DOI: 10.23977/socsam.2022.030205
ISSN 2523-5796 Vol. 3 Num. 2

LDA-based Topic Mining Research on China's Government Data Governance Policy

Qian Yang

Xi'an Polytechnic University, Xi'an, Shaanxi, 710048, China

Keywords: Government Data Governance Policy, LDA Model, Topic Similarity, Topic Hotspot

Abstract: Mining China's government data governance policy themes and analyzing the theme evolution paths are helpful to discover the trend of government data governance policy evolution. Using 443 government data governance policies at or above the provincial and ministerial levels in China as data sources, divided into different time windows by years, the policy themes under different time windows are identified by LDA topic model, and the classification effect is judged by using visualization tools, and the topic hotspots are obtained by categorizing them according to the similarity of each phase. Through the analysis of China's government data governance policies under different time windows thematic hotspots clustering, in order to provide corresponding reference for China's government data governance field.

1. Introduction

With the improvement of computer storage capacity and the development of complex algorithms, the volume of data has developed exponentially in recent years, and we are gradually moving from the information age into the era of big data. In the era of big data, data resources have become an important basic strategic resource in China. Most of the data resources are stored in government departments[1] for collection and analysis by government,enterprises and various sectors of society. Along with the data types continue to increase, and the data volume continues to increase, which will force the government to carry out effective data governance.

"Government data governance" is an important element of government governance and aninevitable requirement for modernizing the national governance system and capacity[2], and one of the roles of the government in data governance is the policy maker[3], and the implementation of data governance policy formulation has a good guiding and standardizing effect on government data governance[4]. National policy support can effectively strengthen the data governance and actively guide the public and enterprises to access and utilize government data. Therefore, it is important to dig into the hotspots of government data governance policy topics in China.

2. Literature Review

The literature related to government data governance policies at home and abroad is reviewed and analyzed through literature research method.

The study of domestic government data governance policy research is mainly conducted from three aspects: introduction and borrowing from foreign policies, quantitative analysis of policies,and comparative analysis of policies. In terms of the introduction and borrowing of foreign policies, Li Chongzhao[5] focuses on the data governance policy and governance structure of the UK central government and summarizes the main areas covered by its policies; Zhou Wenhong[6]analyzes the policy textsof the U.S. federal government and proposesconstruction strategies for China's governmentdata governance system. In terms of the quantification of policy texts, Liu Binfang[7] constructed a two-dimensional policy analysis framework of policy tools and the core concept system of big data governance to quantify data governance policy texts; Song Yi[8] used the policycontent analysis method to content code the big data governance-specific policies of the United States, the United Kingdom and Australia . In terms of comparative policy analysis, Yao Gan[9] constructed a government big data governance system and coded the text of big data governance policies in Shanghai, Beijing and Shenzhen with the help of NVIVO12 qualitative analysis tool, and mapped and analyzed the components of the governance system; Liang Zheng[10] compared the data governance policies of the US, EU, UK and Chinabasedon the characteristics of the data governance policy framework.

Foreign government data governance policy research mainly focuses on two aspects: policy analysis and policy framework research. In terms of policy analysis, Napoli P M[11] for principles based on public policy formulation, assessed the challenges faced by communication policy-related data in improving transparency and accessibility and made recommendations for improvement; Y annoukakou A[12] analyzes the synergy of the Right to Know (RIT) and OpenGovernment Data (OGD) movements in an attempt to present access to government information to maximize the dissemination of government information. In terms of policy framework research, Bertot JC[13] based on the analysis of the U.S. policy framework,sorts out the problems of government data governance and proposes corresponding principles; Khatri V[14] analyzes the overall framework of data governance, and points out that data governance is the integration of policy making, process management, technology and responsibilities.

