Computer Life (CPL) ISSN: 1819-4818
DELIVERING QUALITY SCIENCE TO THE WORLD

Content Analysis of Science and Technology Projects Based on Topic Model

Gaopei Li

Faculty of Economics and Management, Xidian University, Xi'an, China

wangyigpll @163.com

Abstract: In this paper, we try to apply LDA topic model to the analysis of the content of science and technology projects. We try to improve the rough and subjective problems in the past analysis. This paper reveals the theme of the project and the distribution of the corresponding content words. The LDA is used to analyze theme of Shaanxi soft science project, revealing the research content under the project. The paper verify that LDA can be effectively used to analyze scientific and technological project data, reveal the research theme of Shaanxi soft science project.
Keywords: Science and technology projects; topic model; LDA.

1. INTRODUCTION

Technological innovation is receiving more and more attention. Science and technology projects play an indispensable role in the advancement of science and technology innovation. The analysis of the project content can obtain the theme of scientific research, at the same time, it is the basic information for judging the hot topics in the research field and identifying the frontiers of scientific research.

Since the 1990s, information mining based on scientific and technological project data has received extensive attention from the academic community. Most of the researches use data statistics and measurement methods to analyze the status and characteristics of science and technology projects through structured statistics.

In terms of project topic content analysis, the number of existing researches is small but is increasing year by year. The research directions include the following categories:

The first method is using basic statistical methods analyzes the distribution of projects under different themes. For example, Wang Jiqiang et al. statistic the national social science fund project data through the scope and frequency of the subject content, and then analyze research status in the field of library and information literature in China[5]. Lu Yang analyzes the research content of overseas Chinese projects by counting the number of projects in different years under different themes [6].

The second method is using the co-occurrence analysis of topic words and other methods reveals research hotspots and evolution trends in a certain field. Most of the existing researches focus on this part, such as Li Huafeng's co-occurrence network on the project name and the analysis of the country. The scope of research on key social science projects [7]. Zhao Rongying obtained the research topic of library, intelligence and archives management subject by constructing a social network fund project keyword co-occurrence network [8]. Tang Ting is constructing a co-occurrence matrix of fund

project data subject heads to analyze the research situation in the field of knowledge management [9]. Yucaihua analyzes the academic research of China's manufacturing industry through the co-occurrence network of the fund project's key words [10]. Du Chaonan et al. built a common word network for the NSFC project and analyzed the research hotspots of privacy issues [11]. Zhang Li et al. constructed a co-occurrence network of social science fund project keywords, and analyzed the research hotspots of journalism and communication [12].

We can find that the second type of research is a common form of content analysis of project topics—building a co-occurrence network of subject terms and analyzing existing research through comprehensive analysis. The co-word network is used to reveal the relationship between the keywords, revealing the research hotspots in the field. The limitations of the current research are reflected in the following points: First, the source of the subject words, the artificially given subject words are greatly influenced by the subjective consciousness, and the results given by different scholars with different text data of the same project are not the same. Secondly, the analysis of the subject content of the project is mainly focusing on the word frequency statistics of the topic words, the network relationship analysis between the topic words, the mining of the project text data is not enough and deep, and the subject matter of the project is less in-depth research.

In order to mine the subject content of science and technology projects objectively and deeply, this paper uses a method based on the LDA topic model to analyze the content of the science and technology project data, combing with the characteristics of the unstructured data length of the text of the science and technology project. In the data preprocessing stage, this paper constructs a word dictionary and stop word dictionary suitable for science and technology projects, optimizes the effect of project text data segmentation, and solves the problem that the subject content analysis of the previous project is too subjective and lacks scientific. In this way, the project text content can be analyzed in depth and the project text data can be fully analyzed. Through analyzing the text of the science and technology project, it not only enriches the project content analysis method theoretically, but also provides richer and more valuable information for the science and technology project management work.

