on Big Data

Wangdong Jiang[1], Taian Yang[1, *], Guang Sun[1, 3], Yucai Li[1], Yixuan Tang[2], Hongzhang Lv[1] and Wenqian Xiang[1]

Abstract: In order to study deeply the prominent problems faced by China's clean government work, and put forward effective coping strategies, this article analyzes the network information of anti-corruption related news events, which is based on big data technology. In this study, we take the news report from the website of the Communist Party of China (CPC) Central Commission for Discipline Inspection (CCDI) as the source of data. Firstly, the obtained text data is converted to word segmentation and stop words under preprocessing, and then the pre-processed data is improved by vectorization and text clustering, finally, after text clustering, the key words of clean government work is derived from visualization analysis. According to the results of this study, it shows that China's clean government work should focus on 'the four forms of decadence' issue, and related departments must strictly crack down five categories of phenomena, such as "illegal payment of subsidies or benefits, illegal delivery of gifts and cash gift, illegal use of official vehicles, banquets using public funds, extravagant wedding ceremonies and funeral". The results of this study are consistent with the official data released by the CCDI's website, which also suggests that the method is feasible and effective.

Keywords: Big data, anti-corruption, text clustering, visualization.

1 Introduction

The topic of Integrity and anti-corruption has always been one of hot issues in China [Dai (2010)], especially after the 18th CPC National Congress, 'Chinese-style anti-corruption', such as the two campaigns-'Fox Hunt' and 'Sky Net', has raised strong repercussions both at home and abroad [Blancke (2018)]. Since the Party's eight-point frugality code carried out, and till the end of December 2018, from the quantitative structure, there were 241,407 mental problems investigated, which violated the Party's eight-point frugality code, and a total of 385,192 people dealt, besides the highest monthly number of people had reached 13,411 and in average it was 5,669; From the position structure, the proportion of officials at the grassroots accounted for nearly 90% among all officials, once up to 97%. It shows that the trend of anti-corruption is constantly deepening to the grassroots, and gradually ensure that net more 'the tiger, fox and flies'; from the

[1] School of Information Management and Technology, Institute of Big Data, Hunan University of Finance and Economics, Changsha, 410205, China.

[2] Housheng School of International Education, Hunan University of Finance and Economics, Changsha, 410205, China.

[3] School of Engineering, The University of Alabama, Tuscaloosa, 35487, USA.

[*] Corresponding Author: Taian Yang. Email: enron_yta@163.com.

geographical structure, the provinces, autonomous regions and municipalities and the Xinjiang Production and Construction Corps, central and state organs, central enterprises and central financial enterprises do not leave any dead ends. Obviously, in the aspect of investigation and punishment, the anti-graft campaign is large, deep, and wide-ranging, and it is easy to see that the task of China's integrity and anti-corruption is more significant than Mount Tai in the imminently and complexly current situation.

Judging from the published anti-corruption data, it may not observe more regulations through the data from one year or one region, but with more and more data published from different years and regions, the characteristics of group behavior will present certain orders, correlations, and also more regularity are going to appear thereupon [He (2000)]. In fact, there exist an inextricable link among all corruption phenomena, and those corrupt behaviors do have common rules to follow. Once the rules can be found, to some extent it means finding a sword against corruption. However, we must analyze the data of integrity and anti-corruption in depth, which aims to grasp the regularity. With continuously moving forward anti-corruption work, the anti-corruption data has been increasingly accumulated. According to these data, we can deeply explore the outstanding problems under anti-corruption work and give targeted suggestion in the light of the problems.

Nowadays, along with the arrival of the big data era [McAfee, Brynjolfsson, Davenport et al. (2012)], it is limited for the traditional data processing technology to grasp the data's whole characteristics and developing trend, let alone discovering the regularity of data. Therefore, traditional technology can no longer meet the requirements of the current era. For the purpose of solving the defects of traditional technology, big data technology emerges at the historic moment and has been applied widely in many fields. Owing to its advancement, big data technology can be used to do researches on news network information regarding China's clean government and anti-corruption.

In the field of anti-corruption, there are few precedents for the use of big data, and its scope of research includes the integrity supervision mechanism, clean government system, internet anti-corruption, early warning mode of clean government, the value of clean government, the measurement of anti-corruption and other more fields. Nowadays, big data technology is still in the ascendant, and it has positive significance for broadening the research in the field of integrity and anti-corruption by related network information.

In the present China's society, the anti-corruption campaign is in full swing, and people around the world are paying increasing attention to this topic through. With the rapid development of the Internet, the Internet has become the main carrier of information dissemination. However, the information obtained from the Internet has the characteristics of huge amount, complex structure, difficulty in extracting semantic information, etc. Therefore, this article offers a proposal to analyze the related network information of anti-corruption on the basis of big data technology, study its internal rules and propose effective strategies, which plays a practically vital role for the further development of China's anti-corruption work. Meanwhile, it provides a feasible strategy for academic colleagues to study the issue of integrity and corruption.

2 Data source and research method

In view of the authoritativeness and representativeness of the data published by the CCDI's website, this article adopts the data from the "supervised exposure" column on the website, which reports the news of the mental problems of the eight central regulations as the data source. The CCDI's website is an authoritative publishing platform in which its central task is the building of governance and anti-graft fight, supported by the CCDI and the discipline inspection and supervision organs at all levels. On September 2, 2013 the website was officially launched. Since then, it has been dedicated to collecting and reporting typical news incidents of corruption across the country. With regard to the column 'supervised exposure', it has covered the quantities of news exceeded 1,000.

Although the news on the CCDI's website is numerous and comprehensive, it exists in the form of web links which cannot be directly used as experimental data. Therefore, it is necessary to collect the text contents of each link and save them in an easy-to-handle format. Under this concept, this article takes advantage of the web crawler technology to gather the content of each news link and completes the collection of news events reported by the "Supervised Exposure" column of the CCDI's website through the Octopus Collector. At the end of December 2018, there were a total of 9260 items effectively collected, and then according to the website's monthly report data on the Party's eight-point frugality code, we made a summary of the monthly report data from January 2014 to December 2018.