In summary, the existing domestic and international studies mainly focus ongovernment data governance policy introduction, policy quantification, policycomparison and policy framework analysis studies. It can be found that the current research on government data governance policy themes needs to be improved. Policy themes are the core content of policy texts and the focus of policy evolution research[15], and few scholars have used theme mining techniquesto study the clustering of government data governance policy themes. Therefore,this paper takes 443 government data governance policies in China since 2002-2021 as the object, and uses LDA model and theme similarity technique to perform theme clustering and hotspot clustering, so as to improve the comprehensive and targeted analysis in the field of government data governance policies in China, with a view to providing corresponding reference for the field.

3. Research Process and Methodology

3.1 Research Ideas

This paper uses the keyword "data governance" to search in the website www.pkulaw.com and the official government website as the data source, mainly including programs, decisions, opinions, measures, plans, standards and other documents promulgated by the government at the provincial and ministerial levels in China from 2002 to 2021. The research process starts from three aspects: data collection and data pre-processing, LDA topic mining, and topic hotspot clustering. Based on the time dimension, the policy texts are divided into budding, exploring and developing periods, and the policy topics under different periods are mined by LDA, and the similarity values between topics are calculated, so as to cluster the topic hotspots in each period. The research framework is shown in Figure 1.

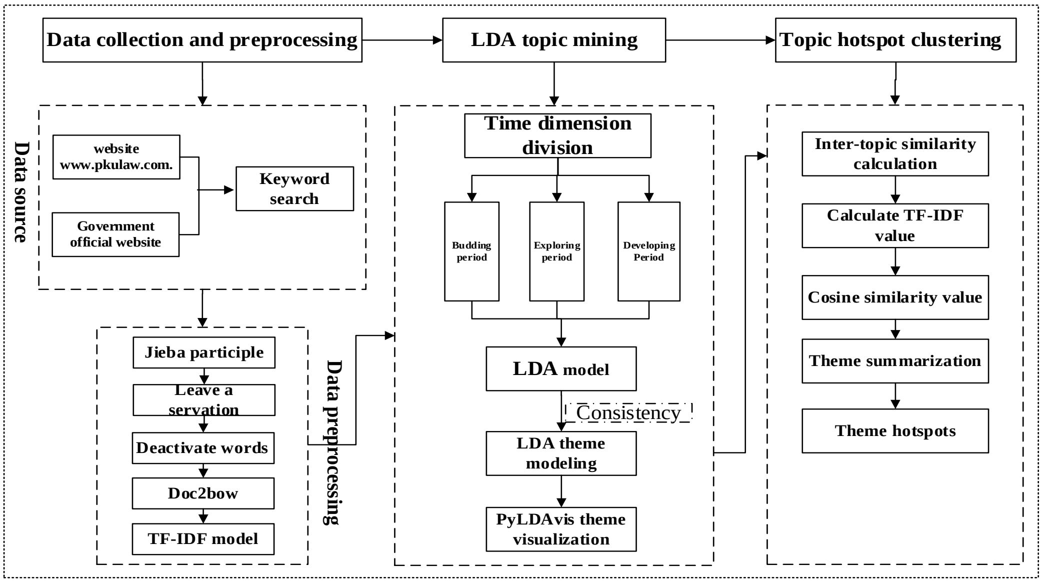

Figure 1: Research framework diagram

3.2 Research Methodology

3.2.1 Data Pre-Processing Methods

The text pre-processing is divided into three stages, first, the text initial screening stage. Based on the website www.pkulaw.com and government official website to collect government data governance policies, remove unnecessary documents such as approvals, letters, replies, announcements. Secondly, the split-word coding and re-screening stage. First, using jieba policy text for word separation, establish reserved words word list, and also select Chinese, Baidu, Harbin University, Sichuan University Machine Intelligence Laboratory deactivation word list thesaurus to achieve deactivation word processing; furthermore, delete words with length less than 2[16], and use gensim.corpora.dictionary to encode the result after word separation to ensure that each word corresponds to unique code. Third, the keywords are extracted and transformed into word vectors. The words are converted into word vectors using the doc2bow model in the encoded dictionary.doc2bow, and finally the doc2bow word vectors are converted into TF-IDF vectors using the TfidfModel model in gensim.models, in preparation for LDA topic clustering.