2. LDA MODEL CONCEPT

The LDA model was proposed by Blei et al. [13], and its modeling process is represented by a directed probability graph:

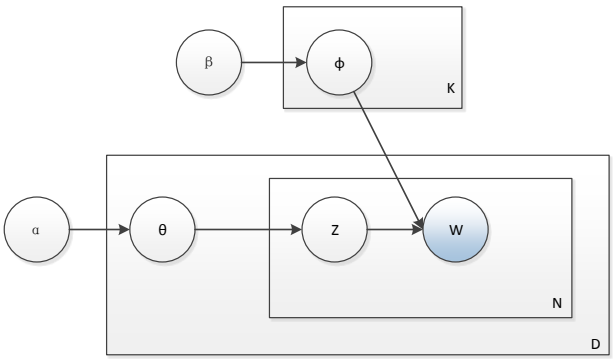

Fig. 1 The directed probability model of the LDA model

The LDA modeling process (simulating a document generation process) is shown below:

1. Number N of words selected for document d, N~Poisson(ε)
2. Select a topic distribution θ for document d, θ~Dirichlet(α)
3. generate a single word w of document d. This process includes:

(1) First select an attribute z for the word w from the topic distribution of this document, z~Multinomial(θ);

(2) After selecting the subject z which the word w belongs, the word w is generated with a probability of $p(w_n|z_n,\beta)$.

In addition, because the Gibbs Samplimg [14] method requires less memory and better solution, the algorithm is often used for parameter reasoning. For the number of topics, the optimal number of topics is determined mainly by calculating the perplexity and the method of calculating the similarity value between the topics. The calculation formula for the perplexity is shown below:

$$Perplexity\ (\mathrm{D}) = exp\left\{-\frac{\sum_{d=1}^{D} log\, p\,(w)}{\sum_{d=1}^{D} N_d}\right\}$$

3. EXPERIMENT AND RESULT ANALYSIS

3.1 Experimental data

In this study, the data of the soft science project in the Shaanxi Science and Technology Project Database was selected as the research data. Considering the analyzability of project data items, this paper finally selects 1,325 soft science project data from 2005-2015 (except 2014) as research data. The number of project data for different years is shown below:

Fig. 2 Project Year - Number Display

3.2 Result Analysis of LDA Model

In this study, the two data items, project name and project profile, are combined, and the combined text is used as a corpus for subsequent topic analysis. In the pre-processing stage of the data, the Chinese vocabulary software is applied in the word segmentation, de-stopping, and noise reduction work, combining self-constructed stop word list.

Next, the optimal number of topics is determined. This study is determined by the perplexity method.

After the experimental corpus is run by the corresponding Python program, the optimal number of topics is determined, K=10. The results are shown below:

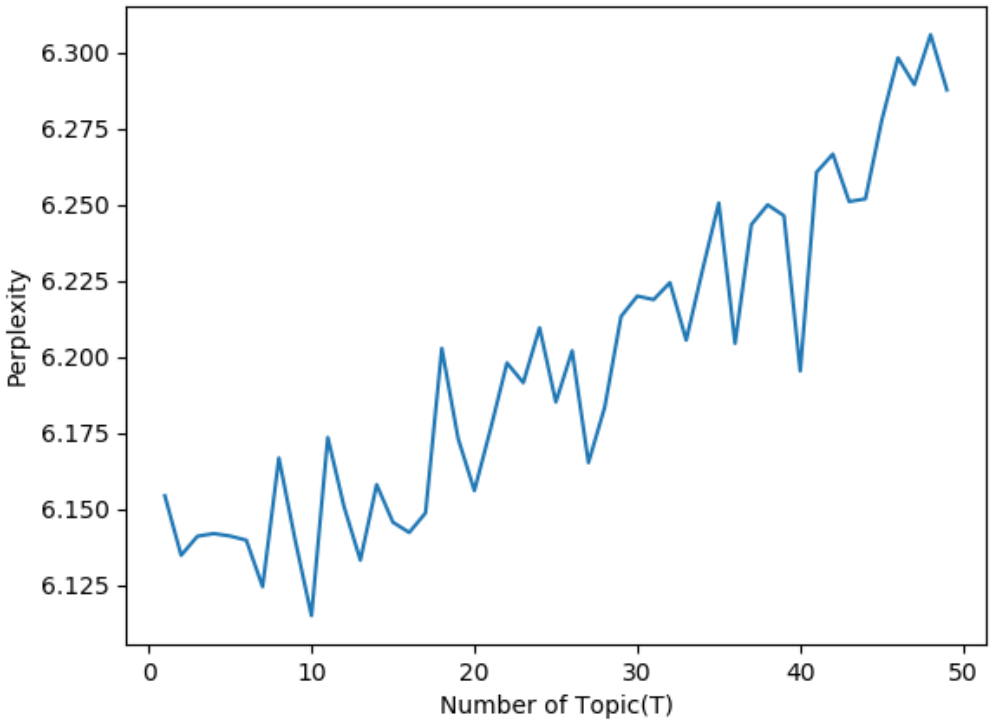

Fig. 3 Number of Topic – Perplexity

Next, the corpus is processed using the LDA model. The whole process is implemented in python. We set the parameter $\alpha = 50/K=5.0$ and the parameter $\beta = 0.01$. Finally, we get the top 10 words and their probability distributions under 10 topics:

Table 1 10 topics - feature word distribution

Topic 0		Topic 1		Topic 2		Topic 3		Topic 4	
power	0.075	finance	0.078	urban and rural planning	0.148	urban group	0.131	region	0.119
technology	0.073	financing	0.071	farmer	0.121	Central Shaanxi area	0.049	collaboration	0.059
innovation	0.066	technology	0.057	financing	0.077	region	0.048	industrial clusters	0.044
push	0.058	countryside	0.045	industrialization	0.061	policy mechanism	0.031	supply chain	0.041
Shaanxi	0.041	innovation	0.036	hospital	0.025	economic society	0.028	competitiveness	0.039
economy	0.036	urbanization	0.035	Shaanxi	0.023	innovation	0.026	equipment manufacturing	0.021
farmland	0.024	mechanism	0.031	poverty	0.016	SME	0.025	innovation	0.017
industry	0.021	financial	0.028	science resources	0.014	constraints	0.025	industrial park	0.017
system	0.017	Shaanxi	0.024	city	0.009	urbanization	0.022	status	0.016
platform	0.008	consumption	0.019	coordinating resources	0.007	modernization	0.021	evaluation	0.008
Topic 5		Topic 6		Topic 7		Topic 8		Topic 9	
economic development	0.073	ecology	0.110	informatization	0.058	knowledge	0.086	Shaanxi	0.073

prevention and control	0.052	sustainable development	0.096	industrialization	0.054	capital	0.067	society	0.068
Shaanxi	0.035	industry	0.052	agriculture	0.053	human capital	0.057	pension	0.060
economic growth	0.026	energy	0.039	risk	Intellectual Protection	Intellectual Protection	0.039	innovation	0.057
Western Region	0.021	tourism	0.035	enterprise	0.037	economic development	0.035	dynamic	0.049
medical	0.017	technology	0.033	countryside	0.036	technology	0.034	urban and rural	0.047
countryside	0.015	reform	0.027	government	0.025	evaluation	0.028	public hospital	0.017
service industry	0.014	Circular Economy	0.026	supply chain	0.024	knowledge	0.028	security mechanism	0.015
innovation	0.013	economic development	0.024	Internet of Things	0.023	higher education	0.027	residential	0.008
legislation	0.012	financial	0.021	medical	0.022	capital	0.026	education	0.003

According to the topic-word probability distribution results, combined with the meaning of individual words under each topic and the inter-words, this paper summarizes the 10 topics:

Topic 0 is about the research and development of science and technology in Shaanxi Province, mainly including the system construction of scientific and technological innovation, technological innovation to promote economic development, and government science and technology management.

Topic 1 is about the financial innovation of Shaanxi Province, which mainly includes the rural financial development model, the impact of finance on the industrial structure, the financing of different subjects such as farmers and enterprises, and the relevance of financial and economic development.

Topic 2 is the study of urban and rural development in Shaanxi Province. It mainly involves the development path and countermeasures of urban and rural development, the development of rural industrialization, and the research on scientific and technological resources.