In this paper, texat clustering and data visualization are adopted as main research method, and the two methods are utilized for achieving the purpose of in-depth study of the outstanding problems of China's integrity and anti-corruption. The research process framework is divided into data acquisition layer, cluster analysis layer and visual display layer, specifically speaking, the data acquisition layer is used for collecting, cleansing and storing data, while vectoring and clustering the stored data is dealt with the cluster analysis layer, besides the visual display layer is completed by means of visual tools such as Gephi. For the detailed data processing flow, please refer to Fig. 1.

Clustering refers to the process of dividing the data set composed of physical objects or abstract objects into several classes or clusters, so that data objects belonging to the same class or cluster can be as similar as possible, while make different classes or clusters as different as possible [Jain, Murty and Flynn (1999)]. Cluster analysis helps to analyze the information, which is hard to obtain directly from the data, in particular, it is suitable for mining trends, regularity and other characteristics in the data. In addition, as an unsupervised machine learning algorithm, cluster analysis does not require training of the machine and manual classification of the preprocessed data, which effectively avoid the errors reliable on classification by experience and expertise, and then it also gets rid of the subjective limitations. Therefore, cluster analysis makes the data analysis results more objective and accurate. Text clustering is an effective text mining method based on clustering technology, and also a vital research direction in the field of text mining and information retrieval [Biswas, Weinberg and Fisher (1998)].

Text clustering in this research is performed using the K-means algorithm, which was proposed by Steinhaus and used and named first by MacQueen in 1967 [MacQueen

(1967)]. Because of its simple implementation and low time cost, the algorithm is widely utilized by people in text mining, data analysis and other fields.

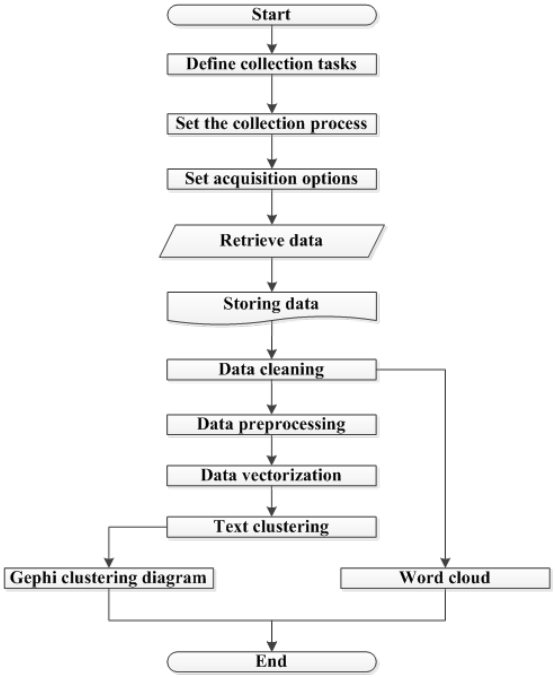

Figure 1: Data processing flow chart

Text clustering can be classified into two major processes: text vectorization representation and clustering. The former one includes the preprocessing of text segmentation and removing stop words, and then serialize and vectorize the preprocessed data. Besides, the latter involves calculating cosine distance, constructing a similarity matrix, and then completes the cluster analysis.

After the era of data visualization [Friendly (2008)], scientific visualization [Defanti and Brown (1991)], and information visualization [Card (1999)], the analysis and display of data has been gradually moving toward the era of big data visualization [Keim, Qu and Ma (2013)]. Big data visualization makes use of advanced visualization methods and human-computer interaction technologies to conduct information mining in any forms that humans can perceive, enabling humans to gain insight into large-scale complex data sets [Chen (2008)]. Big data is regarded as the oil of future development [Hirsch (2013)], in pace with data explosion growth and data opening [Paine (2015)], the demand for visualizing data is becoming increasingly intense. For displaying the characteristics of big data era more intuitively, it is necessary to further study the visualization of big data.

The visualization tools selected in this article cover Gephi and Python-based Wordcloud. As an open source, cross-platform JVM-based complex network visualization tool, Gephi can be used to explore data analysis and flexibly manipulate, display and diversify data, which is to meet the users different needs [Jun (2014)]. Meanwhile, Gephi has powerful view

indication and dynamic analysis, which can effectively guide users to observe and discover the relationship between nodes [Bastian, Heymann and Jacomy (2009)], and due to this feature, it can be served to analyze complex and changeable anti-corruption information.

3 Data process

3.1 Data collection

This paper uses the web crawler software Octopus Collector to collect data. When using the software to collect data, Firstly, you should create a new collection task in the software, type the name of the collection task and the network link address; Next, designing the collection process, setting the collection options, and collecting the title, source, time and body of the news event. Finally, you only need to follow the process to execute your own set of plans, and then you can obtain the data. A total of 14171 news items were collected in this study. From September 2013 to December 2018, there were the monthly report data covered 64 months, which was related on the Party's eight-point frugality code. Besides, the data was stored in '.xls' format.

The data collection process of the Octopus Collector is shown in Fig. 2.

