ISSN: 2664-9640

Text Analysis Based on Python: Research on Management Discipline Research Topics and Evolutionary Laws

Hui Wang[1], Yuanqi Chen[1], Zhendong Pan[2]

[1]School of Management Science and Engineering, Anhui University of Finance and Economics, Anhui 233000, China;

[2]School of Architecture, Anhui Science and Technology University, Anhui 233000, China.

Abstract

This project aims to extract information from the management subject projects funded by the National Social Science Fund and the National Natural Science Fund in China, identify topics, and mine hidden information through topic analysis based on the LDA algorithm. Compared with traditional research methods, this project utilizes Python programming for text data mining, which can significantly improve the efficiency of studying the commonalities, differences, and trends of related social science projects. The steps involved in this project include literature review, text data acquisition using Python web crawling, topic extraction from short text data files using the LDA model, clustering analysis and evolutionary analysis of social science fund project topics, and the formation of research trend orientations for social science fund projects. Further research on the hotspots in the social science field is conducted, which has significant implications for future research on scientific data and the application, organization, and management of topics in the National Social Science Fund and the National Natural Science Fund.

Keywords

Management, LDA, Python, Text Analysis, Research Topics.

1. Introduction

With the advent of the artificial intelligence and big data era, network and information technology have pervaded all aspects of human daily life, and real-world textual information is increasingly being presented in electronic form. Consequently, text mining has emerged as a research hotspot and a central focus in the information field. In this context, the utilization of computers to realize the recognition and analysis of massive text has gained considerable research interest. Meanwhile, with the continuous development of the digital economy, the value of data has gradually surfaced, becoming a vital driving force for social development.

The National Social Science Foundation of China (NSSC) represents the highest level of research funding in the field of humanities and social sciences, endowed with strong authority and representativeness. By applying text mining to the selected topics of NSSC projects, researchers can extract high-value topics in terms of academic value, application value, novelty, and project data. This approach facilitates the identification of project topics, significantly alleviating the problem of "difficult topic selection".

Since its establishment in 1983, the project guidelines and the projects funded by the National Social Science Foundation have consistently addressed major real-life issues and disciplinary development, objectively reflecting the focus, hot spots, current situation, and development direction of research in various disciplines. The National Social Science Foundation of China (NSSC) formally added the discipline of management in 2009 and began to accept applications in 2010, with a focus on theoretical and practical issues in China's economic and social management.

Through the use of text mining technology, researchers can reflect the key text and key data of NSRF projects in a clear and straightforward manner, thereby extracting critical information. However, compared with the world's top countries, there is still a significant gap in text mining of NSRF projects in China. For instance, the probability that key projects have been text-mined is less than 1%. Many developed countries worldwide have begun to prioritize text mining of research projects, especially by utilizing the simple and efficient Python language.

In the realm of national research fund projects, text mining using the Python language holds tremendous potential. Presently, the number of articles related to management disciplines retrieved on the Internet is small, and research is primarily focused on SCISSCI, A&HCI database papers, or master's and doctoral dissertations. As the highest level research fund project in humanities and social sciences, the National Social Science Foundation carries strong authority and representativeness. Thus, utilizing it as the research object can effectively capture the hot information of tourism discipline research, providing researchers with the necessary basis and foundation for conducting theoretical and practical research, possessing robust academic and application value.

2. Current Research Trends

2.1. Research trends abroad

Research on text mining started earlier abroad, and the United States is one of the countries with the earliest standardization of scientific research evaluation activities. In 1914, the Congressional Research Service (CRS) was established in the United States, and the "Project Evaluation Standards" was promulgated in 1975. In the late 1950s, H.P.Luhn conducted pioneering research in this field and proposed the idea of word frequency statistics for automatic classification. In 1960, Maron published the first paper on automatic classification. Subsequently, many scholars represented by K.S. Park, G. Salton, and K.S. Jones have also carried out fruitful research in this field. Currently, research on text mining abroad has moved from the experimental stage to the practical stage. The project evaluation agencies in the United States mainly include Congress, unofficial organizations, and universities themselves. The evaluation of scientific research in American universities mainly uses expert review and quantitative methods. The main method is based on quantitative indicators, and researches corresponding data of various scientific research results to analyze the impact of quantitative data on scientific research [1]. With the multi-dimensional infiltration of disciplines, and the increasingly integrated architecture of the connotation and extension of knowledge systems, the establishment and improvement of the evaluation system for universities is of great importance [2].