3.2.2 Subject Extraction Method

(1) LDA model

LDA model is a document topic generation model that contains a three-layer Bayesian structure of text-topic-word to achieve topic and vocabulary generation in documents, and is also an unsupervised machine learning technique that can identify latent topic information in large-scale corpora [17]. And LDA can effectively analyze large-scale unstructured document sets [18], which has great advantages for policy text topic research. Therefore, the LDA topic model is used for topic extraction in this paper, and the LDA model research process is as follows: data preprocessing→ optimal topic number determination→building topic model→obtaining topic words→performing topic induction.

(2) Optimal theme number selection

There is no unified guideline on the selection of the optimal number of topics for LDA, and topic confusion and topic consistency scores provide two metrics to discern the degree of merit of LDA topic models. The topic consistency value refers to whether the high probability words corresponding to each topic generated by the model are semantically consistent, and a higher value indicates a better model, and the topic consistency index score can better judge the model strengths and weaknesses[19,20]. Therefore, in this paper, we choose topic consistency to determine the optimal number of topics.,take 1-15as the range of topic number, and call the get_Coherence method of LDA topic clustering model to calculate the consistency value, then the peak consistency corresponds to the number of topics as the optimal number of topics.

3.3 Analysis of Research Subjects

3.3.1 Policy Text Sources

The data were obtained from website www.pkulaw.com and government official websites, and were selected according to the following three criteria: ① the subjects of policy issuance were central ministries and commissions and 31 provinces, municipalities and autonomous regions (excluding Hong Kong, Macao and Taiwan); ② the data governance policies promulgated at the provincial and ministerial level or above in China during the 20 years from 2002 to 2021 were selected, and the titles and contents were filtered according to the keywords "data governance" and "governance data"; ③ remove unnecessary documents such as approvals, letters, replies and announcements. In this paper, a total of 443 policy texts were collected after the initial screening, which were mainly temporary regulations, programs, decisions, opinions, approaches, plans, standards and other policy documents promulgated by the government at the provincial and ministerial levels. Among them, 51 are central-level policies and 392 are from provinces, municipalities directly under the Central Government and autonomous regions.

3.3.2 Policy Phasing

The current classification of policy phases under time windows can be divided into: judgment based on landmark events [21], based on national standards and related reports [22], and based on quantitative development trends in the time dimension [23]. Referring to existing studies, this paper classifies the development of government data governance policy themes from 2002-2021 into the nascent, exploratory, and developmental periods.

4. Government Data Governance Policy Theme Hot Topics Mining

4.1 Optimal Number of Topics Determination

The clustering training is performed using the LdaModel model in the Gensim library, and the consistency values corresponding to different numbers of training topics are derived by the get_Coherence method in the training process, and the consistency peaks correspond to the optimal number of topics, and the optimal numbers of topics are finally selected as 5, 8, and 10 for the budding, exploring, and developing periods, as shown in Figure 2.

