TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 03/10 pp680–691
DOI: 10.26599/TST.2021.9010059
Volume 27, Number 4, August 2022

PG-CODE: Latent Dirichlet Allocation Embedded Policy Knowledge Graph for Government Department Coordination

Yilin Kang, Renwei Ou, Yi Zhang, Hongling Li*, and Shasha Tian

Abstract: Government policy-group integration and policy-chain inference are significant to the execution of strategies in current Chinese society. Specifically, the coordination of hierarchical policies implemented among government departments is one of the key challenges to rural revitalization. In recent years, various well-established quantitative methods have been proposed to evaluate policy coordination, but the majority of these relied on manual analysis, which can lead to subjective results. Thus, in this paper, a novel approach called "policy knowledge graph for the coordination among the government departments" (PG-CODE) is proposed, which incorporates topic modeling into policy knowledge graphs. Similar to a knowledge graph, a policy knowledge graph uses a graph-structured data model to integrate policy discourse. With latent Dirichlet allocation embedding, a policy knowledge graph could capture the underlying topics of the policies. Furthermore, coordination strength and topic diffusion among hierarchical departments could be inferred from the PG-CODE, as it can provide a better representation of coordination within the policy space. We implemented and evaluated the PG-CODE in the field of rural innovation and entrepreneurship policy, and the results effectively demonstrate improved coordination among departments.

Key words: policy knowledge graph; department coordination; topic diffusion; latent dirichlet allocation; rural revitalization

1 Introduction

The rural revitalization strategy in China is firmly associated with the national economy and people's livelihoods. Rural innovation and entrepreneurship have attracted much policy attention[1] ever since the Chinese government proposed a new strategy of public entrepreneurship and innovation during the Summer World Economic Forum in 2014. Under this strategy, both central and local government policies have been guided by the "innovation-driven" and "rural revitalization" policies to constantly promote the new development of rural innovation and entrepreneurship with the transformation of labor and information technology from cities to rural areas.

At the same time, the rural policy advancement faces the dilemma of government department coordination as well as management and talent development[2]. On the one hand, there is an increase in the number of rural innovation and entrepreneurship entities, including scientific and technological personnel as well as Chinese migrant workers, to name a few[3]. On the other hand, the related departments may not have an adequate understanding of the coordination efforts involved in rural innovation and entrepreneurship policies because the latter has been operating in a fragmented manner for a long time[4]. A survey carried out by China Association for Science and Technology Third-Party Evaluation Task Force (2016) revealed fragmentation and isolation of

• Yilin Kang, Renwei Ou, Yi Zhang, and Shasha Tian are with the School of Computer Science, South-Central University for Nationalities, Wuhan 430074, China. E-mail: ylkang@mail.scuec.edu.cn; 1195821448@qq.com; 1031271461@qq.com; shashatian77@mail.scuec.edu.cn.
• Hongling Li is with the School of Public Management, South-Central University for Nationalities, Wuhan 430074, China. E-mail: lihl723@163.com.
* To whom correspondence should be addressed.
Manuscript received: 2021-05-15; revised: 2021-07-02; accepted: 2021-07-30

policies between the central and local governments during the introduction of the "Entrepreneurship and Innovation" policy, thus indicating insufficient inter-departmental coordination.

In light of such problems, it is necessary to organize the existing rural innovation and entrepreneurship policy system, with the aim of improving the coordination and linkage of the rural innovation and entrepreneurship policies and promoting the government policy-chain inference and policy-group integration.

In recent years, several quantitative methods have been proposed to evaluate policy coordination. For instance, Hughes et al.[5] adapted eight internationally recognized principles of good governance to determine the criteria of drug policy coordination, combined with expert scoring methods from stakeholders. Meanwhile, Peng et al.[6] provided specific operation manuals for policy quantitative standards to improve the evaluation of policy coordination. Although these well-established works have been completed for many years, several aspects hint at subtle subjectivities. Due to the obscure definition of entrepreneurship itself and the incomplete conventional datasets, most of the methods proposed to evaluate rural entrepreneurship policy lacked sufficient rigor.

Thus, in the current paper, we have proposed a novel latent Dirichlet allocation (LDA) embedded policy knowledge graph called "policy knowledge graph for the coordination among the government departments" (PG-CODE) for government policy-group integration and policy-chain inference. Similar to a knowledge graph, a policy knowledge graph is a policy base that uses a graph-structured data model to integrate policy discourses. The LDA is a topic model for data mining that may occur when we attempt to discover abstract topics from a collection of policy discourses. With policy topics and government departments as entities as well as the similarity among departments as weights, the PG-CODE is able to infer the coordination among government departments, thus providing a better representation of overall coordination within the policy space.

In the current study, we have implemented and evaluated the LDA embedded policy knowledge graph based on the rural innovation and entrepreneurship policy discourse from the central government (e.g., State Council) to the local governments (e.g., Hubei Provincial People's Government (HPPG)). Our evaluation results demonstrate the effective coordination among government departments.

The rest of the paper is organized as follows. In Section 2, we describe the related work in the field of policy. Section 3 presents the details of the PG-CODE and discusses in detail the definition of coordination strength (CS) and topic diffusion (TD) in detail. In Section 4, we present the results of our implementation and evaluation of the policy knowledge graph based on the policy discourse on rural entrepreneurship and innovation from 2014 to 2020 in China. Section 5 concludes and highlights future research directions.

The main contributions presented in this paper are as follows:

- We propose a novel policy knowledge graph with an embedded LDA topic model.
- We propose two computational metrics for evaluating coordination among departments: namely, CS and TD.

2 Related Works

2.1 Research framework for rural entrepreneurship

Developed countries, such as the members of the European Union[7] and the United States[8], as well as developing countries, such as India and Bangladesh, are undergoing rural policy reforms and structural adjustments in terms of rural entrepreneurship and innovation. Thus far, related studies have focused on a research framework based on the factors affecting rural entrepreneurship.