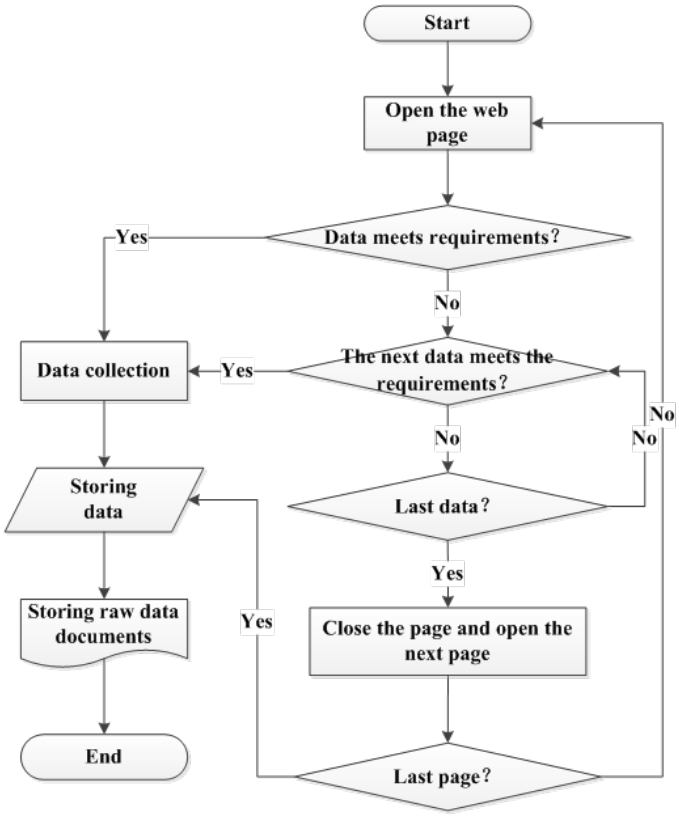

Figure 2: Octopus collector task flow chart

3.2 Data cleaning

From the collected data, manually delete the data in the text that are inconsistent with the anti-corruption theme, and obtain 9260 valid news data; due to the data reporting format of the eight monthly reporting system was officially confirmed in January 2014 and is still in use today, so the data before January 2014 was deleted and 60 months of valid data was obtained. The purpose of data cleaning is to improve the quality of the text and prepare for data pre-processing, meanwhile, the integrity of data will be effectively boost.

3.3 Data preprocessing

Data preprocessing includes word segmentation and stop word processing. The text should be preprocessed before data processing, which means randomly extracting some documents for word segmentation. For the fact that the original text contains specific topics, whereas the computer only depends on the set algorithm program to perform word segmentation, so it will result in a situation where the word segmentation is not effective. Therefore, according to the pre-processing result, the user needs to artificially add the unsatisfactory word to the user-defined dictionary, thereby improving the accuracy and validity of the word segmentation, which facilitate reducing the error.

Since the content of collected news data is basically all Chinese, the word segmentation process can adopt python to call the jieba word segmentation package to deal with the data. Jieba is a Chinese word segmentation module developed by domestic programmers through Python. It is mostly used in text mining and Chinese word segmentation in search engines, and the word-segmentation dealt by jieba has the characteristic of high precision and fast speed. The jieba word segmentation method apply dynamic programming to find the maximum segmentation combination based on word frequency [Sun (2019)], via looking up the maximum probability approach. For the unregistered words, the HMM model based on Chinese character abilities is adopted, and the Viterbi algorithm is used for calculation.

In order to better perform text clustering, therefore during the process of deleting stop words, this study synthesized the Sichuan University Machine Intelligence Laboratory Stop Words, Baidu Stop Words, Harbin Institute of Technology Stop Words, etc. [Zhou and Cao (2011)]. To establish a full stop word list, and excluded words that could not represent text features, considering the clustering effect of fusing multi-stop vocabulary is better than that of single stop vocabulary.

The data preprocessing flowchart is shown in Fig. 3.

3.4 Data vectorization

Text data vectorization is the mathematization of words in text, in other words, the representation of a piece of text as a vector. This study used a TF-IDF-weighted BOW (bag of words) model. The BOW model divides a piece of text into words, ignoring its word order, grammar and other elements. It is only regarded as a collection of several words. The appearance of each word in the text is independent and does not depend on whether other words appear. And then constructs a matrix consisting of 0 and 1 according to whether the feature word appears and represents the matrix as a vector. This study uses

the BOW model as a prototype to replace the non-zero values appearing in the text vector with the weights of the TF-IDF, which is more effective than the traditional BOW model in text classification [Lilleberg, Zhu and Zhang (2015)].

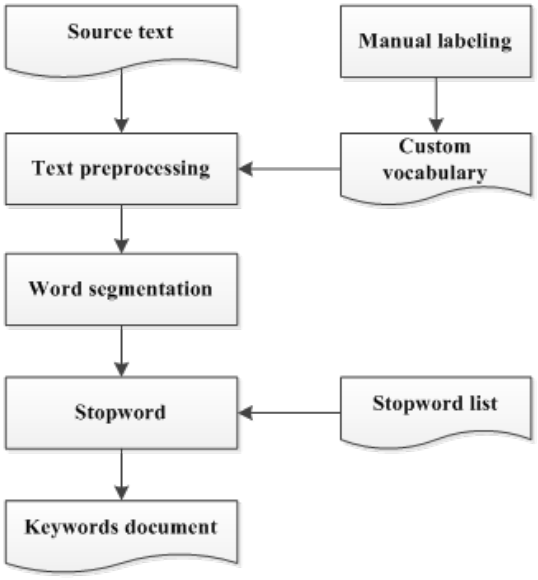

Figure 3: Data preprocessing flow chart

The data vectorization process is completed by using python to call the sklearn library. Sklearn is a machine learning library that includes methods such as regression, dimensionality reduction, clustering, and classification. This study uses sklearn's Feature extraction library to complete the vectorization of data.

TF-IDF is one of the most commonly used features in keyword extraction monitoring methods [Salton and Buckley (1988)], which is to measure the importance of a word for a document or a part of a corpus. As for the word X, its feature extraction function is:

$$f(x) = TF(x) * IDF(x) \tag{1}$$

where TF(x) represents the frequency of the word x and IDF(x) stands for the inverse document frequency of the word x.

$$TF(x) = \frac{n(x)}{\sum n} \tag{2}$$

where n(x) represents the number of times the word x appears in the text, and ∑n represents the total number of occurrences of all words in the text.

As a classic Chinese text analysis method, word frequency statistics refers to extracting a certain amount and length of corpus and calculating the number of occurrences of different words. A word appears more in a piece of text, the entropy of the word covers greater, which means that the word is more representative. Therefore, it is suitable for the feature words as text classification, which is convenient to perform further analysis to the key content and main purpose. Considering the length of the document, the same word

has a higher word frequency in the long document than in the short. The word frequency needs to be standardized, that is, the number of occurrences of the word is divided by the total number of words in the document, in case it is biased towards the long document.

