Topic Distribution of China's Data Governance Policies: A Full-text Highlighted Clue Word Approach

Bikun Chen [1], Yuxin Liu [2] and Kuan Bai [2]

[1] *Soochow University, No. 199 Renai Road, Suzhou, Jiangsu, China*
[2] *Nanjing University of Science and Technology, No. 200 Xiaolingwei Street, Suzhou, Jiangsu, China*

Abstract

In order to expand application scopes of full-text approaches and enrich topic analysis of policy documents, this paper comprehensively collects China's data governance policy texts as sample data and applied statistical analysis and text mining to extract and analyze the full-text highlighted clue word distribution of different data governance policies. It is found that the objects of data governance policies have been evolving to various kinds, and the scope of its coverage has expanded. Data governance policies have shown three tendencies: more attention paid to the protection of relevant subjects' rights, more effort at data development and utilization, and more emphasis on the role of data in socio-economic development.

Keywords

Data policy, data governance, topic distribution, highlighted clue word, full-text mining

1. Introduction

The increasing availability of full text from scientific articles in machine readable electronic formats is an opportunity to greatly impact scientometrics. In-text citations and entity metrics are typical examples of full-text analysis in scientometrics [1]. Full-text analysis in scientometrics mainly focused on the following topics. (1) Firstly, in-text citations, such as reference position [2-3], proximity of cited references [4-6], citation contexts or sentiments [7-10], and citation motivation or behavior [11-14]. (2) Secondly, entity metrics, such as concepts [1, 15], datasets [16], software [17-19] and algorithms [20]. (3) Finally, linguistic complexity of scientific writing styles and scientific impacts [21-22], characteristics of a highly cited article [23] and highly viewed or downloaded article [24]. Most above researches focus on academic articles, whether their full-text approaches can be extended or properly applied to other full-text literature sources is a critical topic for their robustness and flexibility.

Quantitative analysis of policy literature (especially the scientific technology and academic policies) is a hot research topic in scientometrics and public management field, which aims to explore policy topics, intentions, evolutions or relationships among government entities and can be summarized into three categories: (1) Bibliometrics-based analysis of policy documents, such as mentions of academic papers in policy documents [25]; (2) Basic statistics of policy elements, such as temporal and spatial distribution [26-27], policy instruments [28-29], linguistic characteristics [30] or topic analysis [31]; (3) Network analysis of policy elements, such as networks of policymaker [32] or joint policy-issuing network [33]. Most above researches applied classical bibliometrics or network methods to analyze external or content attributes of policy documents, but their quantitative approaches needed to be enriched or expanded, for example, topic analysis of policy documents are mainly implemented by global keyword

3rd Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2022), June 20-24, 2022, Cologne, Germany and Online
EMAIL: bkchen@suda.edu.cn (Bikun Chen);283958279@qq.com (Yuxin Liu); 1772217791@qq.com (KuanBai)

frequency or keyword co-occurrence, which needed to be conducted in fine-grained or detailed level.

In order to expand application scopes of full-text approaches and enrich topic analysis of policy documents, this paper applied full-text highlighted clue word approach to analyze topic distribution and evolution of China's data governance policies. China's data governance policies are selected based on the following reasons: Data is the new oil and a growing number of countries have issued data governance policies to support data development and utilization. Detecting the topic distribution of policies can provide insights for the improvement and innovation of policies at present and in the future. As an important part in the globe, governments at all levels in China has issued a chain of data governance policies. At present, few studies have comprehensively analyzed topic distribution of these policies in China.

Specifically, some critical questions are needed to be addressed in data governance policies: (1) Whether the relevant subjects have data rights or legal rights (Who); (2) To what extent can the data (the relevant objects) be utilized (What and How); (3) What is the purpose of utilizing the data (Why).

2. Data and method

2.1. Data

Data governance policies cover a broad scope, ranging from government data, public data, industry data, big data and so on. In this study, data governance policies are acquired by searching a series of data governance related keywords (eg. government data, public data, industry data, big data, government information, public information, social information, digital regulation, data regulation and so on) in PKULaw Database (https://www.pkulaw.com), general search engine (eg. Google, Baidu and Bing) and academic literatures. Then, every crawled policy is manually read and evaluated by five experienced domain experts. Finally, 258 policy documents are kept as the sample.

Compare with academic literatures, policies are usually issued by governments or other organizations with limited number but without organization-assigned keywords. While, they both are structured with basic metadata, such as title, time, location, main body and so on.