In particular, many studies have proposed different kinds of frameworks to theoretically analyze entrepreneurship in rural areas from a wide variety of disciplines, including "structuration theory" in sociology as well as "actors network theory"[9] and "culture economy" in economic geography. The proposed studies are mostly concerned about the entrepreneurial processes in rural areas and aimed to achieve the integration of entrepreneurial policy into the wider context of rural development policies and strategies. Stathopoulou et al.[7] presented a coherent framework of entrepreneurship in rural areas in Europe, depicting rural entrepreneurship as a three-stage sequential process that is greatly influenced by distinct territorial characteristics; in their work, they emphasized rurality as a dynamic entrepreneurial resource. Fully understanding the process will facilitate the effective design, delivery, and implementation of competent entrepreneurial policies in rural areas. Meanwhile, Goetz et al.[8] employed a

neo-classical entrepreneurial decision equation, pointing out that government policy can influence economic start-up activities especially in rural areas. However, due to the fuzzy definition of entrepreneurship itself and the incomplete conventional datasets, most of the proposed methods of evaluating rural entrepreneurship policy lacked sufficient rigor.

2.2 Evaluating policy coordination

Unlike the establishment of an analytical framework to carry out the the rural entrepreneurship process, some researchers have attempted to use econometric methods to carry out the qualitative analysis of policies. For example, Hughes et al.[5] adapted eight internationally recognized principles of good governance proposed by the Economic and Social Commission for Asia and the Pacific to determine the criteria for evaluating of drug policy coordination; they administered a pilot survey to 36 stakeholders from top Australian advisory bodies to score each criterion. Peng et al.[6] provided specific operation manuals for policy quantitative standards, considering policy entities, policy objectives, and policy measures as detailed criterion. By further subdividing the elements under each criterion, they took the method of expert scoring for quantitative evaluation. Combined with the quantitative measures of legal change in the policy contents proposed by Libecap[10], he then quantified the degree of coordination of policies and explored the coordinated evolution of policies.

Based on the aforementioned information, it can be seen that most scholars use qualitative analysis and expert scoring methods to quantify policy coordination. However, these may be subject to slight subjectivity with a low degree of automation. In addition, the existing research on the evaluation of policy coordination mostly examined policy entities at the same level; however, evaluating hierarchical policy coordination is equally important in gauging the effects of policy governance[11].

2.3 LDA and topic modeling

Topic modeling is widely used in natural language processing and text mining studies. LDA is one of the most popular methods in the field of topic modeling, as it facilitates in topic discovery and semantic mining from unordered documents[12].

First proposed by Blei et al.[12], LDA is a generative probabilistic model for the collection of discrete data, such as text corpora; it is also a topic model that is used for discovering abstract topics from a collection of documents. In the LDA model, each document is represented as a finite random mixture over latent topics, where a topic is characterized by a distribution over words, aiming at constructing the implicit "document-topic-word" relationship. Each word belongs to different topics with different probabilities and those words with the highest probabilities in each topic usually give a good idea of what the topic is[13].

Researchers have applied the LDA model in various fields such as political science[14–16], economics, software engineering, geography science, and so on. For example, in political science, Greene and Cross[16] analyzed political interactions in the European Parliament by detecting latent themes in legislative speeches over time based on the new two-layer matrix factorization methodology, their proposed LDA-based model can also be applied to other more traditional forms of political discourses.

We also investigated highly scholarly articles related to the LDA topic model in the field of policy. Using the LDA model, Shirota et al.[17] conducted topic extraction from meeting notes generated from January 2013 to June 2014. The extracted topics clearly showed the monetary policy of the Central Bank of Japan. To track the evolution of policy and development of new energy vehicles (NEVs) in China, Jia and Wu[18] investigated 5185 articles on NEV obtained from the China National Knowledge Infrastructure by means of LDA-based text mining. Zhao et al.[19] analyzed the topic evolution of the coordinated development of Beijing-Tianjin-Hebei Provinces by analyzing the 14 235 journal articles collected by China National Knowledge Infrastructure in the past ten years.

2.4 Knowledge and topic models

Recently, several knowledge-based topic models have been proposed. Xie et al.[20] built a Markov Random Field regularized the LDA model to incorporate the external word correlation knowledge, guided by the aim of improving the coherence of topic modeling. Yao et al.[21] incorporated human knowledge in the form of a probabilistic knowledge base named "Probase" into topic models to improve semantic coherence. Xie et al.[22] proposed a novel representation learning method for knowledge graphs, including continuous bag-of-words and deep convolutional neural models. To process fact-oriented triple knowledge in knowledge graphs, Yao et al.[23] proposed a novel knowledge-based topic model by incorporating knowledge graph embeddings into topic

modeling. Most of the relevant works have focused on embedding knowledge into topic models. In comparison, our work focused on incorporating topic modeling into a policy knowledge graph that is similar to a knowledge graph.

3 PG-CODE

Google first proposed the concept of a knowledge graph in 2012. The knowledge representation of the knowledge graph mainly describes the relationship among entities, and this can be especially helpful when conducting massive unstructured data analysis. At present, knowledge graphs are mostly stored and managed in the form of resource description framework triples. Similar to a knowledge graph, we propose in this paper a policy knowledge graph representing critical information from unstructured policy discourse as a form of policy knowledge. Specifically, the entities are departments and topics, and the relationships among entities represent the frequency of the department's attention to topics. Therefore, using a policy knowledge graph is helpful in effectively analyzing the coordination among policy-making departments.

3.1 Proposed framework

The schematic diagram of the PG-CODE is shown in Fig. 1. As illustrated in the diagram, we collected the relevant policy discourse from the official websites of the government departments, and preprocessed the collected policy discourses. This process included deleting the invalid contents, such as the title, stop words, punctuation marks, and index numbers at the beginning of each policy. Then key phrases were extracted from the standard policy discourse set, after which all policy discourses were transformed into bag-of-words models according to the key phrases. Before training the LDA model, the number of topics of the model (K) was determined. Afterward, we trained all bag-of-words models of discourses to obtain the LDA model for mining policy discourse topics. From such information, a policy knowledge graph could be built. In turn, using the generated policy knowledge graph, the coordination relations among departments of central and local governments could be inferred.

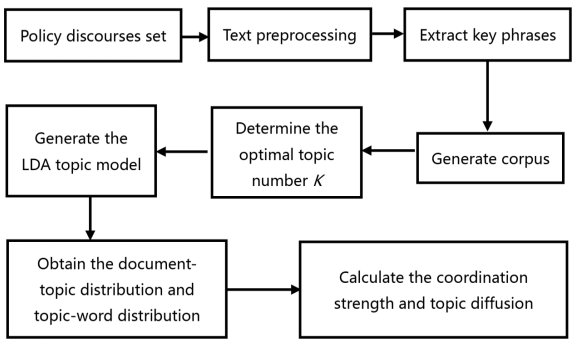

Fig. 1 Schematic diagram of the proposed PG-CODE.