2.2. Domestic research dynamics

2.2.1. Text Mining Research

The evaluation of scientific research projects in China began late, but with the continuous development of science and technology, technological innovation has gradually become the main driving force for economic and social development [3]. Under the high attention of society, the funding for research projects has continued to increase [4]. Since the reform of the scientific research system in the 1990s, the evaluation of research projects has gradually been given attention. At the onset of the Fourth Industrial Revolution, China put forward the concept of "Made in China 2025," which posed higher requirements for China's research work. Research assessment plays a critical role in scientific research activities and talent education. Similar to foreign evaluation methods, China uses qualitative analysis based on peer review and quantitative analysis based on econometric methods. However, there is still a certain gap with foreign evaluation methods in terms of indicator system, standardization, and scientificity, which is related to the fact that the development of China's research evaluation work is still in the stage of improvement [5].

In recent years, research on TDT technology has emerged both domestically and internationally, with language models being the main focus in text topic discovery. Among them, LDA [6] has achieved many results in application by implementing topic allocation based on a generative model. Li Chang [7] et al. introduced the WI-LDA model, which is based on the technical words and IPC context, and achieved better results in task-specific topic generation. Tan Xu [8] et al. fused the ARMA model and LDA for dynamic presentation and fine-grained division in sentiment analysis. In addition, to compare the application effects of experiments more intuitively, it is also necessary to determine a topic number K as a control variable. However, there are many methods for determining the number of LDA topics. Blei [9] et al. used the Perplexity index to measure the model's uncertainty about topic model allocation, but this method tends to be biased towards high topic numbers. Wang Xiwei [10] et al. used the Occam's razor principle to determine the minimum curve inflection point as the topic number, but this method lacks stability and is difficult to ensure that the solution obtained is the optimal solution. Griffiths [11] et al. used Bayesian statistical criteria, replacing the Perplexity index with the method of logarithmic marginal likelihood. Guan Peng [12] et al. used JS divergence to calculate the variance of each topic-word distribution parameter around its mean and combined it with the Perplexity index to propose the Perplexity-Var index. Although these methods have made improvements based on the original indicators, they still fall within the scope of multiple training. Teh [13] et al. proposed the Hierarchical Dirichlet Process (HDP), which independently extracts each sample from the mixture distribution, and generates the final mixture component score by completing the sampling process. The experimental results showed that the best mixture component score was consistent with the results obtained by the Perplexity method.

Through text mining technology, the information of national scientific research fund projects can be mined to reflect the key words and data of these projects in a simple and clear manner, thereby extracting important information from them. In 2021, China's National Natural Science Foundation project was divided into 14 major categories with a total of more than one million projects approved, including 3,917 key projects with a huge amount of funding. However, in the text mining of these projects, China lags far behind world-class countries. For example, less than 1% of these key projects have been text-mined, whereas many developed countries have already begun to attach importance to text mining of research projects, especially using the concise and efficient Python language to conduct text mining of their own research fund projects. Therefore, in the national scientific research system, more attention should be paid to the promotion and application of text mining technology, which can provide strong support for the optimization

2.2.2. Topic Analysis in Tourism Studies

The research theme and evolution of tourism management as a discipline is still in its infancy stage in China, and there is a lack of research findings in this area. Currently, the main topics and hotspots of related disciplines in China are focused on tourism management, tourism education, tourism ecology, and health tourism. Yishao Hua summarized and collated the development, research themes, and methods of tourism studies both domestically and internationally through literature review [14]. Zhang Lingyun and his team conducted a study on 176 articles published in the 2009 "Ren Da Photocopy Journals" by combining data statistics to explore the current status and hotspots of tourism management theory research, predicting the development trend of tourism research in China [15]. Zhang Juan and her team used the 8,162 tourism management research literature in CSSCI as the data source and conducted visual analysis using knowledge mapping tools [16]. Jin Minglei used research tools such as EndNote, Excel, and NVivo to sort and collate the time distribution, country and institution distribution, and highly cited literature distribution of 1,189 research papers on tourism education published in the SCI, SSCI, and A&HCI databases from 2009 to 2019 [17]. Some scholars focused on tourism education in the tourism discipline, and analyzed the keywords of 536 master's and doctoral theses related to tourism education in the full-text database of China National Knowledge Infrastructure using software such as Bicomb, SPSS, and Excel to understand the current research status of tourism education in China. Zhong Linsheng and Li Meng used CiteSpace software to analyze, sort, and explore the research literature on long time-series of Chinese ecological tourism, providing a visualization reference and reference for the theoretical exploration and practical development of ecological tourism in China.