Topic 3 is the study of urban agglomeration development, focusing on regional innovation issues, constraints in the development of urban agglomerations, and the role of SMEs in the development of urban agglomerations.

Topic 4 is the study of industrial clusters and innovation development, mainly involving industrial cluster research in different industries such as equipment manufacturing and tourism industries, and research on the status quo and development evaluation system of industrial clusters.

Topic 5 is the study of economic development in Shaanxi Province, focusing on economic growth, economic development in the western region, economic development in rural areas, the impact of different industries such as service industry on economic development, and the study of corresponding policies and regulations.

Topic 6 is the study of sustainable development, including the focus on ecological governance, the sustainable development of the energy industry, the development of the ecotourism industry, and the research on the mechanism of circular economy.

Topic 7 is the research on informatization construction, including the development of Shaanxi Sanhua (informatization, industrialization, urbanization), information construction and economic development.

Topic 8 is the study of knowledge talent management, which mainly involves the research and development of talent training and evaluation system, the protection of corporate intellectual property

rights, and the influence of invisible capital such as knowledge and talents on economic development. Topic 9 is the study of social management in Shaanxi Province. It focuses on how to manage and solve social and people's livelihood issues, such as pension issues, medical security mechanisms, housing prices, and coordinated urban and rural development.

The feature words included in the above 10 topics obtained by LDA model are words that appear in the content of the project text with higher frequency and can represent the content of the topic. Each topic is a collection of these words, and each topic can be regarded as a hot spot of the project content. For the popularity of hot content, this paper quantifies the number of researches on the hot content. The more the number of related research projects, the more popular the hot content.

According to another experimental result document-topic probability distribution generated by the model, the probability distribution of each document d under different topics can be known. A larger probability value indicates that the document d is more likely to belong to the topic. Determine the topic that the document most likely belong to, based on the probability value. Then count the number of documents in all project documents under different topics, and then reveal the popularity of these 10 topics in the research of Shaanxi soft science project.

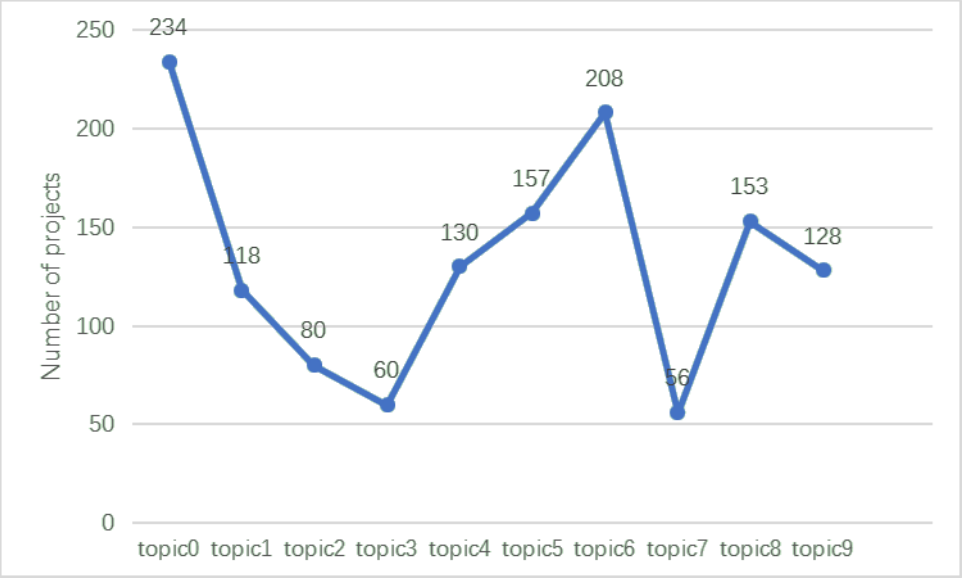

Fig. 4 Heat distribution of project topic

According to the above figure, we can get the following content: In the research content of the Shaanxi Soft Science Project, the researchers pay the highest attention to the two themes of technological development innovation and sustainable development, and the corresponding number of project research is the most. The number of corresponding projects on the theme of economic development and knowledge and talent management ranks second; the number of research on industrial development, social management, and financial innovation topics is not much different, and the popularity of the topic is on average; for the topic of informationization, urban agglomeration, urban-rural integration, the number of research projects under these three topics is relatively small, and the topic's popularity is below average.