3.2 LDA-embedded policy knowledge graph

One of the most popular methods to infer the topics of each policy discourse is to use LDA. An LDA topic model is a probabilistic topic model based on a three-layer Bayesian structure. The model assumes that the document has a probability distribution of potential topics, and that such topics represent the probability distribution of a series of words in the document set. Thus, we can infer the topics of each document by constructing the probability distribution of "document-topic" and "topic-word". The LDA model can be used to show that a document contains multiple topics with different probability values simultaneously and that each word belongs to a different topic with a different probability value. After analyzing some discourses and extracting their topic distributions, we can infer the topics of a policy discourse based on the topic distribution. In this paper, we used the LDA model to infer the topic of each policy discourse, thus allowing us to create topic entities in the policy knowledge graph.

3.3 Determining the optimal topic number K

Before constructing the LDA model, we need to preset the number of recognized topics based on the normalized corpus, which is the key parameter of the LDA model. This is an important task, as setting the number of topics K unreasonably can directly affect the recognition performance of the topic. If K is too small, the topic becomes over-generalized and is unable to reflect the core essence of the policy discourse; if K is too large, the topic becomes too detailed and not systematic. Therefore, it is necessary to adopt some indicators, such as perplexity and Jensen-Shannon Divergence (JSD) distance to evaluate the generalization performance of the model. In turn, this helps us determine the number of topics. In this paper, we take the method of combining perplexity and JSD distance to comprehensively determine the optimal number of topics K.

Perplexity. Used by convention in language modeling, perplexity is a recognized indicator to

evaluate the generalization ability of the model[12]. Here, a lower perplexity score indicates better generalization performance. In the LDA model, perplexity refers to the degree of uncertainty that a document d belongs to a certain topic determined by the trained model. The lower the perplexity score, the more accurate the model's inference on the topic of the document d. For a test set of M documents, the perplexity is given below.

$$\text{Perplexity}(D) = \exp\left[-\frac{\sum_{d=1}^{M}\sum_{i=1}^{N_d} P\left(w_i \mid d\right)}{\sum_{d=1}^{M} N_d}\right] \quad (1)$$

where D represents the collection of all documents, M is the number of documents, N_d refers to the number of words in the d-th document, and $P(w_i|d)$ represents the probability of the word w_i in the d-th document. According to the conditional probability formula, if the number of topics is K, then the following equation is given:

$$P\left(w_i \mid d\right) = \sum_{j=1}^{K} P\left(w_i \mid t_j\right) P\left(t_j \mid d\right) \quad (2)$$

where t_j represents the j-th topic, $P\left(w_i \mid t_j\right)$ refers to the probability of the word w_i in the j-th topic, and $P\left(t_j \mid d\right)$ represents the probability of the j-th topic in the d-th document.

Jensen-Shannon divergence distance. As the number of topics K increases, the perplexity of the model generally shows a downward trend. However, there are usually some obvious inflection points, which indicate that the generalization performance of the model is significantly improved after a certain number of topics[24]. Often, we select the number of topics when K increases and the perplexity value decreases only slightly. Therefore, the number of topics can be approximately estimated based on these inflection points, although some studies have shown that simply determining the number of topics through perplexity may not be sufficiently accurate. Therefore, other factors must be considered, such as the average similarity of topics, which refers to the average degree of difference among all the identified topics, usually measured based on JSD.

Kullback-Leibler Divergence (KLD) is a measure of the asymmetry of the difference between two probability distributions. The JSD distance is calculated according to the KLD distance with some notable differences, including that JSD is symmetric and it is always a finite value raging from 0 to 1. The JSD distance can be used to measure the similarity of two multinomial distributions P and Q. The larger the value of JSD distance is, the smaller the similarity among two multinomial distributions is.

The calculation equations of the JSD distance are as follows:

$$\text{JSD}(T_i \| T_j) = \frac{1}{2}[\text{KLD}(T_i \| M) + \text{KLD}(T_j \| M)] \quad (3)$$

$$\text{KLD}(T_i \| T_j) = \sum_{k=1}^{N} T_i\left(w_k\right) \log\left[\frac{T_i\left(w_k\right)}{T_j\left(w_k\right)}\right] \quad (4)$$

$$M = \frac{T_i\left(w_k\right) + T_j\left(w_k\right)}{2} \quad (5)$$

where $\text{JSD}\left(T_i \| T_j\right)$ represents the distance between the topic T_i and the topic T_j; the larger the value of JSD, the smaller the similarity between T_i and T_j; and N refers to the number of words in the corpus, while $T_i\left(w_k\right)$ represents the probability of the word w_k appearing in the i-th topic. Thus, we have Eq. (6) below.

$$\text{JSD}(T_i \| T_j) = \frac{1}{2}\sum_{k=1}^{N} T_i\left(w_k\right) \log\left[\frac{2T_i\left(w_k\right)}{T_i(w_k)+T_j(w_k)}\right] + \frac{1}{2}\sum_{k=1}^{N} T_j(w_k) \log\left[\frac{2T_j\left(w_k\right)}{T_i\left(w_k\right)+T_j\left(w_k\right)}\right] \quad (6)$$

We can obtain the average similarity of all topics of the model by calculating the average value of the JSD distances of all topics. The equation for calculating the average similarity of the topics is as follows:

$$\text{Avg\_Sim}(D) = \frac{2\sum_{i=1}^{K}\sum_{j=i+1}^{K} \text{JSD}\left(T_i \| T_j\right)}{K(K-1)} \quad (7)$$

where D is the document set.

Here, the smaller the average similarity distance of the topics, the greater the difference among topics, and the stronger the generalization performance of the model.

In our work, we determine the number of topics K by thoroughly considering the perplexity and JSD distance, thus allowing us to train the well-constructed LDA model.

3.4 Coordination strength

We can use the well-established LDA model to infer the topics of interest related to policy discourse. A variety of

government departments are responsible for publishing different policies. Thus, if a policy involves a topic, it can be considered that the corresponding government department of this policy pays attention to the topic as well.