Currently, the number of related articles retrieved from the internet is relatively small, and the research objects mainly focus on papers from the SCI, SSCI, and A&HCI databases or master's and doctoral theses. However, the National Social Science Fund Project is the highest-level scientific research funding project in the field of humanities and social sciences, with strong authority and representativeness. Using it as a research object can effectively grasp the hot information of tourism research, provide a basis and reference for theoretical and practical research for researchers, and has strong academic and practical value.

3. Methods and Data

3.1. Data Crawling Mining Method

Firstly, we collected and organized data from the literature of projects funded by the Chinese Social Science Foundation (CSSF) using Python. We employed a stop-word strategy and utilized web crawling technology to retrieve information on over 1,292 National Social Science Fund projects and 1,263 management-related projects. Next, we utilized the jieba library to segment, organize, rank, and cluster short texts related to research topics, laying the foundation for our next step of using the LDA algorithm for data analysis in this project.

3.2. LDA Algorithm Topic Extraction Method

The LDA (Latent Dirichlet Allocation) algorithm is a text topic model that decomposes text data into several topics and calculates their distributions, revealing the latent topic structure behind the text. Specifically, LDA views each document as a mixture of multiple topics, each topic consisting of multiple words whose distributions follow a Dirichlet distribution. LDA algorithm estimates the topic distribution of each document and the word distribution of each topic through Bayesian inference, thereby achieving topic modeling of text data.

In this project, LDA algorithm was employed to analyze text materials such as grant proposals and research reports to reveal information such as the research topics, keywords, and directions of the applicants. Specifically, the text data was preprocessed, including removing stop words, tokenization, and stemming. Then, the LDA algorithm was used to model the preprocessed text data, and the topic distribution of each document and the word distribution of each topic were calculated to analyze the text data. Finally, the LDA algorithm was used to obtain information such as the keywords of each topic, the similarity between topics, and the research directions and interests of the project applicants, providing a scientific basis for project evaluation and decision-making.

4. Analysis and Results

4.1. Data Collection and Visualization

We utilized Python web scraping techniques to acquire data from the National Social Science Fund Project Database, simulating the manual search process. The Selenium module was used to control the browser and select project categories, subject categories, and funding years, followed by clicking the search button to obtain the HTML source code of the search results page. By looping through different years, we extracted data from the database and saved it to a CSV file.

In the main program, we read the saved dataset and selected project names from a specified year for analysis. The project names were processed to retain only Chinese characters, and then segmented using the Jieba tokenizer for frequency analysis. We further extracted features using TF-IDF and applied the Latent Dirichlet Allocation (LDA) model for topic modeling. Finally, the top 10 most frequent words for each topic were outputted, and we created bar charts for the topics and word cloud graphs for the high-frequency words. The theme word cloud analysis and distribution of management social science fund project names for the years 2018, 2019, 2020, and 2021 are depicted in Figure 1 and Figure 2.