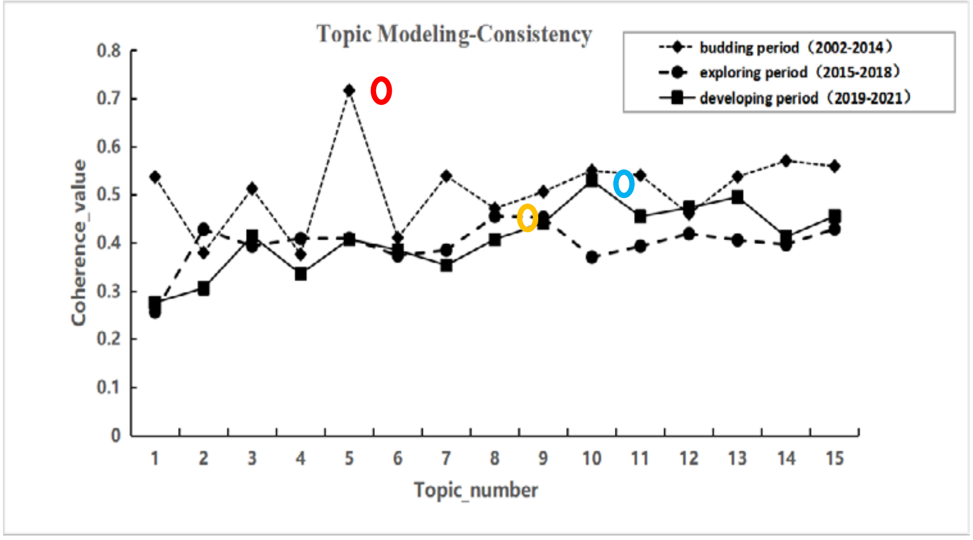

Figure 2: Theme consistency scores

4.2 Theme Classification Effect

In this paper, using gensim's pyLDAvis visualization tool [24], it is possible to obtain visual classification effect maps for three periods, to control the output LDA topic words and to perform topic clustering. Among them, the adjustment parameter 1 is selected, the distribution of circles on the left represents different topics, the size of the circles represents the number of policies containing that topic, and the distance between the circles represents the correlation between topics.

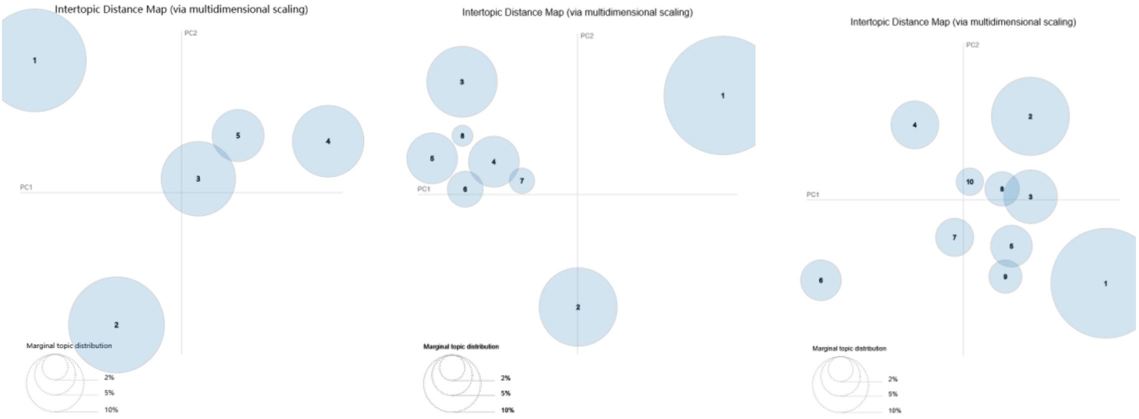

Figure 3: Topic clustering visualization

From the above visualization result graph (see Figure 3), it can be seen that there is cross-fertilization but relatively independent of each theme within each period, based on the open data environment, government data are rapidly transmitted and effectively integrated, and government big data rapidly develops and penetrates into various industries, which makes data and various fields deeply integrated to release the maximum data value. At the same time, the recurrence of characteristic words under different themes fully demonstrates the importance of data governance in each field, which indicates that the theme classification is effective.

4.3 Topic Hotspot Clustering

Based on LDA topic clustering and topic visualization, to improve the accuracy of topic naming, this paper selects the first 10 to 15 keywords under each topic category as the main basis of topic induction, and other remaining words as the secondary auxiliary of topic induction to get the topic-word distribution of each period, and inductively get the final topic. Then, according to the similarity between themes and combined with the focus of theme induction results to get the theme hotspots in the corresponding period.