However, only a small amount of text contains feature words that are more important than the feature words contained in a large amount of text, and it is more advantageous to distinguish the categories of the text, so IDF (inverse document frequency) is selected as another evaluation factor. If the number of texts containing the feature word x in the text set is less, the better the class discrimination of x is. IDF can reduce the importance of feature words contained in a large amount of texts and can also enhance the importance of feature words contained in only a few texts. Therefore, the TF is usually used in combination with the IDF.

The following formula is commonly used to calculate IDF:

$$\mathrm{IDF(x)} = \log\left(\frac{N}{N(x)+1}\right) \tag{3}$$

where N is the total number of texts and N(x) is the number of documents containing x.

3.5 Text clustering

In this study, K-means algorithm is used for text clustering [Jain (2010)]. K-means is a dynamic clustering algorithm which is based on the smallest squared error between the data to be clustered and the cluster center. In the K-means algorithm, the degree of similarity between data is represented by Euclidean distance, and the steps of k-means algorithm for clustering are shown in Fig. 4.

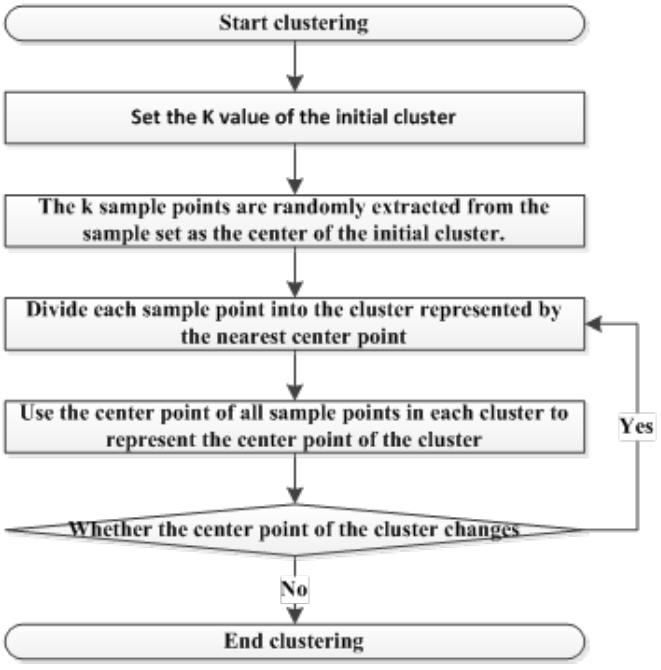

Figure 4: K-means clustering flow chart

The Euclidean distance between the data point x and the center point c is shown as following:

$$d(x, c)^2 = \sum_{i=1}^{n}(x_i - c_i)^2 \quad (4)$$

where n is the dimension, xi and ci are the ith attribute values of x and c. In addition, SSE (sum of squares due to error) is usually used to evaluate the quality of clustering results. The calculation formula of SSE is as follows:

$$SSE = \sum_{i=1}^{k}\sum_{x \in s_i} d(x_i, c_i) \quad (5)$$

where d is the Euclidean distance between the data point x and the center point c, k is the number of clusters, x is the data point, ci is the center of the ith cluster, and s_i is the set of data points in the ith cluster.

For the fact that the performance of the algorithm is affected greatly by the selection of the k value, so how to correctly determine the k value is the key of the k-means algorithm. Therefore, in this paper, different clustering k values are set for pre-experiment, and the line graph between k value and SSE is obtained, as shown in Fig. 5, then calculating the rate of decline of SSE, and next draw a line chart between the rate of decline and the value of K as shown in Fig. 6. As can be seen from Fig. 5, with the cluster K value increasing, SSE is decreasing smoothly. However, when the cluster k value is greater than 40, the SSE value begins to fluctuate up and down. The phenomenon shown in Fig. 5 is more intuitive. In Fig. 5, when the cluster k value is over 40, the SSE rate decline above and below zero. The reason for this phenomenon is that the randomness of K-means' initial cluster center selection leads to the clustering result falling into a local optimal deadlock. In addition, it can be found from Fig. 6 that when the clustering k value is 39, the best clustering effect can be achieved. Nonetheless, there is no local optimal deadlock occurring and the decline rate of SSE is greater than zero at this time.

From the above, the cluster k value of this paper is taken as 39.