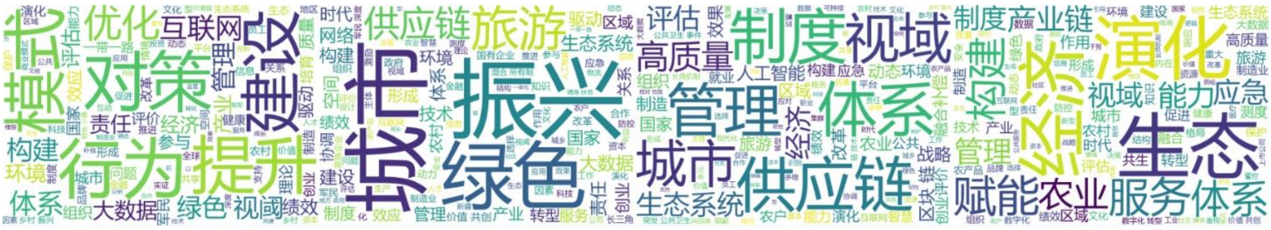

Figure 1: Word Cloud of Management Social Science Fund Projects

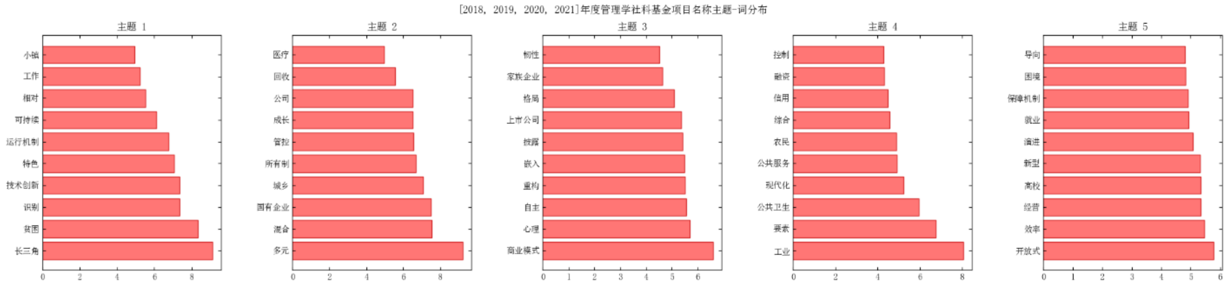

Figure 2: Distribution of Project Themes and Keywords

4.2. Topic Content Analysis

Based on the results of the topic modeling, we summarized a series of key words under different topics, and summarized the topic name of the topic, forming a list of topic words in the field of management research, see Table 1.

Table 1: Topic Words of Management Research

Theme	Name	Keywords(Top 3)
1	Development Connotation	Interdisciplinarity, Cross-Disciplinary, Integration,
2	Technological Background	Artificial Intelligence, Digital Technology, Big Data,
3	Establishment of Humanities Laboratory	Practice, Science and Technology, Modernization

4	Historical context	Textbooks, Disciplinary Systems, Discursive Systems
5	Discipline positioning	Discipline Construction, Interdisciplinary, Digital Humanities
6	Revolution of Teaching	Teaching Mode, Practice, Talent Cultivation
Theme	Name	Keywords(Top 3)
7	Construction philosophy	Course Ideology, Talent Cultivation, Moral Education
8	curriculum system	Talent cultivation, Applied Type, Practice

The analysis of the eight themes reveals several important trends in the field of management studies.

In the theme of Development Connotation, the top three keywords are Interdisciplinarity, Cross-Disciplinary, and Integration. This suggests that the integration of different disciplines is a major area of focus in management studies, with an emphasis on developing cross-disciplinary solutions to complex problems.

In the theme of Technological Background, the top three keywords are Artificial Intelligence, Digital Technology, and Big Data. This indicates that the incorporation of technology into management studies is increasingly important, with an emphasis on leveraging the latest advancements in AI, digital technology, and big data to drive innovation and progress.

In the theme of Establishment of Humanities Laboratory, the top three keywords are Practice, Science and Technology, and Modernization. This suggests that there is an increased focus on modernizing the field of management studies by integrating scientific and technological advancements and emphasizing practical, hands-on learning experiences.

In the theme of Historical Context, the top three keywords are Textbooks, Disciplinary Systems, and Discursive Systems. This indicates that there is a growing awareness of the historical context in which management studies have evolved, with a focus on analyzing the role of textbooks, disciplinary systems, and discursive systems in shaping the field.

In the theme of Discipline Positioning, the top three keywords are Discipline Construction, Interdisciplinary, and Digital Humanities. This suggests that there is a need for a more defined positioning of management studies as a discipline, with a focus on interdisciplinary approaches and digital humanities methodologies.

In the theme of Revolution of Teaching, the top three keywords are Teaching Mode, Practice, and Talent Cultivation. This indicates that there is a growing recognition of the need for new and innovative teaching methods, with a focus on incorporating practical, hands-on learning experiences and cultivating talent to meet the demands of the changing business landscape.

In the theme of Construction Philosophy, the top three keywords are Course Ideology, Talent Cultivation, and Moral Education. This suggests that there is an increased emphasis on the role of course ideology, talent cultivation, and moral education in shaping the field of management studies.

Finally, in the theme of Curriculum System, the top three keywords are Talent Cultivation, Applied Type, and Practice. This indicates that there is a growing focus on the development of curriculum systems that prioritize talent cultivation, applied learning, and practical experience.