4.3.1 Budding Period Theme Hotspots

Table 1: Distribution of themes in the budding period (2002-2014)

Time stage	Topic summarization	Subject words
Budding period (2002-2014)	Data standards	Internet, Definitions, Data standards, Technical support, Refurbishment, Database, Modification, Website, Functio-nality, Development, Code
	Banking industrydata quality	Banking industry, Data quality, On-site, Audit, Financial institutions, Regulators, Indicators, Anomalies, Risk, Changes
	Data validation	Computer information, Certification, Testing, Users, Citizens, International, Institutions, Violations, Statistics, Collection, By Law
	Government inf-ormation sharing	Government information, System,Exchange, Collection, Sharing, Public, Directory, Infrastructure, Encourage, encourage
	Banking risk management	Risk, Finance, Lending, Commercial banks, Financial institutions, Promotion, Transaction, Platform, Enhance, Encourage

As shown in Table 1, the number of budding policy texts is 14 and the number of themes is 5. After integrating similar themes, focus on the following thematic hotspots. ① Data management guidelines: Theme 1 (data standards), Theme 3 (data validation). ② Banking data management: Theme 2 (Banking industry data quality), Theme 5 (Banking risk management). ③Government information sharing: Theme 4 (government information sharing).

4.3.2 Exploring Period Theme Hotspots

As shown in Table 2, The number of policy texts in the exploration period is 68 and the number of themes is 8. After integrating similarity themes, the focus is on the following thematic hotspots: ① Artificial intelligence technology: Theme 2 (Artificial intelligence powers data governance), Theme 8 (banking industry and government services intelligence). ②Digitalization of government services: Theme 1 (data resource platform), Theme 5 (healthcare), Theme 6 (platformization of government services). ③digitalization of industry: Theme 3 (digital financial pilot area), Theme 4 (Industry digitization), Theme 7 (digital transformation of the logistics industry).

Table 2: Distribution of themes in the exploring period (2015-2018)

Time stage	Topic summarization	Subject words
Exploring period (2015-2018)	Data resource Pl-atform	Public resources, Acquisition, E-Commerce, Risk management, Trading platforms, Operator, Cross border, Institutions, Statistics, Data applications, Cloud platforms
	Artificial intellige-nce powers data governance	Command, Artificial intelligence, Finance, Smart manufacturing, Logistics, Agriculture, Legal, Data management, Transportation, Technology industry, Open Sharing
	Digital finance P-ilot zone	Experimental zone, digital, finance, Credit, Website, Intensification, Financial sector, Government services, Command, Market players, Credit collection
	Industry digitization	Operators, Engineering projects, Construction industry, Highways, Husbandry, Information centers, Animal husbandr, Public data, Transportation
	Healthcare	Government information, Health care, E-government, Resource sharing, Statistics, Interim measures, Information industry data sharing, Government services, Urban, Fitness, Medicine
	Platformization ofgovernment servi-ces	Government Services, Credit, Public announcement, E-commerce, E-government, Cloud platform, Top level design, Rule of law, Rewards and punishment, Electronic seal, Electronic license, Encryption
	Digital transform-ation of the logi-stics industry	Logistics, Public data, Digital transformation, Logistics parks, Intermodal transport, Dumping transport, Commissioning, Supply chain, Freight, Transport
	Banking industry and Government Services Intellige-nce	Banking, Government Services, Artificial Intelligence, Higher Education, Health Care, E-Government, Administration, Financial Institutions, International, Schools, Information Technology

4.3.3 Developing Period Theme Hotspots

As shown in Table 3, The number of policy texts in the development period is 361 and the number of themes is 10. After consolidating similar themes, the focus is on the following thematic hotspots: ① Blockchain technology: Theme 1 (blockchain enabled data governance), Theme 6 (one network for all). ②Digital transformation: Theme 4 (digital economy), Theme 7 (industry digital transformation), Theme 10 (digital government construction). ③Government big data management: Theme 3 (medical institutionsand industrial big data), Theme 5 (government data), Theme 8 (public data application) ④Data security management: Theme 2 (data security), Theme 9 (traffic data security management).

From the clustering results of government data governance policy theme hotspots, the theme hotspots include artificial intelligence technology, digital industrialization, blockchain technology, digital transformation, etc., which are consistent with the relevant plans such as "widely applying digital technology to government management services", promoting industrial "digital transformation", "building new advantages of digital economy", and accelerating "digital industrialization" mentioned in the outline of the 14th Five-Year Plan and 2035 Vision released in March 2021[25], which also proves that the trend of policy theme hotspots in this paper is consistent with the relevant plans and national strategies in China.