Here, we define the CS based on the value of the cosine distance similarity among the topic frequency vectors of the hierarchical departments. Moreover, we can obtain the topic frequency vector of each department by calculating the frequencies of the topics concerned based on the hierarchical departments. In particular, for one department D_i, D_i consists of K topics, that is, $D_i = \{x_1, x_2, \ldots, x_K\}$. Then, we calculate the cosine distance similarity of the topic frequency vector of the central and local government departments. The calculation equation of the CS of local departments is as follows:

$$\mathrm{CS}(D_1, D_2) = \frac{\sum_{i=1}^{K}(x_i y_i)}{\sqrt{\sum_{i=1}^{K} x_i^2}\sqrt{\sum_{i=1}^{K} y_i^2}} \tag{8}$$

where D_1 and D_2 are the central and local government departments, respectively; x_i and y_i represent the frequency of the i-th topic of concern by the central and local government departments, respectively.

The greater the value of cosine similarity, the more similar the topics of concern given attention by the departments of central and local governments. This demonstrates the stronger coordination relationship between the two government entities.

Next, we set the criteria to evaluate the coordination among departments based on CS. The value of CS greater than 0.30 can be divided into three levels: weak [0.30, 0.50), medium [0.50, 0.75), and strong [0.75, 1.00) coordination. Using these values, we can judge the coordination degree between local and central government departments.

3.5 Topic diffusion

After obtaining the topic frequency vectors of all departments, for one topic, we can count the number of policy discourses that pay attention to it, which we call topic diffusion. The equation for calculating TD is as follows:

$$\mathrm{TD}(t_j) = \sum_{i=1}^{N}\left(x_{ij}\right) \tag{9}$$

where $\mathrm{TD}(t_j)$ is TD of the j-th topic, N is the number of departments, and x_{ij} is the frequency of the i-th department's attention to the j-th topic (x_{ij} can be obtained from the topic frequency vector of all departments, as shown in Section 3.4).

The greater the TD of one topic, the greater the focus of policy discourses on this topic. In turn, this indicates that this is currently a hot topic.

4 Experiment

In this section, we evaluate the performance of our LDA embedded policy knowledge graph using two experimental tasks. Specifically, we aim to answer the following questions:

- Can our model clearly demonstrate TD?
- Can our model contribute to the evaluation of the coordination among hierarchical government departments?

4.1 Data collection and corpus

4.1.1 Data sources

In this paper, we ran our experiments from the official websites of departments of the central government and Hubei Province as the data source. Then, we queried the policy discourses related to rural innovation and entrepreneurship in public information channels. The policy release period is from 2014 to 2020. After performing manual selection, we finally obtained 57 central policies and 46 policies of Hubei Province. The query date was February 10, 2021.

4.1.2 Generating the corpus

First, we preprocessed each policy discourse from the obtained policy data by deleting the invalid contents, such as the title, stop words, punctuation marks, and index numbers at the beginning of each policy. We only retained the main content of each policy.

Then, on the basis of mutual information and information entropy, we adopted the algorithm of phrase extraction to extract the key phrases of 57 central policies. The extracted phrases were then used to make a phrase dictionary. Only the phrases that appeared in more than two documents were retained, while those that appeared in 50% of the documents were deleted. Finally, a dictionary containing 1936 phrases was obtained. Some phrases and corresponding indexes in the dictionary are shown in Table 1. Finally, using this dictionary, 57 central policies were transformed into

Table 1 Part of a dictionary containing 1936 phrases based on 57 central policies.

Index	Phrase
1	Entrepreneurship and innovation
2	Network security
3	Artificial intelligence emerging industry
4	Internet finance
⋮	⋮
1933	Cross-border electronic-commerce (E-commerce)
1934	Coordinated development
1935	Industry convergence
1936	Agricultural science and technology (Sci-Tech)

a bag-of-words model to generate a standard corpus for model training.

4.2 Determining the optimal number of topics *K*

Initially, it is estimated that the optimal number of topics would range from 10 to 25. According to the method introduced in Section 3.3, the optimal number of topics was selected by calculating the perplexity and the average topic similarity of the model under different values of K. Using the corpus obtained in Section 4.1, we generated LDA models with different K. The change of perplexity and the average similarity of topics with K ranging from 10 to 25 are shown in Table 2. As can be seen, when $K = 19$, the average JSD distance of topics reaches to the local maximum, while perplexity reaches the global minimum when $K = 21$.

Then we draw the JSD distance bubble chart (each bubble in the bubble chart represents one topic) of the LDA model when K equals to 19, 20 and 21, as shown in Fig. 2. The distance among the bubbles in the chart represents the JSD distance among the topics. It can be found that there is less overlap among the topics when

Table 2 Value of perplexity and topic average similarity with *K* ranging from 10 to 25.

K	Perplexity	Topic average similarity
10	320.34	0.095
11	320.93	0.095
12	298.69	0.111
13	302.93	0.122
14	289.58	0.124
15	275.21	0.131
16	281.11	0.128
17	287.03	0.134
18	260.55	0.142
19	269.86	0.153
20	262.47	0.146
21	240.44	0.135
22	264.59	0.162
23	246.28	0.163
24	256.38	0.161
25	259.21	0.167

$K = 20$, so the optimal number of topics K of the model is determined to be 20. Finally, we generate the LDA model with K equal to 20 by using the corpus obtained in Section 4.1, and the training time does not exceed 2000 ms in total.

4.3 Obtaining the global topics

We can obtain the topic-word distribution after generating the LDA model. Next, we selected the top five phrases with the highest probability under each topic, as shown in Table 3.