4.3. Key Topic and Trend Analysis

Based on the above analysis, it can be concluded that the above 8 themes are hot topics in the field of management studies. Therefore, we choose 5 representative themes randomly for further analysis , and the search condition is set to include both new humanities and the theme label in the topic search, in order to screen out relevant literature. LDA topic modeling is performed on these articles, and the topic word distribution is further analyzed. See Figue 3.

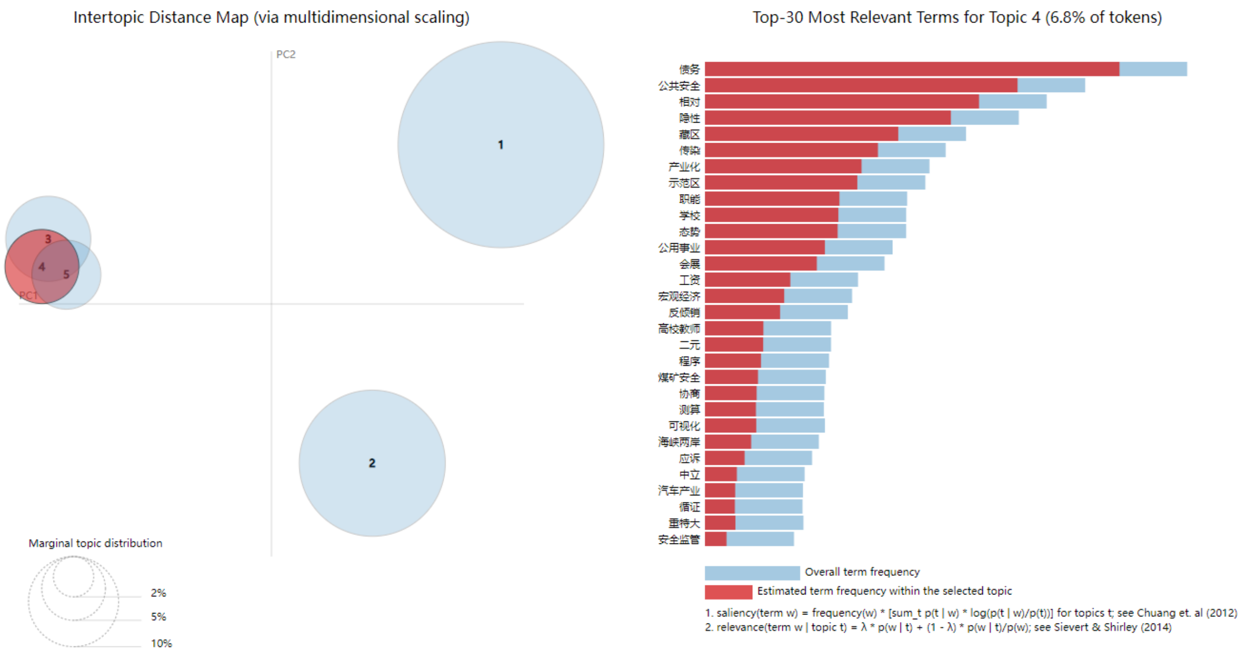

Figure 3: LDA Topic Modeling

Based on the analysis of the example provided above, we can conclude that the hot topics in the field of management studies are debt and public safety.

Debt and public security have become the research hotspots in this field today. With the increasing globalization and interdependence of national economies, debt has become a major challenge to economic development and financial stability, attracting significant attention from scholars and policymakers. Meanwhile, public security issues have become more complex and diversified, including terrorism, cybercrime, and social instability. Understanding the nature and impacts of debt and public security and proposing effective solutions are crucial for promoting sustainable development and ensuring social stability. Therefore, research in this field has attracted increasing attention from academia, government, and society at large, making it a critical area of study in the current academic landscape.

4.4. Areas for Improvement

Despite its usefulness in text mining and topic analysis in management research, the LDA algorithm has several limitations. Firstly, it is a topic modeling algorithm that is based on the Bayesian probability model, which assumes that each document comprises several topics and each topic contains multiple words. However, the complexity of text in management research can make it difficult to simply divide into individual topics, which could lead to LDA algorithm failing to accurately capture the topic information of the text. Secondly, the LDA algorithm's computational complexity is high when dealing with large-scale text data, resulting in a slow running speed. Additionally, the algorithm requires high pre-processing and parameter setting requirements for the text, which necessitates researchers to possess a high level of professional knowledge and skills. In conclusion, there is still significant scope for improvement in the methods employed in this paper.