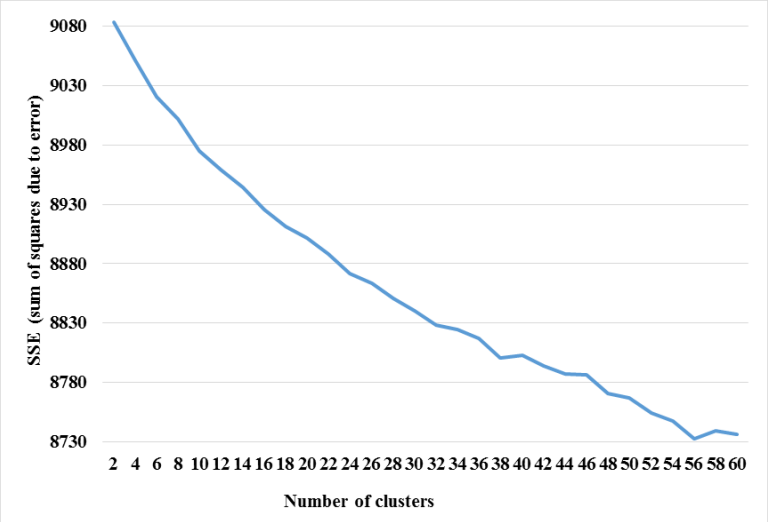

Figure 5: The line graph between SSE and clustering k value

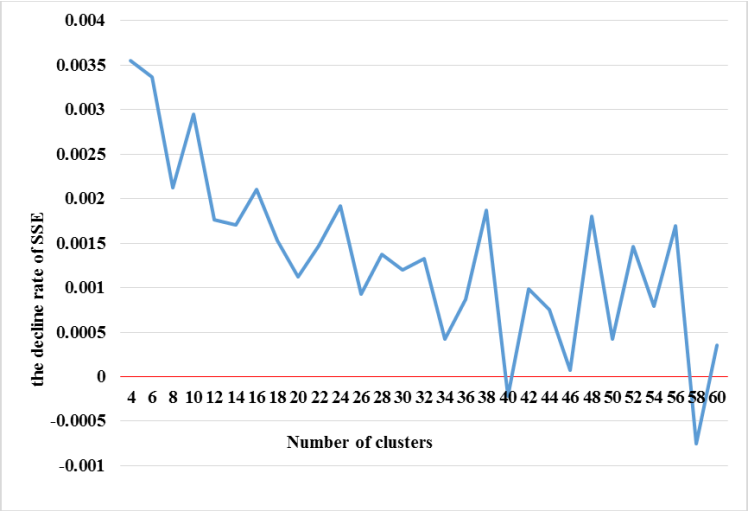

Figure 6: The Line chart between SSE decline rate and cluster k value

4 Data visualization and analysis

4.1 Data visualization

The cleaned text data is generated into a word cloud map, and the data for completing the clustering is visualized by means of Gephi tools. In addition, the Party's eight-point frugality code monthly summary data of the regulations can also be visualized to summarize the experience of the past anti-corruption work, and analyze the outstanding problems faced by the current anti-corruption work, and thereby propose targeted recommendations for the future.

The process of generating the word cloud map is completed by adopting Python to call the Wordcloud package. Wordcloud, as the third party library of Python word cloud display, which can draw the word cloud map as the parameter frequency of the words in the text, and actually it is widely used since it can be customized by the user from many aspects, such as the word cloud image background, color and size. Furthermore, the keywords frequency bar chart can be derived from the data preprocessed results.

The clustering result network diagram is completed by Gephi. Since Gephi supports inputting csv documents in a specific format and the csv document is composed of node data and side data, therefore, the data need to be manually processed before the data imported. Firstly, extracting the key words from the class documents after the clustering, and the keywords whose word frequency is not less than 10 are selected; Next, the keywords between the documents are compared to calculate the similarity between the documents and the documents; Finally, the similarity between the documents is used as the weight of the data imported by Gephi, and it is judged whether it belongs to the same big class according to the distance and the thickness of the edge. In general, the smaller the distance between classes, the thicker the class is connected to the class, and then the more likely it belongs to the same class. The clustering result network diagram is

summarized as follows: there are 39 nodes with 741 edges, the average of each node is 19, and the network diameter is 1.

The key words network graph is established on the basis of the clustering network graph, and each key word is connected with its class. The key words frequency is taken as the weight of the node data, and the proportion of the node in this class is determined according to the font size and color depth of the node label. Generally speaking, words with larger font sizes and darker colors are more likely to be representative words of the class. The overview of the network diagram is as follows: there are 2976 nodes and 2976 edges, the average of each node is 1, and the network diameter is 5. Some of the results of this visualization are shown below:

Figure 7: Anti-corruption keyword cloud map

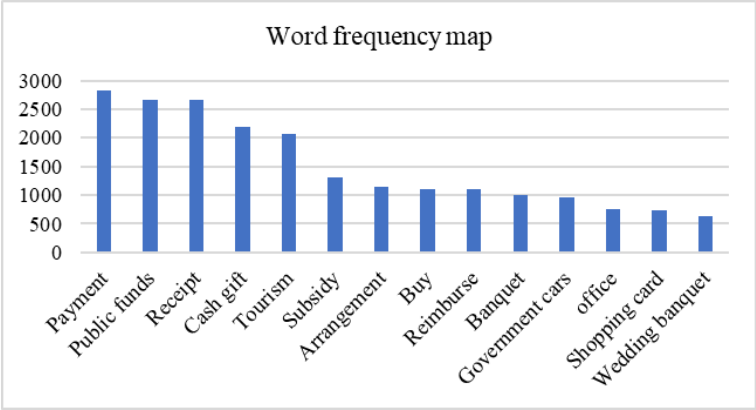

Figure 8: Word frequency map

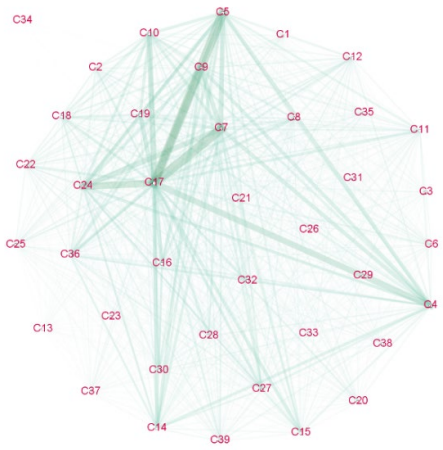

Figure 9: Clustering result network diagram

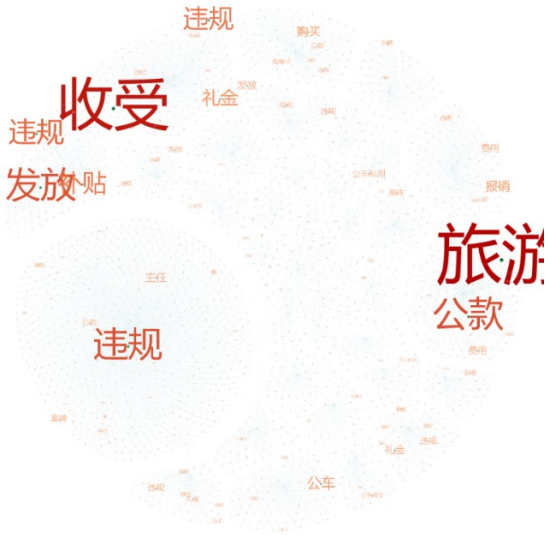

Figure 10: Keywords network diagram (overall)