According to the related phrases under different topics in Table 3, the name of each topic can be summarized with the help of a policy interpretation expert, as shown in Table 4. All 20 topics in Table 4 are topics of government concern in the field of rural innovation and

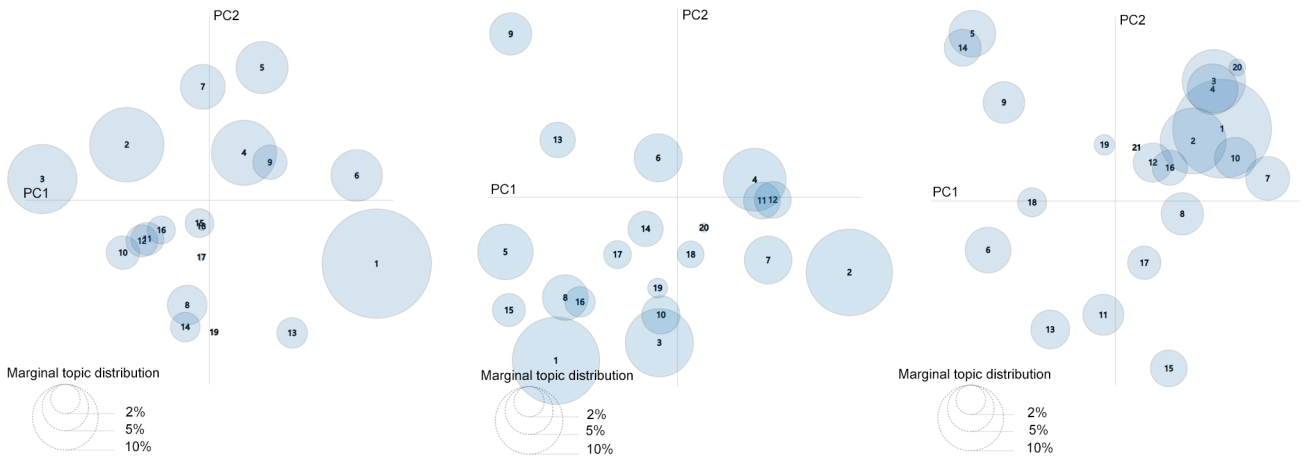

Fig. 2 Bubble chart of the LDA model with 19, 20, and 21 topics.

Table 3 Top 5 phrases with the highest probability under each topic.

Topic No.	Relevant phrase
1	Entrepreneurship and innovation; social capital; farmers cooperative economic and social development; public service capability
2	Rural revitalization; entrepreneurship and innovation; development of cultural industry cultural enterprises; public service platform
3	Resumption of work and production; public service platform; cultural technology; E-commerce development; village revitalization
4	Employment and entrepreneurship; digital agriculture; village revitalization; mass entrepreneurship; entrepreneurship service platform
5	Innovation and entrepreneurship; intellectual property law; public service platform; rural revitalization; rural innovation
6	Technological innovation; village revitalization; innovation driven; agricultural production; national food security
7	Innovation and entrepreneurship; village revitalization; business model innovation basic public services; E-commerce enterprise
8	Resumption of work and production; cultural science and technology; technological innovation; information service platform; high-tech enterprise
9	Village revitalization; farmer cooperative; deep poverty area; enterprise training; environmental improvement
10	Innovation and entrepreneurship; village revitalization; entrepreneurship service university research institute; cultural industry development
11	Village revitalization; innovation and entrepreneurship; poverty alleviation deep poverty area; integrated development
12	Innovation and entrepreneurship; agricultural science and technology; public service platform; development of modern agriculture; agricultural production
13	Village revitalization; social capital; agricultural science and technology; poverty alleviation; credit information
14	Innovation and entrepreneurship; entrepreneurship leader; rural innovation village revitalization; entrepreneurship service
15	Cultural industry development; village revitalization; cultural enterprise innovative development; online interaction
16	Vocational skills training; innovation and entrepreneurship; E-commerce enterprise; E-commerce development; public service platform
17	Village revitalization; cultural technology; poverty alleviation; deep poverty areas; building well-off society
18	Resumption of work and production; information service platform; university graduates; village revitalization; entrepreneurship training
19	Social capital; cultural industry development; public service platform; cultural enterprise; innovation and entrepreneurship
20	Training organizations; human resources market; special fund management; entrepreneurship training; workers vocational skills

Table 4 Names of topics concluded from relevant phrases.

Index	Name of topics	Index	Name of topics
1	Rural informatization strategy	11	Social security system of the people going to rural and back to rural
2	Developing the cultural of MEIIs in rural areas	12	Promotion of agricultural science and technology
3	E-commerce in rural areas	13	Business incubator
4	Agricultural industrial chain	14	Service and support system for entrepreneurship
5	Local characteristic agriculture	15	Rural cultural industry development
6	Land-use planning	16	Developing the rural MEIIs park
7	The online and offline (O2O) agriculture development	17	Science and technology engine in Chinese rural development
8	Agricultural Sci-Tech innovation	18	Heading-back-to-rural college student start-up
9	Cultivating innovation and start-up subjects	19	Innovation and start-up service system in rural areas
10	Improving the innovation and start-up eco-environment	20	Training of new professional peasants

entrepreneurship.

4.4 Analysis of TD

TD refers to the number of policies that pay attention to the same topic. Here, we use the LDA model to infer the topics of all policies (running time of no more than 1000 ms) and counted the number of policies that paid attention to the same topic. In doing so, we were able to obtain the topic frequency vector of each department and draw the policy knowledge graphs of Hubei Province and central government departments, as shown in Figs. 3 and 4, respectively. The blue node in the graph represents the topic, while the number on the line represents the frequency of a department's attention to a certain topic. The meanings of the department abbreviations in Figs. 3 and 4 are as follows:

SC: State Council;

MIIT: Ministry of Industry and Information Technology;

MST: Ministry of Science and Technology;

MA: Ministry of Agriculture;