2.2. Method

2.2.1. Classification of policy documents

Every policy is manually read and evaluated by five experienced domain experts in sample data collection. It is found that policies titled "government information publicity" are firstly issued, then policies titled "government data publicity", "public information or data publicity", "big data development" and "certain industry data development" (eg. scientific data and transportation data) are released by governments at all levels in China. Therefore, based on the policy titles and trajectory of policy issuance, data governance policies in this study are classified into four categories: government data, public data, industry data and big data, shown in Table 1.

Table 1
Classification of policy documents

Type	# of policy documents
government data policy	129
public data policy	31
industry data policy	79
big data policy	19

2.2.2. Extraction of policy elements

This study is mainly based on two policy elements: policy-issuing time and location of policy-issuing agency. Policy-issuing time is the specific time when a policy was released to the public; location of policy-issuing agency refers to China' provinces where the government department that formulated and released the policy located. Policy-issuing time and location are the metadata listed in the policy document and can be directly extracted. Temporal and spatial distribution of different data governance policies are shown in Figure 1.

2.2.3. Extraction of full-text highlighted clue words (FHCW)

Inspired by entity metrics [1], full-text highlighted clue words are proposed to solve the

research questions and defined as follows: they are a series of notional words in full text of literatures with certain logic (eg. "subjective-objective", "theory-method-application", "structural-dynamical") and significance (eg. representation of sentiments, motivation, behavior or scenario), primarily selected by experienced domain experts. Traditional term frequency, n-gram and co-word analysis usually focus on author-assigned, database-assigned or bibliography-extracted keywords with high frequency, TF-IDF method focus on unique term in each document, while FHCW approach emphasize on notional words in full text with any frequency, no matter how common or unique. Besides, FHCW approach is similar to expert content analysis because notional words are selected and arranged logically by experienced domain experts, but FHCW are automatically extracted by software (or programming language) and expert content analysis are usually conducted by manually reading and coding.

In this study, FHCW are selected by five experienced domain experts in terms of "policy subjects, policy objects and application scenarios" logic and orderly arranged on the basis of rights of policy subjects (mainly from low intensity to high intensity), openness degree of policy objects (from low degree to high degree) and specific application scenarios (from specific to general), shown in Table 2. Five experienced domain experts specializes in government information policy research and practice. Operatively, FHCW in each policy document are extracted by quanteda package in R language [34].

In order to evaluate the advantage of FHCW approach, comparative experiments between FHCW approach and traditional global keywords analysis (unigram, bigram and TF-IDF methods) are also conducted, shown in Table 4. For TF-IDF method, top10 keywords in each policy document are extracted and then aggregated globally.

Table 2

Selection of full-text highlighted clue words

Type	Highlighted clue words
rights of policy subjects	reserve the right; confirm the right; authorization; rights; legal rights
openness degree of policy objects	share; openness; development; utilization
application scenarios	data security; data assets; digital economy; digital government; digital society

2.2.4. Quantitative analysis of FHCW

Firstly, FHCW in every policy document are counted and aggregated into each policy type, year and province in China. Then, in order to reduce the influence caused by the unbalanced number in each policy type, year and province, mean value of every FHCW are calculated by dividing the total number of policy documents in each group. Mean value are also normalized between each group (horizontal level) and in each group (vertical level), shown in Figure 2, 3 and 4 (implemented by tidyverse package in R language). Finally, co-occurrence network of FHCW is constructed based on the co-occurrence relationships in each policy document by Gephi software, shown in Figure 5 and Table 3.

3. Results

3.1. Temporal and spatial distribution of different data governance policies

From Figure 1a, it is shown that policies of big data and public data is overall increasing after 2017, but policies of government data and industry data is overall decreasing after 2017. From Figure 1b, it is shown that governments in Shandong, Beijing (provincial level municipality directly under the central government) and Zhejiang issued the maximum number of policies, then Jiangsu, Guangdong and Fujian, Qinghai and Tibet are the least.