4.2 Data analysis

It can be observed from Fig. 7 that the word 'violation (违规)' is particularly eye-catching, and in turn it can be noticed the following words, such as 'payment (发放)', 'public funds (公款)', 'receipt (收受)', 'cash gift (礼金)', 'tourism (旅游)', 'subsidy (补贴)', 'arrangement (操办)', 'banquet (宴请)', 'government cars (公车)' , these words can play a vital role in analyzing the issue of integrity and anti-corruption. From keywords frequency map in Fig. 8, it can be observed the word frequency of the words 'payment', 'public funds', 'receipt', 'cash gift' and 'tourism' are all greater than 2000, which indicates that in the anti-corruption problem there are many related cases investigated,

such as 'illegal delivery of gifts', 'illegal payment of subsidies or benefits', 'public funds travel', and 'banquets using public funds'. Thus, it is necessary to focus on the above issues in future work.

Looking at Fig. 9, it is found that the sides connected between C17, C24, C7, C5, C4, etc., are relatively thick, and the distances between C17, C24, C7, C19, etc. are short, so they can be regarded as the same class. As can be seen from Fig. 10, the overall keyword network diagram is still highlighted by the keywords with higher word frequency.

Because of space reasons, this paper makes use of the intercepted part of ithe overall keywords, as presented in Figs. 11 and 12.

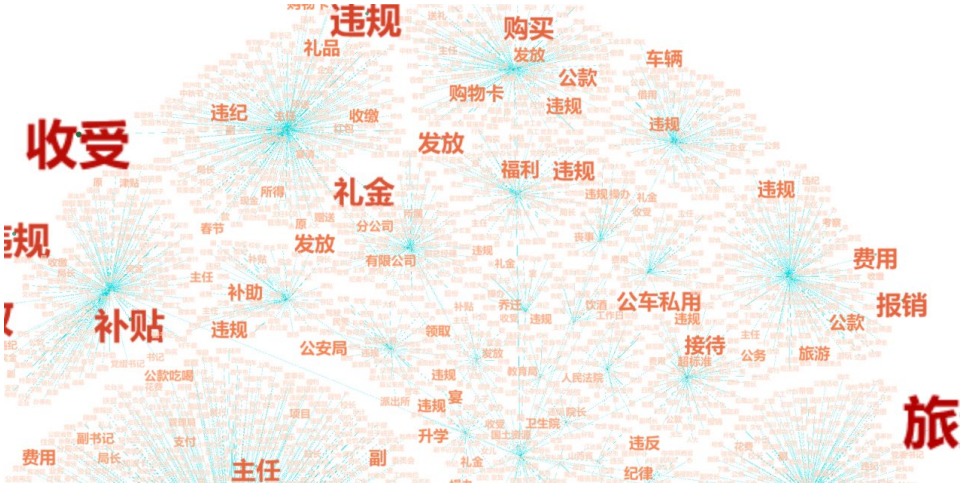

Figure 11: Keywords network diagram (part 1)

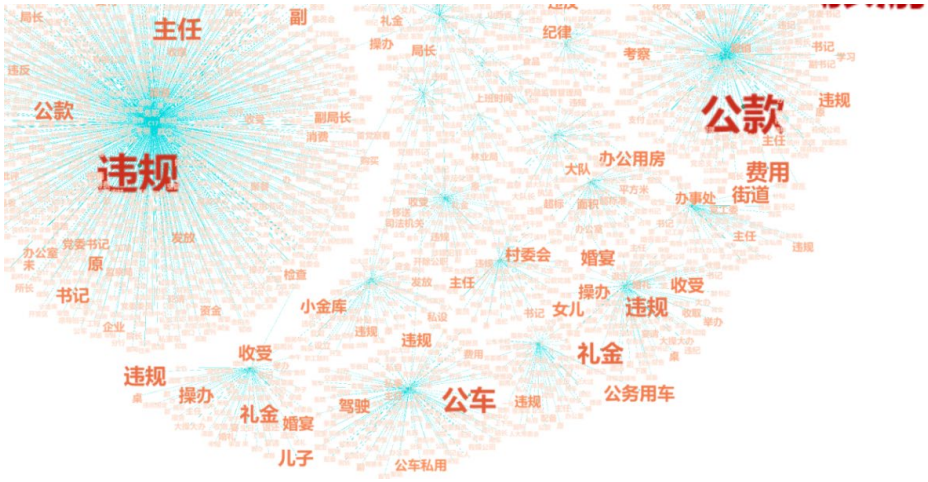

Figure 12: Keywords network diagram (part 2)

From Figs. 11 and 12, which presents the keywords are particularly prominent, such as receipt (收受), tourism (旅游), violation (违规), subsidy (补贴), cash gift (礼金),

shopping cards (购物卡), public funds (公款), government cars (公车), wedding banquet (婚宴), etc. Therefore, various violations in the picture can be classified as following 'illegal payment of subsidies or benefits, illegal delivery of gifts, illegal use of official vehicles, banquets using public funds, extravagant wedding ceremonies and funeral, public funds travel, and other major categories. In addition, the size of the keywords in the figure also reflects the number of events represented by the word. Thus, it is not difficult to find out from the above two figures that there are more issues represented by words such as receipt, travel, violations, public funds, subsidies, and cash gifts.

To put it another way, from the official monthly report data of the CCDI, the pie chart for drawing eight rules for processing type is shown in Fig. 13.