MHRSS: Ministry of Human Resources and Social

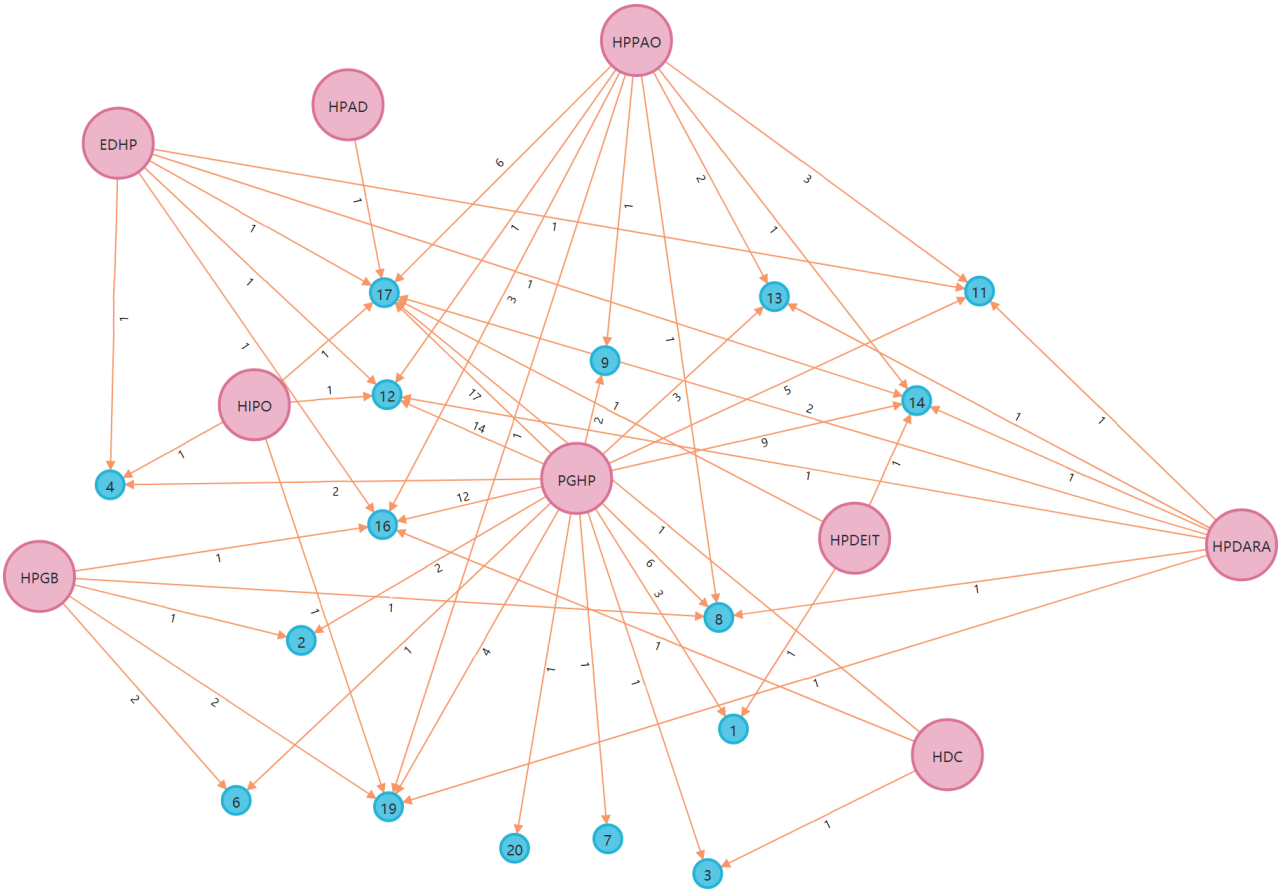

Fig. 3 Policy knowledge graph of Hubei Province.

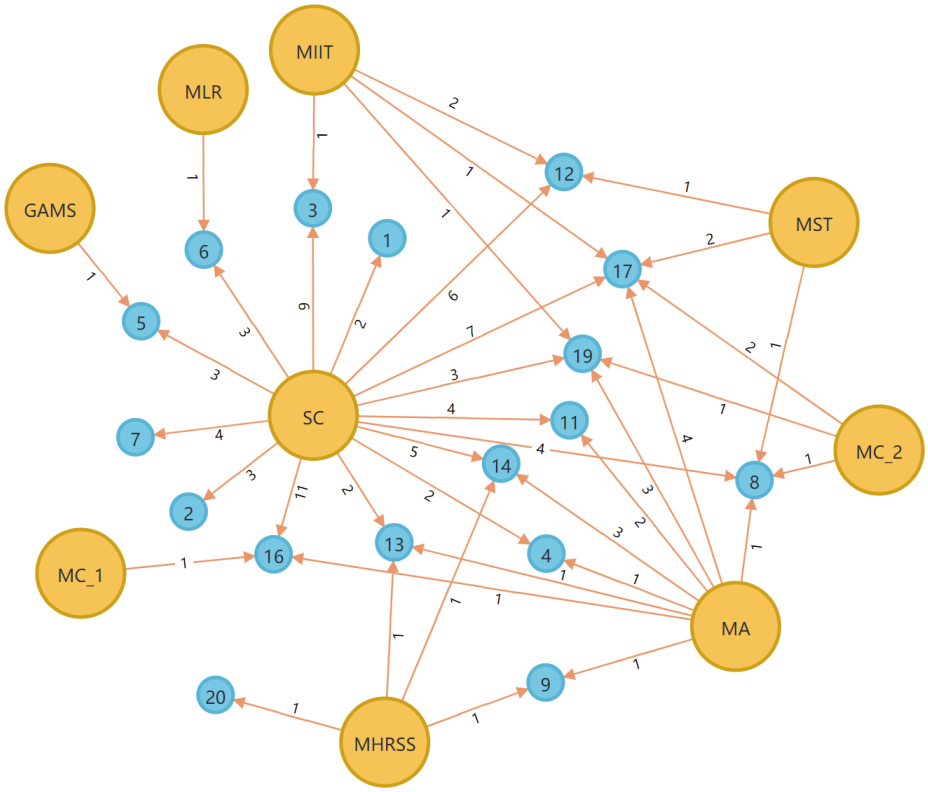

Fig. 4 Policy knowledge graph of the central government.

Security;

MC_1: Ministry of Commerce;

MC_2: Ministry of Culture;

MLR: Ministry of Land and Resources;

GAMS: General Administration of Market Supervision;

PGHP: People's Government of Hubei Province;

HPPAO: Hubei Provincial Poverty Alleviation Office;

HPAD: Hubei Provincial Audit Department;

HIPO: Hubei Intellectual Property Office;

HPDARA: Hubei Provincial Department of Agriculture and Rural Affairs;

HDC: Hubei Department of Commerce;

HPGB: Hubei Provincial Grain Bureau;

EDHP: Education Department of Hubei Province;

HPDEIT: Hubei Provincial Department of Economy and Information Technology.

We can calculate the frequency of attention of all topics according to the policy knowledge graphs of Hubei Province and the central government. Some of the topics with the highest frequency of attention are shown in Fig. 5. As can be seen, the departments of the Hubei Province paid more attention to the following topics: 12 (promotion for agricultural science and technology), 14 (service and support system for the entrepreneurship), 16 (developing the rural mass entrepreneurship and innovation initiatives (MEIIs) park), and 17 (science and technology engine in China rural development), which are the same as the topics the central government departments paid attention to. Thus, this finding indicates that Topics 12, 14, 16, and 17 are the hot topics of rural vitalization in China's rural areas from 2014 to 2020.