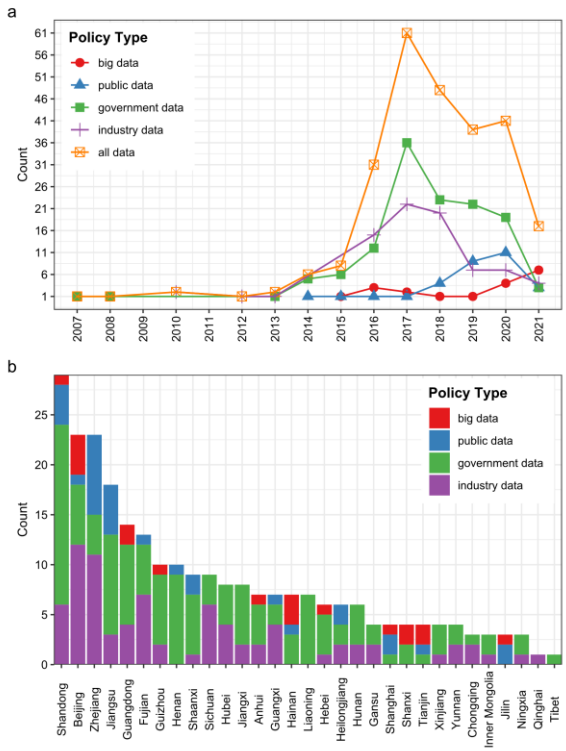

Figure 1: Temporal (a) and spatial (b) distribution of different data governance policies

3.2. Temporal distribution of FHCW

From Figure 2a, it is shown that: (1) in FHCW concerning rights of policy subjects, "reserve the right" is proposed the earliest (in 2008) and shows "increase-decrease-increase" trajectory, "authorization" is also proposed the earliest (in 2008) and shows overall growing trend, "confirm the right", "rights" and "legal rights" are mentioned relatively late and indicate overall growing trend; (2) in FHCW concerning openness degree of policy objects, "share" is proposed the earliest and shows "increase-decrease" trajectory but "openness", "development" and "utilization" are mentioned relatively late and indicate overall growing trend; (3) in FHCW concerning application scenarios, "data security", "data assets", "digital economy", "digital government" and "digital society" appear successively and indicate overall growing trend.

From Figure 2b, it is shown that: (1) "share" is mostly mentioned in almost every year, then "openness", "utilization" and "data security"; (2) "share" is solely accounting for a large proportion in early years but proportions of "openness", "utilization" and "data security" are gradually increasing in recent years.

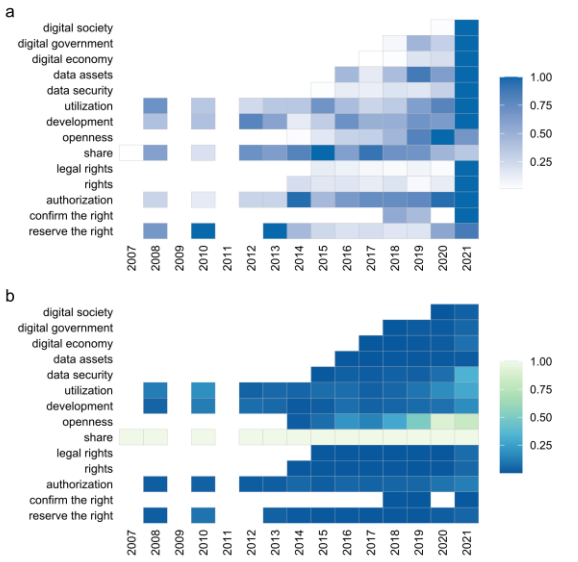

Figure 2: Temporal distribution of FHCW (a. normalized between any year; b. normalized in any year)

3.3. Spatial distribution of FHCW

From Figure 3a, it is shown that: (1) in FHCW concerning rights of policy subjects, "authorization" is mentioned in almost every province, then "rights" and "reserve the right", "legal rights" and "confirm the right" are the least; (2) in FHCW concerning openness degree of policy objects, "share" is mentioned in almost every province, then "development", "utilization" and "openness"; (3) in FHCW concerning application scenarios, "data security" is mentioned in most provinces, then "data assets" and "digital economy", "digital government" and "digital society" are the least. On the whole, provinces in east China (eg. Tianjin, Shanghai, Shandong, Beijing, Guangdong, Zhejiang and Jiangsu) mention the majority of FHCW, introduce the emerging FHCW (eg. confirm the right, digital government and digital society) and shift their focus from "share" to "development" and "utilization".

From Figure 3b, it is shown that "share" is mostly mentioned in most provinces. On the whole, compared with other provinces, proportions of "openness", "utilization", "development" and "data security" is increasing in provinces of east China.