Table 3: Distribution of themes in the developing period (2019-2021)

Time stage	Topic summarization	Subject words
Developing period (2019-2021)	Blockchain enabled data governance	Information center, Data center, Blockchain, Geohazard, One network for all, Green, Management office, Trial, Commodity, Spatial planning
	Data security	Public data, Data centers, Data security, Big data industry, Engineering projects, Tourism, Construction, Disasters, Digital economy, Rule of law, Rescue
	Medical institutions and industrial big data	Medical institutions, Industry big data, Disease types, One network for all, Score, Payment, Agriculture and rural areas, Civil defense, Indicators, Processing, Public hospitals, Career
	Digital economy	E-licensing, Digital economy, Engineering projects, Public data, Tourism, Websites, Mineral resources, Public resources, Disasters, Transactions, Functional departments
	Government Data	Campus, Medical institutions, Industrial big data, Government data, Restoration, Arable land, Net Signature, Public data, Education Bureau, Economic and social, Land, Payment
	One network for all	One network for all, Cross-border, Help, Chief data officer, Processing, Electronic license, Water supply, Website, Digital transformation, Pension, Hospital, Material reserve
	Industry digital transformation	Finance, Manufacturing, Financing, Lending, Public data, Digital transformation, Communities, Smart city, Registration, Digital economy, Transaction, Processing, Financial services, Private enterprises
	Public data applicat-ion	Public data, Data center, One network for all, Services industry, Material reserves, Banking industry, Earthquake prevention, Education field, Processing, Business management, Data quality, Financing
	Traffic data securitymanagement	Transportation, Transportation, traffic, Data security, Engineering projects, firefighting, Public data, Law enforcement, Rescue, Management, Command, Rule of law
	Digital governmentconstruction	Digital transformation, Government information, Solutions, Industrial big data, Health care, E-commerce, Digital economy, Cloud platform, financial industry, Construction, Internet of things, Transportation, Big data industry, Manufacturing, Talent, electronic license

5. Conclusions

Based on LDA theme mining and pyEchart visualization tool, this paper conducts theme hotspot mining on government data governance policies above provincial and ministerial levels in China, which can provide some reference for government departments' policy decisions and extend research ideas for scholars in the field of data governance. However, this paper has the following shortcomings: firstly, the data sources are selected from government data governance policies at the provincial and ministerial levels in China, but not from municipalities, which may be less relevant to the corresponding topics at the micro level; secondly, the focus of topic hotness is based on the similarity between subjects, which lacks actual measurement data and may affect the accuracy of summarizing topic hotness. The next step will be to improve the research object, add municipal cities, and introduce hotness and novelty indicators to accurately measure the thematic hotspots.

References

[1] Chen Zhenghan. A review of China's government data governance research in the last decade[J]. Social Science Dynamics, 2022(02):59-67.

[2] Liang Y, Li Xiaoxang, Liu ZHENG, Zheng YIPING. Research on the core elements and cultivation paths of China's government data governance talent capacity [J]. Library, 2022(04):34-41.

[3] Zheng Lei. The two roles of government in data governance: policy makers and data users[J]. Exploration and Controversy, 2020(11):21-23.

[4] Niu Lixue, Bai Xiangyang. Research progress and future research trends of government data governance at home and abroad[J]. Hebei Science and Technology Tu Yuan, 2020,33(01):21-27.DOI:10.13897/j.cnki.hbkjty. 2020.0004.

[5] Li Chongzhao, Huang Juan. Policy and governance structure of government data governance in the UK[J]. E-Government,2019(01):20-31.DOI:10.16582/j.cnki.dzzw.2019.01.003.

[6] Zhou Wenhong, Wu Qiong, Tian Xin, Zhong Ruiling. A study on the practical framework of data governance in the U.S. federal government--a policy-based analysis and insights [J]. Modern Intelligence, 2022,42(08):127-135.