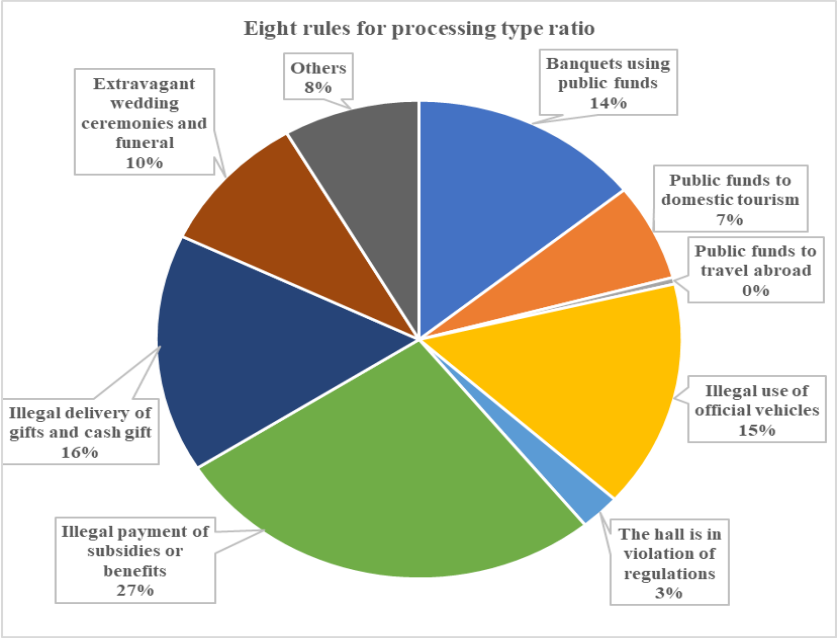

Figure 13: Eight rules for processing type ratio

The analysis of the data in Fig. 13 shows that illegal payment of subsidies or benefits, illegal delivery of gifts and cash gift, illegal use of official vehicles, banquets using public funds, extravagant wedding ceremonies and funeral are prominent problems in anti-corruption work. These five types of issues accounts for 82% of the total number of eight required issues. This is the same as the conclusion obtained from the analysis of the keyword network diagram, which proves the scientific and practicability of the research method.

According to Fig. 14 classified in yearly units, in the number of people enumerated, it can be known that the number of people surveyed shows a trend of rising month by month, indicating that the concept of anti-corruption and honesty is gradually deepening into the hearts of the people with the continuous development of anti-corruption work. From the overall trend of the past, the trend of the curve has experienced rapid rise, slow rise, peak, and sharp decline, which shows a 'time lag' effect in this procedure.

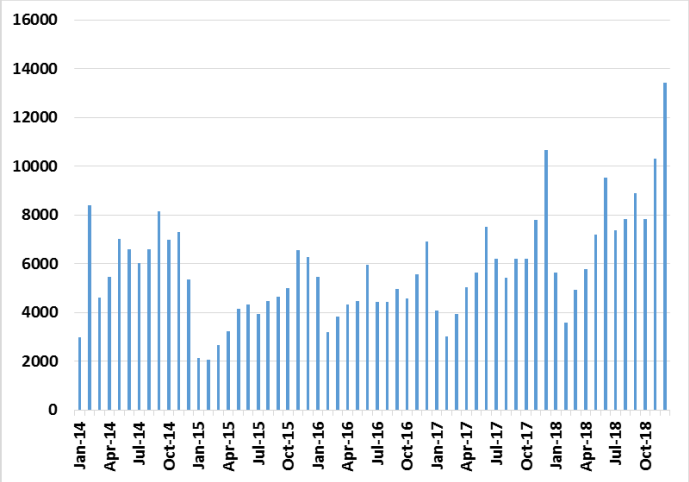

Figure 14: Histogram of the number of investigations

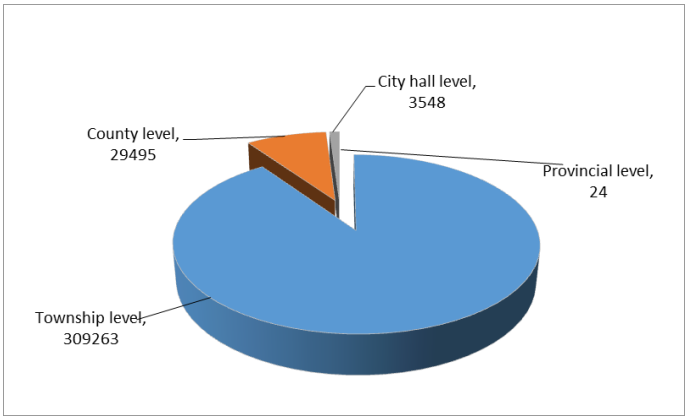

Figure 15: Level of the number of investigations

From Fig. 15, it can be analyzed that the relationship between the level composition of the investigated officials and the number of people. In this research data, as the official position is higher, there are fewer number of investigated and punished, which constitutes a low-centered anti-corruption strategy. In terms of the composition of the number of people, the proportion of the township level accounts for 90%, while the proportion of the provincial and provincial level is less than 0.01%. Therefore, President Xi Jinping pointed out: "It can be seen from the large number of cases that have been investigated and dealt with, some party members and cadres have turned a deaf ear to the disciplinary regulations, disregard the four forms of decadence, and take any chances to corrupt. Therefore, the discipline must be further stricter." [Xi (2014)].

5 Conclusion and future work

5.1 Conclusion

From the data analysis of this study, the following conclusions can be drawn:

First of all, the outstanding problems faced by China's anti-corruption struggle work are illegal payment of subsidies or benefits, illegal delivery of gifts and cash gift, illegal use of official vehicles, banquets using public funds, extravagant wedding ceremonies and funeral. Therefore, it should be given priority to the allocate the resources of supervision verification, discipline inspection, etc., which is so as to perform accurate anti-corruption work.

Secondly, the number of people involved in various types of corruption incidents presents a cyclical pattern of rapidly rising, slowly rising, reaching peaks and gradually falling back on a yearly basis, which is caused by the "time lag" effect of policies. The reason for the change in the trend of the pattern is that the relevant policies of the anti-corruption campaign from start to implementation needs time to adapt, so there are a 'time lag' effect in this process.