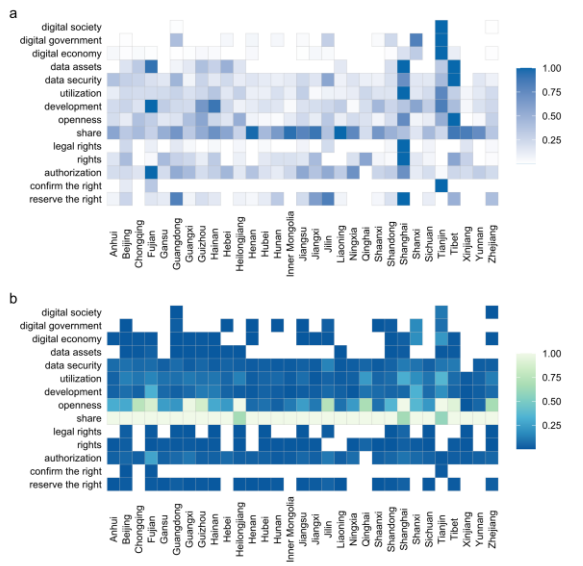

Figure 3: Spatial distribution of FHCW (a. normalized between any province; b. normalized in any province)

3.4. FHCW distribution of different data governance policies

From Figure 4a, it is shown that: (1) in FHCW concerning rights of policy subjects, big data, public data and industry data policies mentioned all FHCW, then government data, big data policies underlined "reserve the right", "confirm the right", "rights" and "legal rights" more than other policies, government data mentioned "authorization" more than other policies; (2) in FHCW concerning openness degree of policy objects, government data mentioned "share" more than other policies, public data policies mentioned "openness" and "utilization" more than other policies, big data policies mentioned "development" more than other policies; (3) in FHCW concerning application scenarios, big data policies mentioned all FHCW more than other policies.

From Figure 4b, it is shown that: big data policies mentioned "share" and "openness" more, then "data security", "utilization" and "development"; public data policies mentioned "openness" and "share" more, then "utilization"; industry data policies mentioned "share" and "openness" more, then "utilization", but did not mention "digital government" and "digital society"; government data policies mentioned "share" and "openness" more but did not mention "confirm the right" and "digital society".

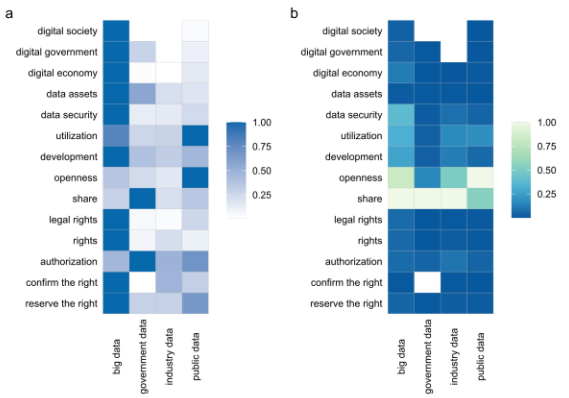

Figure 4: FHCW distribution of different data governance policies (a. normalized between any kind of data governance policies; b. normalized in any kind of data governance policies)

3.5. Global co-occurrence network of FHCW

In Figure 5, network is drawn by force-directed layout algorithm, each node indicates a FHCW and node size denotes weighted degree, lines between any two nodes means their co-occurrence value, node label is proportional to its node size, node color means different types of FHCW. From Figure 5 and Table 3, it is found that "share" and "openness" are placed in the center of the network with the maximum value; "utilization", "data security", "development" and "authorization" are placed near the center of the network with the lower value than "share" and "openness"; "confirm the right" and "digital society" are placed in the border of the network with the lowest value.

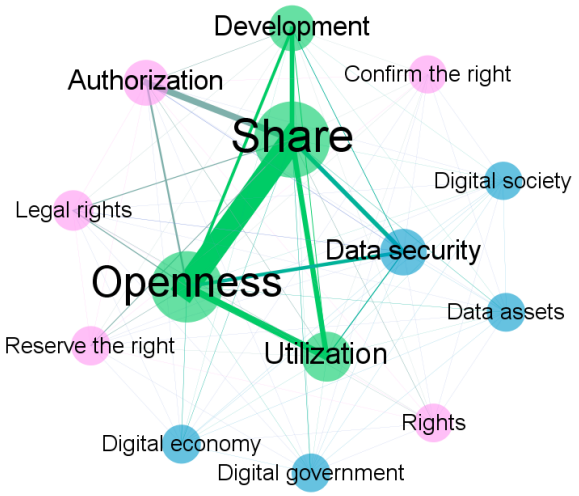

Figure 5: Global co-occurrence network of FHCW (pink: rights of policy subjects; green: openness

degree of policy objects; blue: application scenarios)