5. Conclusion

The current study aimed to investigate the topics and evolution of Chinese National Social Science Fund (NSSF) projects from 2016 to 2022. Python was used to perform topic text mining, and various analytical methods, including LDA-based topic analysis, clustering analysis, and evolution analysis, were employed to identify the hot topics and compare them with previous research. The study found that, among others, debt and public security are current hot topics in NSSF projects and have the potential to become future research areas.

The multi-topic analysis provides an objective and comprehensive understanding of NSSF hot topics, which can be an effective reference for promoting innovation and development in NSSF. However, the study's use of LDA algorithm has limitations, such as the complexity of text management in research, which makes it difficult to simply classify topics. The subjective selection of stop words may also cause some bias. Therefore, further verification is needed for the evaluation of research trends in this study.

Acknowledgements

This project is sponsored by the Anhui University of Finance and Economics Scientific Research Fund (Project No. 90084) in 2022, titled "Research on Management Discipline Research Topics and Evolutionary Laws - Based on the Perspective of National Two-class Projects".

References

[1] Lu Yiyi, Guo Shengwei. Evaluation of scientific research in American universities and its reference [J]. Management Observation, 2016(21).

[2] Gross PF. A critical review of some basic considerations in post-secondary education evaluation [J]. Policy Sciences, 1973, 4(02): 171-195.

[3] Feng Yueqiang, Qi Wei. Analysis of the development law of technological originality [J]. China Science Foundation, 2007, 1: 14-16.

[4] Gu Quan. A comparative study of scientific research project funding management in China and Britain [J]. Science Research Management, 2012, 33(1): 120-126.

[5] Zhou Peng, Zhang Min, Guo Shengwei. Historical evolution, current situation, and countermeasures analysis of scientific research evaluation in China [J]. Management Observation, 2016, 32: 173-176.

[6] Li Yu, Wang Miao. Analysis of research achievements in humanities and social sciences in Shaanxi Province [J]. Information Exploration, 2015(05).

[7] Wang Chenguang. Bibliometric analysis of humanities and social sciences research on the Chinese Arctic: Based on statistics of CSSCI journals [J]. Journal of Ocean University of China (Social Sciences), 2017(02).

[8] Wang Xiaoxia. A quantitative analysis of National Social Science Fund projects in ethnology in the past decade [J]. Comparative Studies of Cultural Innovation, 2020(35).

[9] Ma Cunyong, Wang Yongbin. Research progress on Marxist theoretical disciplines in western China from the perspective of National Social Science Fund projects: Based on data analysis of projects approved from 2009 to 2018 [J]. Journal of Lanzhou Jiaotong University, 2020(03).

[10] Pei Zhenwei. Analysis report of National Social Science Fund projects on religious studies from 2008 to 2019 [J]. World Religious Culture, 2019(05).

[11] Zhou Yuan, Liu Huailan, Du Pengpeng, Liao Ling. Research on text classification model based on improved TF-IDF feature extraction [J]. Journal of Intelligence, 2017, 35(5): 111-118.

[12] Liu Xiaohui, Li Changling, Feng Zhigang. Analysis of research hotspots in disciplines based on improved TF*IDF method: Taking library and information science as an example [J]. Journal of Intelligence, 2017, 35(7): 82-87.

[13] Li Chang, Yi Huifang, Wu Hong, Ji Fangyan. Theme analysis of patent technology for autonomous driving cars based on WI-LDA topic model [J]. Journal of Intelligence, 2018, 37(12): 50-55.

[14] Ma Sidan, Liu Dongsu. Research on text classification method based on weighted Word2vec [J]. Journal of Intelligence, 2019, 37(11): 38-42.

[15] Cao Yimei, Li Zhenqi. Influence of Weibo public opinion on the development of hot events [J]. News Enthusiasts, 2020, 0(1): 47-49.

[16] Tan Xu, Zhuang Muni, Mao Taitian, Zhang Qian. Analysis of sentiment evolution in large-scale network public opinion based on LDA-ARMA hybrid model [J]. Journal of Intelligence, 2020, 39(10): 121-129.

[17] Wang Xiwei, Zhang Liu, Huang Bo, Wei Yanan. Construction and Empirical Study of Weibo Users' Topic Graph Based on LDA: Taking the "Ethiopian Airlines Crash" as an Example. Data Analysis and Knowledge Discovery, 2020, 4(10): 47-57.