Finally, in terms of the composition of the number of people, the number of people in the township level accounts for 90%, so it is the main target of the anti-corruption struggle. There are three reasons for this phenomenon: First, the structure of cadres at the township level are relatively complex and large in number, and comparatively speaking, it is relatively high probability to investigate this levels; Second, there is the work style of grassroots party inherited for a long time, which becomes the grassroots officials' dependence, and it is difficult to break the path dependence in a certain period of time. Third, the rule information transmission is lagging behind. Since the township level cadres work at the grassroots level of the government. However, the communication of policies needs to be uploaded and delivered in turn, resulting in loss of energy transmission. Therefore, the implementation of policies is relatively lagging, which causes grassroots cadres to have a chance to be lucky, so they have become the hardest hit areas.

5.2 Future work

In future research methods, other clustering algorithms or machine learning algorithms can be considered to realize the prediction function of big data. In the future data acquisition, more comprehensive information can be considered to acquire and the collection unit can be expanded from the province to the city, even towns and villages, in order to collect more and all-sided data; in future's data processing, we can consider using the model with a better clustering effect. In addition, the data can be considered for dimensionality reduction during processing [Napoleon (2011)].

The core of big data is prediction. Through the analysis of historical data, after a period of machine learning, a corruption previous warning model can be built to apply big data technology to the prediction of corrupted behavior. Data information can be obtained on the Internet through formal and informal channels. Therefore, the collected information can be accurately analyzed, identified, judged, and evaluated, and then draw relevant conclusions. Moreover, with the construction and use of cloud computing platforms [Sun, Fan, Jiang et al. (2019)], this model of the big data era will certainly become a reality.

Acknowledgement: This research is funded by the Open Foundation for the University Innovation Platform in the Hunan Province, grant number 16K013; Hunan Provincial Natural Science Foundation of China, grant number 2017JJ2016; 2016 Science Research Project of Hunan Provincial Department of Education, grant number 16C0269. Accurate crawler design and implementation with a data cleaning function, National Students innovation and entrepreneurship of training program, grant number 201811532010. This research work is implemented at the 2011 Collaborative Innovation Center for Development and Utilization of Finance and Economics Big Data Property, Universities of Hunan Province. Open project, grant number 20181901CRP03, 20181901CRP04, 20181901CRP05. We also thank the anonymous reviewers for their valuable comments and insightful suggestions.

References

Bastian, M.; Heymann, S.; Jacomy, M. (2009): Gephi: an open source software for exploring and manipulating networks. *Third International AAAI Conference on Weblogs and Social Media*, pp. 361-362.

Biswas, G.; Weinberg, J. B.; Fisher, D. H. (1998): ITERATE: a conceptual clustering algorithm for data mining. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 28, no. 2, pp. 219-230.

Blancke, S. (2018): Tigers, flies and crocodiles: hunting season for chinese intelligence. chinese anti-corruption campaigns in an aggressive foreign policy. *Zeitschrift Für Außen- und Sicherheitspolitik*, vol. 11, no. 3, pp. 343-364.

Card, M. (1999): *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Press, USA.

Chen, C. (2008): An information-theoretic view of visual analytics. *IEEE Computer Graphics and Applications*, vol. 28, no. 1, pp. 18-23.

Dai, C. Z. (2010): Corruption and anti-corruption in China: challenges and countermeasures. *Journal of International Business Ethics*, vol. 3, no. 2, pp. 58-70.

Defanti, T. A.; Brown, M. D. (1991): Visualization in scientific computing. *Advances in Computers*, vol. 33, pp. 247-307.

Friendly, M. (2008): A brief history of data visualization. *Handbook of Data Visualization*, pp. 15-56. Springer, Berlin, Heidelberg.

He, Z. (2000): Corruption and anti-corruption in reform China. *Communist and Post-Communist Studies*, vol. 33, no. 2, pp. 243-270.

Hirsch, D. D. (2013): The glass house effect: big data, the new oil, and the power of analogy. *Social Science Electronic Publishing*, vol. 66, pp. 373.

Jain, A. K. (2010): Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651-666.

Jain, A. K.; Murty, M. N.; Flynn, P. J. (1999): Data clustering: a review. *ACM Computing Surveys (CSUR)*, vol. 31, no. 3, pp. 264-323.

Jun, D. (2014): Comparative study of the social network analysis tools: ucinet and gephi. *Information Studies: Theory & Application*, no. 8, pp. 27.

Keim, D.; Qu, H.; Ma, K. L. (2013): Big-data visualization. *IEEE Computer Graphics and Applications*, vol. 33, no. 4, pp. 20-21.

Lilleberg, J.; Zhu, Y.; Zhang, Y. (2015): Support vector machines and word2vec for text classification with semantic features. *IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing*, pp. 136-140.

MacQueen, J. (1967): Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, no. 14, pp. 281-297.

McAfee, A.; Brynjolfsson, E.; Davenport, T. H.; Patil, D. J.; Barton, D. (2012): Big data: the management revolution. *Harvard Business Review*, vol. 90, no. 10, pp. 60-68.

Napoleon, D.; Pavalakodi, S. (2011): A new method for dimensionality reduction using k-means clustering algorithm for high dimensional data set. *International Journal of Computer Applications*, vol. 13, no. 7, pp. 41-46.

Paine, D. (2015): Book review the data revolution: big data, open data, data infrastructures & their consequences. *Computer Supported Cooperative Work*, vol. 24, no. 4, pp. 385-388.

Salton, G.; Buckley, C. (1988): Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, vol. 24, no. 5, pp. 513-523.

Sun, G.; Fan, X. P.; Jiang, W. D.; Li, F. H.; Jiang, Y. W. (2019): Obfuscation-based watermarking for mobile service application copyright protection in the cloud. *IEEE Access*.

Sun, J. Y. (2019): Project description-jieba. https://pypi.org/project/jieba/.

Xi, J. P. (2014): *The Govermence of China*. Foreign Languages Press, China.

Zhou, Y.; Cao, Z. W. (2011): Research on the construction and filter method of stop-word list in text preprocessing. *2011 Fourth International Conference on Intelligent Computation Technology and Automation*, vol. 1, pp. 217-221.