Table 3

Top 10 normalized co-occurrence value of node pairs

Node pairs	Normalized co-occurrence value
share - openness	0.333
utilization - openness	0.095
share - authorization	0.09
share - utilization	0.074
share - development	0.065
share - data security	0.062
data security - openness	0.051
development - openness	0.042
authorization - openness	0.027
openness - legal rights	0.015

4. Comparative Experiments

Limited by required pages, traditional keywords analysis in global level (without dividing policy documents into different temporal and spatial groups) and FHCW Approach are compared. From Table 4, it is shown that unigram and bigram methods focus on fulltext-extracted keywords with high frequency, TF-IDF method focuses on unique term in each document. But it is difficult to seize the logic among the keywords. While FHCW approach emphasize on notional words arranged in certain logic in full text with any frequency, no matter how common or unique. It might be practical and traceable to explore policy topics and intentions.

5. Discussion and Conclusion

Based on the results above, it is concluded that: (1) Objects of data governance policies has been expanded from government data to industry data, public data and big data; (2) Concerning rights of policy subjects, policy orientation has shifted from data access to protection of data rights; (3) Concerning openness degree of policy objects, policy orientation has shifted from data share and openness to data development and utilization; (4) Concerning application scenarios, policy orientation has shifted from data-oriented governance to society-oriented governance, paying more and more attention to digital economy, digital government and digital society.

Limited by required pages, only temporal and spatial elements are extracted to explore FHCW and only traditional global keywords analysis (unigram, bigram, TF-IDF and co-occurrence methods) is conducted. In further study, more policy elements (eg. policy instruments and policymakers) and more advanced method (eg. word embedding and dynamic network analysis) will be introduced and compared.

Table 4

Comparison between top14 global keywords analysis and FHCW approach

Unigram	Bigram	TF-IDF	FCHW
share	healthcare	government section	reserve the right
information	information resources	government information resources	confirm the right
data	share of government information resources	sharing platform	authorization
department	medical big data	openness	rights
management	catalogue of government information resources	government data	legal rights
government information resources	scientific data	big data	share
operation	legal person	public data	openness
agency	perform duty	medical care	development
big data	public credit	health	utilization

construction	administrative region	administrative agency	data security
resources	agency of public administration & service	service agency	data assets
service	geographic space	open platform	digital economy
catalogue	national secret	leading group	digital government

6. ACKNOWLEDGMENTS

This article is supported by the National Social Science Foundation of China (No. 21BTQ013), Scientific Research Innovation Program for Jiangsu Graduate Student (No. KYCX21_0419 & No. SJCX21_0158), Research Project Team of Humanities and Social Sciences, Soochow University (No. 22XM0002) and Youth Interdisciplinary Research Team of Humanities and Social Sciences, Soochow University (No. 202205).

7. References

[1] Ding Y, Song M, Han J, et al. 2013. Entitymetrics: measuring the impact of entities. PLoS ONE, 8(8): e71416.

[2] Hu Z, Chen C and Liu Z. 2013. Where are citations located in the body of scientific articles? A study of the distributions of citation locations. Journal of Informetrics, 7(4): 887-896.

[3] Kevin B, Nees J E, Giovanni C, et al. 2018. Characterizing in-text citations in scientific articles: a large-scale analysis. Journal of Informetrics, 12(1): 59-73.

[4] Liu S and Chen C. 2012. The proximity of co-citation. Scientometrics, 91(2): 495-511.

[5] Kevin B, Henry S, Richard K. 2013. Improving the accuracy of co-citation clustering using full text. Journal of the American Society for Information Science and Technology, 64(9): 1759-1767.

[6] Kim H, JeongY, Song M. 2016. Content- and proximity-based author co-citation analysis using citation sentences. Journal of Informetrics, 10(4): 954-966.

[7] Henry S. 2011. Interpreting maps of science using citation context sentiments: a preliminary investigation. Scientometrics, 87(2): 373-388.

[8] Liu S and Chen C. 2013. The differences between latent topics in abstracts and citation contexts of citing papers. Journal of the American Society for Information Science and Technology, 64(3): 627-639.

[9] Ding Y, Zhang G, Tamy C, et al. 2014. Content-based citation analysis: the next generation of citation analysis. Journal of the Association for Information Science and Technology, 65(9): 1820-1833.