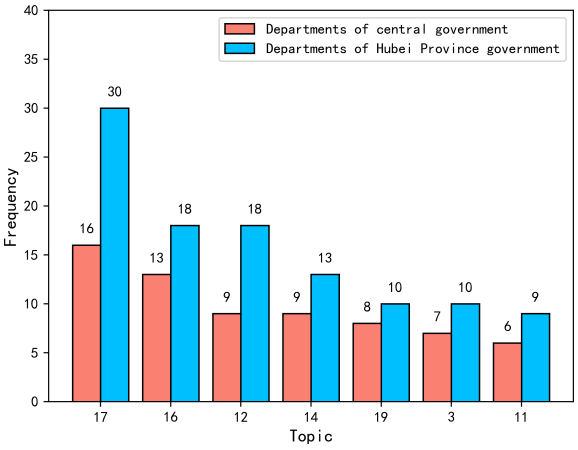

Fig. 5 The topic diffusion in the departments of Hubei Province and the central government.

4.5 Analysis of CS

Next, we calculated the CS between the departments of Hubei Province and the central government according to the topic frequency vector of each department obtained in Section 4.4, and by using the formula proposed in Section 3.4. We used the State Council and the Ministry of Agriculture as examples to calculate the strength of the coordination relationship among the departments of Hubei Province and the State Council and Ministry of Agriculture to obtain the coordination graph among departments, as shown in Fig. 6.

According to the three levels of CS in Section 3.4, it can be seen that the PGHP has a strong coordinating relationship with the State Council and the Ministry of Agriculture, while the Hubei Provincial Food Bureau has a weaker coordination relationship with the same government agencies; the degree of coordination among other departments is medium. In general, the policy coordination among the departments of the PGHP, the State Council, and the Ministry of Agriculture is basically consistent.

4.6 Evaluation

Aside from using the policy knowledge graph to evaluate the CS among departments, the accuracy of the model is verified through the expert scoring method combined with the policy quantitative standard[6] in econometrics. Three policy interpretation experts from the School of Public Management were invited to evaluate the CS

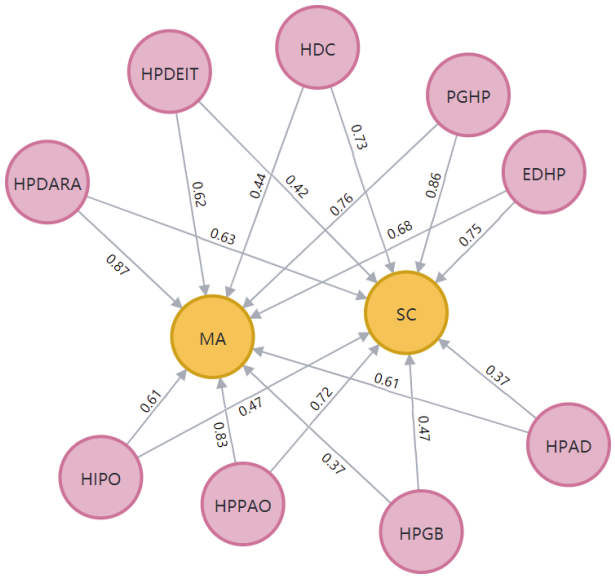

Fig. 6 Coordination graph among departments of Hubei Province, State Council, and the Ministry of Agriculture.

among different departments of Hubei Province and the central departments. Their opinions are consistent with our results. For example, through the experts' interpretation of the policies issued by the State Council and the PGHP, we are able to confirm a strong cooperative relationship in terms of rural innovation and entrepreneurship between the State Council and the PGHP. In other words, the information provided by the policy knowledge graph has a great degree of similarity with the manual grading conducted by experts. Therefore, in the process of evaluating the degree of department coordination, using the policy knowledge graph has a certain reference value with high credibility.

5 Conclusion

This paper examined the gaps among the well-established econometrics theories of policy coordination under the domains of social science and economics. In particular, this work examined the lack of automatic computational methods that can be adapted to evaluate the CS among different departments. To this end, we developed a new approach called PG-CODE, a novel embedded policy knowledge graph for evaluating the CS among hierarchical departments. Then, we conducted experiments on the rural innovation and entrepreneurship policy discourses issued by the departments of the central government and Hubei Province government during the period from 2014 to 2020, Based on our findings, the following conclusions can be drawn:

(1) We were able to combine two indicators of perplexity and the JSD distance among topics to determine the optimal number of topics for the LDA model. Meanwhile, according to the TD, we identified the popular topics that the central government and Hubei Province government departments paid attention to. We were also able to clarify the focus of the central government's rural revitalization work.

(2) The policy knowledge graph can be helpful in assessing whether the local departments and the central departments are able to foster and maintain coordination. By calculating the CS among departments, the policy knowledge graph can be used to determine the degree of coordination among local and central departments, thus providing semi-automatic assistance in the evaluation of departmental coordination.

(3) Finally, the policy knowledge graph can be used to build an intelligent question answering system in future, which can provide suggestions for expert evaluations of various government policies.

Acknowledgment

This work was supported by the National Social Science Fund of China (No. 20BGL231), and the Natural Science Foundation of Hubei Province (No. 2018CFB380).

References

[1] Q. X. Han, The thought of people's co-creation and sharing: Systematic analysis of the new thought of the CPC central committee on state governance, (in Chinese), *Journal of the Party School of the Central Committee of the C. P. C.*, vol. 20, no. 1, pp. 15–27, 2016.

[2] A. G. Li and Q. Zeng, Policy situation and the next prospect of widespread entrepreneurship and innovation, (in Chinese), *Reform*, no. 10, pp. 149–157, 2017.

[3] H. Y. Zeng, Implementation progress and suggestions of entrepreneurship and innovation, (in Chinese), *Macroeconomic Management*, no. 12, pp. 21–23, 2015.

[4] Z. S. Zhang and X. T. Li, Performance appraisal and research on policy of innovation and entrepreneurship based on the DEA model—Take Tianjin business incubators as the analysis objects, (in Chinese), *Journal of Tianjin University (Social Sciences)*, vol. 18, no. 5, pp. 385–391, 2016.

[5] C. E. Hughes, A. Ritter, and N. Mabbitt, Drug policy coordination: Identifying and assessing dimensions of coordination, *Int. J. Drug Policy*, vol. 24, no. 3, pp. 244–250, 2013.