[7] Liu, B.F., Wei, W., An, S.M. Policy analysis of government data governance in the era of big data [J]. Journal of Intelligence, 2019, 38(01):142-147+141.

[8] Song Yi, An Xiaomi, Ma Guanghui. A study on big data governance capacity of US, UK and Australian governments - a content analysis based on big data policies [J]. Intelligence Information Work,2018(01):12-20.

[9] Yao KAM, Xia ZJ. A study on the practice under government big data governance system--a comparative analysis based on Shanghai, Beijing and Shenzhen [J]. Intelligence Data Work, 2020, 41(01):94-101.

[10] Liang C, Wu Peiyi. International comparison of data governance policies: history, characteristics and insights[J]. Science and Technology Herald, 2020, 38(05):36-41.

[11] Napoli P M, Karaganis J. On making public policy with publicly a vailable data: Thecase of US communications policymaking [J]. Government Information Quarterly, 2010, 27(4): 384-391.

[12] Yannoukakou A, Araka I. Access to government information: Right to information and open government data synergy [J]. Procedia-Social and Behavioral Sciences, 2014, 147:332-340.

[13] Bertot JC, Zheng Lei, Xu Huina, Bao Linda. Policy framework for big data and open data: issues, policies and recommendations [J]. E-Government, 2014 (1): 6-14.

[14] Khatri V, Brown C V. Designing data governance[J]. Communications of the ACM, 2010, 53(1): 148-152.

[15] Huabin, Kang, Yue, Fan, Linhao. Research on the characteristics and evolution of the hierarchical nature of Chinese high-tech industry policies--an analysis based on 6043 policy texts from 1991-2020[J]. Science and Technology Management,2022,43(01):87-106.

[16] Cai Wei, Ding Jingda, Liu Chao, Chen Yifan. Comparative analysis of data governance research themes identification and evolution at home and abroad [J]. Intelligence Inquiry, 2022(07):102-109.

[17] Li Zibiao, Zhang Li. Research on the evolution of patent technology themes of steel materials based on LDA model [J]. Science and Technology Management Research, 2020, 40(24):175-183.

[18] Zhou Jian, Zhang J, Qu Ran, Yan Shi. Analysis of domestic and foreign blockchain theme mining and evolution based on LDA [J]. Journal of Intelligence, 2021, 40(09):161-169.

[19] Liu Shiyang, Hua Bolin. LDA-based public culture theme extraction and evolution analysis [J]. Library Intelligence Research, 2021,14(02):28-37.

[20] Cao Chen, Luo Qiansheng, Huang Jun, Sui Daliang, Xiao Zhan. Analysis of the current situation of science and technology innovation cooperation in the twin-city economic circle of Chengdu and Chongqing regions--based on social network and LDA theme model[J]. Soft Science,2022,36(01):98-107.DOI:10.13956/j.ss.1001-8409.2022.01.15.

[21] Tan, Chunhui, Xiong, Mengyuan. A comparative analysis of the evolution of domestic and foreign data mining research hot topics based on LDA model[J]. Intelligence Science, 2021, 39(04): 174-185. DOI: 10.13833/j. issn. 1007-7634.2021.04.023.

[22] Zhang Tao, Ma Haiqun. Analysis of big data policy themes and development trends in China [J]. Intelligence Theory and Practice, 2022, 45(03):72-80. DOI:10.16353/j.cnki.1000-7490.2022.03.011.
[23] Zhang Tao, Ma Haiqun. Analysis of hot topics and evolution of artificial intelligence policy in China[J]. Modern Intelligence, 2021, 41(11):150-160.
[24] Wan Yan, Zhang Minghui, Gao Jinping. A study on policies related to artificial intelligence in China based on text analysis [J]. Journal of Library and Information Studies, 2021, 6(12):54-63.
[25] Huang Qifan. The key stage of the great rejuvenation -- understanding and experience of learning the Fourteenth Five Year Plan for National Economic and Social Development of the People's Republic of China and the Outline of the Vision Goals for 2035 [J]. People's Forum, 2021 (15): 5.