[10] Lu C, Ding Y and Zhang C. 2017. Understanding the impact change of a highly cited article: A content-based citation analysis. Scientometrics, 112(2): 927–945.

[11] Terrence B. 1986. Evidence of complex citer motivations. Journal of the Association for Information Science and Technology, 37(1): 34-36.

[12] Susan B, Howard S. 1991. Motivations for citation: a comparison of self citation and citation to others. Scientometrics, 21(2): 245-254.

[13] Donald C, Georgeann H. 2000. How can we investigate citation behavior? a study of reasons for citing literature in communication. Journal of the American Society for Information Science, 51(7): 635-645.

[14] Zhang C, Ding R, Wang Y. 2018. Using behavior and influence assessment of algorithms based on full-text academic articles. Journal of the China Society for Scientific and Technical Information, 37(12): 1175-1187. (in Chinese)

[15] Kathy M, Hal D, Snigdha C, et al. 2016. Predicting the impact of scientific concepts using full-text features. Journal of the Association for Information Science and Technology, 67(11): 2684-2696.

[16] Belter C. 2014. Measuring the value of research data: a citation analysis of oceanographic data sets. PLos One, 9(3): e92590.

[17] Pan X, Yan E, Wang Q, et al. 2015. Assessing the impact of software on science: A bootstrapped learning of software entities in full-text papers. Journal of Informetrics, 9(4): 860–871.

[18] Pan X, Yan E, Hua W. 2016. Disciplinary differences of software use and impact in

scientific literature. Scientometrics, 109 (3): 1-18.
[19] Pan X, Yan E, Cui M, et al. 2018. Examining the usage, citation, and diffusion patterns of bibliometric mapping software: a comparative study of three tools. Journal of Informetrics, 12(2): 481-493.
[20] Wang Y and Zhang C. 2018. Using full-text of research articles to analyze academic impact of algorithms. In Proceedings of the 13th International Conference (iConference 2018).
[21] Lu C, Bu Y, Wang J, et al. 2019. Examining scientific writing styles from the perspective of linguistic complexity. Journal of the Association for Information Science and Technology, 70(5): 462-475.
[22] Lu C, Bu Y, Dong X, et al. 2019. Analyzing linguistic complexity and scientific impact. Journal of Informetrics, 13(3): 817-829.
[23] Mohame E. 2019. Characteristics of a highly cited article: a machine learning perpective. IEEE Access, 7: 87977-87986.
[24] Chen B, Deng D, Zhong Z, et al. 2020. Exploring linguistic characteristics of highly browsed and downloaded academic articles. Scientometrics, 122(3): 1769–1790.
[25] Lutz B, Robin H, Werner M. 2016. Policy documents as sources for measuring societal impact: how often is climate change research mentioned in policy-related documents? Scientometrics, 109(3): 1477–1495.
[26] Quan W, Chen B, Shu F. 2017. Publish or impoverish: an investigation of the monetary reward system of science in China (1999-2016). Aslib Journal of Information Management, 69(5): 486-502.
[27] Shu F, Quan W, Chen B, et al. 2020. The role of Web of Science publications in China's tenure system. Scientometrics, 122(3): 1683–1695.
[28] Huang C, Yang C, Su J. 2018. Policy change analysis based on "policy target–policy instrument" patterns: a case study of china's nuclear energy policy. Scientometrics, 117: 1081-1114.
[29] Huang C, Yang C, Su J. 2021. Identifying core policy instruments based on structural holes: A case study of china's nuclear energy policy. Journal of Informetrics, 15(2): 101145.
[30] Michael L, Kenneth B, John G. 2003. Extracting policy positions from political texts using words as data. American Political Science Review, 97(2): 311–331.
[31] Huang C, Su J, Xie X, et al. 2014. Basic research is overshadowed by applied research in China: a policy perspective. Scientometrics, 99(3): 689–694.
[32] Yang C, Huang C, Sun J. 2020. A bibliometrics-based research framework for exploring policy evolution: A case study of China's information technology policies. Technological Forecasting and Social Change, 157: 120116.
[33] Yang C and Huang C. 2022. Quantitative mapping of the evolution of AI policy distribution, targets and focuses over three decades in China. Technological Forecasting and Social Change, 174: 121188.
[34] Benoit K, Kohei W, Wang H, et al. 2018. quanteda: An R package for the quantitative analysis of textual data. Journal of Open Source Software, 3(30): 774.