[6] J. S. Peng, W. G. Zhong, and W. X. Sun, Policy measurement, policy coordinated evolution and economic performance: An empirical study based on innovation policy, (in Chinese), *Management World*, no. 9, pp. 25–36, 2008.

[7] S. Stathopoulou, D. Psaltopoulos, and D. Skuras, Rural entrepreneurship in Europe, *Int. J. Entrep. Behav. Res.*, vol. 10, no. 6, pp. 404–425, 2004.

[8] S. J. Goetz, M. Partridge, S. C. Deller, and D. A. Fleming, Evaluating U.S. rural entrepreneurship policy, *J. Reg. Anal. Policy*, vol. 40, no. 1, pp. 20–33, 2010.

[9] J. Murdoch, Networks—A new paradigm of rural development? *J. Rural Stud.*, vol. 16, no. 4, pp. 407–419, 2000.

[10] G. D. Libecap, Economic variables and the development of the law: The case of western mineral rights, *The Journal of Economic History*, vol. 38, no. 2, pp. 338–362, 1978.

[11] X. Y. Liu, Y. R. Pang, W. S. Hou, and X. H. Shan, Research on the coordination of S&T innovation policies between central and local governments from the perspective of relation-content, (in Chinese), *Forum on Science and Technology in China*, no. 12, pp. 13–21, 2020.

[12] D. M. Blei, A. Y. Ng, and M. I. Jordan, Latent dirichlet allocation, *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[13] H. Jelodar, Y. L. Wang, C. Yuan, X. Feng, X. H. Jiang, Y. C. Li, and L. Zhao, Latent dirichlet allocation (LDA) and topic modeling: Models, applications, a survey, *Multimed. Tools Appl.*, vol. 78, no. 11, pp. 15169–15211, 2019.

[14] B. Chen, L. L. Zhu, D. Kifer, and D. Lee, What is an opinion about? Exploring political standpoints using opinion scoring model, in *Proc. 24th AAAI Conf. Artificial Intelligence*, Atlanta, GA, USA, 2010, pp. 1007–1012.

[15] R. Cohen and D. Ruths, Classifying political orientation on twitter: It's not easy! in *Proc. 7th Int. AAAI Conf. Weblogs and Social Media*, Cambridge, MA, USA, 2013.

[16] D. Greene and J. P. Cross, Unveiling the political agenda of the European parliament plenary: A topical analysis, in *Proc. ACM Web Science Conf.*, Oxford, UK, 2015, p. 2.

[17] Y. Shirota, Y. Yano, T. Hashimoto, and T. Sakura, Monetary policy topic extraction by using LDA: Japanese monetary policy of the second ABE cabinet term, presented at 2015 IIAI 4th Int. Congress on Advanced Applied Informatics, Okayama, Japan, 2015, pp. 8–13.

[18] S. S. Jia and B. G. Wu, Incorporating LDA based text mining method to explore new energy vehicles in China, *IEEE Access*, vol. 6, pp. 64596–64602, 2018.

[19] J. Zhao, H. F. Li, and C. G. Li, Analysis of research topic evolution of coordinated development of Beijing-Tianjin-Hebei based on probabilistic topic models, (in Chinese), *Sci. Technol. Eng.*, vol. 19, no. 36, pp. 225–234, 2019.

[20] P. T. Xie, D. Y. Yang, and E. Xing, Incorporating word correlation knowledge into topic modeling, in *Proc. 2015 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, CO, USA, 2015, pp. 725–734.

[21] L. Yao, Y. Zhang, B. G. Wei, H. Z. Qian, and Y. B. Wang, Incorporating probabilistic knowledge into topic models, in *Proc. 19th Pacific-Asia Conf. Knowledge Discovery and Data Mining*, Ho Chi Minh City, Vietnam, 2015, pp. 586–597.

[22] R. B. Xie, Z. Y. Liu, J. Jia, H. B. Luan, and M. S. Sun, Representation learning of knowledge graphs with entity descriptions, in *Proc. 30th AAAI Conf. Artificial Intelligence*, Phoenix, AZ, USA, 2016.

[23] L. Yao, Y. Zhang, B. G. Wei, Z. Jin, R. Zhang, Y. Y. Zhang, and Q. F. Chen, Incorporating knowledge graph embeddings into topic modeling, in *Proc. 31st AAAI Conf. Artificial Intelligence*, San Francisco, CA, USA, 2017, pp. 3119–3126.

[24] J. B. Qu and S. Y. Ou, Analyzing topic evolution with topic filtering and relevance, (in Chinese), *Data Analysis and Knowledge Discovery*, vol. 2, no. 1, pp. 64–75, 2018.

Yilin Kang received the PhD degree in computer science from Nanyang Technological University, Singapore, in 2015. She is currently an assistant professor with the School of Computer Science, South-Central University for Nationalities (SCUN), Wuhan, China. Her current research interests include knowledge discovery, cognitive and neural systems, and brain-inspired computing. Prior to joining SCUN, she was a research fellow with the Nanyang Technological University-University of British Columbia Joint Research Centre of Excellence in Active Living for the Elderly (LILY). She serves as the PC member of AAAI 2019–2022, IJCAI 2021, 2016, and OC member of IEEE WI/IAT 2015. She serves as the reviewer in several major journals, such as *IEEE Trans. on NNLS*, *IEEE Trans. on SMC*, and *JAAMAS*.

Renwei Ou is currently pursuing the BS degree in software engineering at the School of Computer Science, South-Central University for Nationalities, Wuhan, China. His reaserch interests include machine learning, probabilistic modelling, and reasoning.

Yi Zhang is currently pursuing the BS degree in intelligent science and technology at the School of Computer Science, South-Central University for Nationalities, Wuhan, China. Her research interests include machine learning and natural language processing.

Hongling Li received the PhD degree from Huazhong University of Science and Technology, Wuhan, China, in 2009. She is currently an associate professor with the School of Public Management, South-Central University for Nationalities, Wuhan, China. Her current research interests include analysis on public policy, science and technology policy and innovation management, and public human resource management. She serves as a member of International Association for Chinese Management Research (IACMR).

Shasha Tian received the PhD degree from Wuhan University, Wuhan, China, in 2021. She is currently an associate professor with the School of Computer Science, South-Central University for Nationalities, Wuhan, China. Her research interests include robot path planning, intelligent algorithm, and wireless sensor networks. She has published more than 10 research papers in refereed conferences and journals.