4. CONCLUSION

In this paper, the LDA topic model is applied in the project content analysis, which realizes the

subdivision of the project content research topic and the topic popularity analysis, which makes up for the rough analysis of the previous project topic and lack of objectivity.

Applying the LDA topic model to the project topic content analysis can effectively mine the topic content and improve the depth and objectivity of the project topic analysis. Through empirical research, the 10 main topics and corresponding characteristic words of Shaanxi Soft Science Project Research are obtained, which provides more comprehensive and valuable information for scientific understanding of the research situation of the project.

However, the current research is still aimed at the analysis of the content of the project text itself. It is still necessary to expand the analysis dimension. If we can use the research of this project as the basis and analyze other data items of projects in future research, I believe that the project content can be further and comprehensively mined. At the same time, it can provide valuable information for project management.

REFERENCES

[1] Wang Feng. Analysis of Anhui Province' National Natural Science Foundation of China from 2006 to 2013[J]. China Science Foundation, 2015(1):69-72

[2] Li Ning, Gu Lingwei, Yang Yaowu. Analysis and Enlightenment of American Science and Technology Policy Fund Project from 2007 to 2016[J]. Science and Technology Management Research, 2017(18)

[3] Li Huafeng, Yuan Qinwei. Measurement of the National Social Science Fund Major Project Project from 2004 to 2015[J]. Modern Information, 2016(11): 132-139

[4] Ding Zhixiu. Analysis of National Social Science Funds and Literary Studies in the Past 20 Years[J]. Modern Information, 2015, 35(2): 119-123

[5]Wang Jiqiang, Chen Xianlai, Wang Hui. Discussion on the Present Situation and Trends of the China's Research on Library Science, Information Science and Literature Science from the Subjects of National Social Science Fund Objects in Recent 10 Years[J]. Library and Information Service, 2010, 20(19): 95-97

[6] Lu Yang. Research and Development of Overseas Chinese from the National Social Science Fund Project——Based on the Quantitative Analysis of the Overseas Chinese Research Project of the National Social Science Fund from 1991 to 2013[J]. Southeast Asia South Asia Studies, 2014(2): 94-100

[7] Li Huafeng, Yuan Qinwei. Analysis of Management Science Funding Based on NSFC Project Data from 2006 to 2015[J]. Science and Technology Management Research, 2017(6)

[8] Zhao Rongying, Zhao Wei, Chen Bikun. Analysis of the Research Status of the Subject of Library, Information and Archives Management in China*--From the Perspective of National Fund Projects from 2001 to 2012[J]. Journal of Information, 2013(7) :106-112

[9] Tang Ting, He Xiaolan. Analysis of the research topic in the field of knowledge management in national fund projects——based on strategic coordinates [J]. Information Science, 2018

[10]Yu Caihua, Lian Tonghui. Research Progress of China's Manufacturing Industry——Based on Statistical Analysis of National Social Science Fund Projects[J]. Journal of Xuzhou Institute of Technology(Social Science Edition), 2018, v.33; 04): 46-51

[11] Du Chaonan, Yuan Qinwei, Yue Quan. Research Status and Hot Topics Analysis of Privacy Issues in China——Based on the Research of National Natural Science Foundation Project Data from 2004 to 2016[J]. Information Science, 2018, V36(3):99 -104

[12] Zhang Li, Yang Wentao, You Yu, et al. Research on Atlas and Hotspots of Journalism and Communication——Based on the Empirical Analysis of the National Social Science Fund Projects from 2000 to 2017[J]. Journal of Xi'an Jiaotong University: Social Scientific edition, 2018(3)

[13] Blei D M , Ng A Y , Jordan M I , et al. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research, 2003, 3:993-1022

[14]Jing T, Zhi G, Yu H, et al. Automatic image annotation method based on multi-modal topic model [J]. Foreign Electronic Measurement Technology, 2015