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

# Analysis of the Evolutionary Trend of Science Fund Topics Based on LDA and HMM: Taking science and technology management and policy discipline as an example

Fangxin Zhang

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

1300995097@qq.com

---

*Abstract: The topic model is a modeling method that can extract large-scale document implicit topics on a large scale and effectively. This study proposes to use the implicit Dirichlet distribution (LDA) model to mine potential topic clusters in science funds. By means of hot topic extraction, trend analysis and clustering analysis, the forecast and scientific decision making for field hot work can be realized, which helps promote the government business field information intellectualization and knowledge-driving.*
*Keywords: Technology management and policy, LDA model, Topic Evolution and prediction.*

---

## 1. INTRODUCTION

In the research of scientific and technological topics, the National Natural Science Foundation of China is an important source of information and an important factor in measuring the level and achievements of scientific research. However, at present, most scholars only use the relevant data of the science fund-funded projects to conduct some statistical analysis and co-occurrence analysis, etc. When researching hot topics and development trends, statistical and metrology can not find hidden information. this method leads to the results of the research is often not deep enough and has limitations. There is no prediction of the direction and trend of future science funded projects, especially quantitative predictions. In order to determine the direction and trend of research in the National Natural Science Foundation of China, the director of the NSCF, Academician Yang Wei[1] once said: "The Fund Committee often looks for scientists with strategic thinking to give a forward-looking perspective in the form of brainstorming. This guide is not a binding, but a guiding role." Expert experience and knowledge is absolutely important, but such methods cannot avoid the limit of experts in knowledge and the inability to analyze large amounts of data. So this method cannot master the research direction of the world today accurately and comprehensively . With the rapid development of science and technology, the amount of data of the Natural Science Foundation of the project has increased year by year. How to accurately obtain the evolution path of the hot topic and predict the topic from the funded project is of great significance. Research on the evolutionary path and prediction of the theme funded by the National Natural Science Foundation of China is conducive to grasping the development trends and evolutionary trends in the field of science and technology, accurately identifying potential research sites, and potential problems in the process of scientific research management. This is beneficial for researchers to grasp the research hot topics and research frontiers of the disciplines. At the same time, it will provide important support and support for the work of scientific research management departments and relevant scientific and technological decision-making departments When researching scientific research layout, making scientific and technological decisions, and solving potential problems in a timely manner.In turn, it is conducive to promoting the allocation of the funding structure of the China Science Foundation project, and thus promoting innovative research.

The traditional methods of topic evolution analysis mainly include word frequency analysis method, co-citation analysis method and co-word analysis method. Qiu Junping0 used the word frequency analysis method to study the development law of the research papers of digital library disciplines in China, and classified the topic words and keywords, and summarized the research focus of the digital library. Xu Nuo[3] used multivariate analysis and social network analysis to conduct statistical and co-citation analysis of the core authors of anti-competitive intelligence papers, and studied the direct cooperation strength and relationship of authors in the field, and showed the field's status quo. The co-word analysis method uses the internal keywords of the literature to analyze the evolution of subject topics emphasizing the relationship between words and words . Wang Xiaohua[4] studied the topic of popular news headlines on Sina.com based on co-word analysis methods. Compared with the traditional scientometric method, the text mining method pays more attention to the content of the literature and achieve deep, fine-grained and all-round research on the literature. The LDA model makes up for the shortcomings of traditional text mining models that can not reflect the semantic relationship between vocabulary. It has been widely used in the fields of text classification, sentiment analysis, text topic mining, etc. In recent years, it has gradually been applied to literature and patents field to conduct topics discovery and evolutionary research. There are many researches on topic recognition in scientific and technological thematic research, but there are few studies on quantitative prediction of topics. This study proposes to use the LDA model to mine potential topics in science funds. We analyzes the distribution characteristics and evolution rules of the theme funded by the Science Foundation in 2001-2017, and then predict future technology trends. Finally, we take the science and technology management and policy disciplines as an examples. Apply the above combination method[5-6] to analyze the topic distribution, evolution law and future trends in the field of science policy and management.

## 2. METHODOLOGY

### 2.1 Pre-processing operations

Before building LDA model, pre-processing operations are performed, including word segmentation, de-stopping words, and removing noise from text[7].

(1) Word segmentation. Because the expressions of Chinese and English words are different, there are separated spaces in English between the words, but there is nothing between Chinese words. So

this article uses Chinese word segmentation software to perform Chinese word segmentation on the topics, abstracts and Chinese abstracts of the Science Fund.
(2) Remove the stop words. In the study, the stop words were removed by establishing a stop word list. Put the stop words you need to remove into the stop word list and then match them in the authoring program to remove the stop words.

### 2.2 Hot topic recognition

After the data is pre-processed, the pre-processed data is time sliced and classified into each time interval. Then, the thematic fund data of each time interval is analyzed in turn to obtain the distribution of the topics in each time interval .
The LDA (Latent Dirichlet Allocation) model is a three-layer Bayesian network model containing text-theme-terms proposed by Blei[8] in 2003 for statistical modeling of discrete data sets. LDA model can effectively reduce the dimension of text data and explore hidden topic information. The topic result is expressed as a series of word probability distributions. By maximizing the word co-occurrence probability to find word clustering and using Dirichlet distribution to describe the document generation process.It can simulate semantic information better for large-scale corpora. Assume that the science fund subject is subject to the superparametric Dirichlet prior distribution:

$$Dir(\theta / \alpha) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_{dk}^{\alpha_k - 1} \tag{2-1}$$

$\theta_{dk}$ indicates the distribution of the science fund d in the theme k of the science fund.
Each science fund theme k generates the distribution of subject terms, $\phi k$~Dir(β).Each science fund funded project d generates keyword distribution, $\theta d$~Dir(α).Generate a theme item for the nth item in each project *zdn*~Multinomial($\theta d$) and *wdn*~Multinomial($\phi_{z_{dn}}$).Therefore, in this study, the LDA likelihood model is:

$$p(W \mid \alpha, \beta) = \prod_{d=1}^{D} \int p(\theta_d \mid \alpha) \prod_{n=1}^{N_d} \sum p(z_{dn} \mid \theta_d) p(w_{dn} \mid \phi_{z_{dn}}) d\theta_d \tag{2-2}$$

The LDA topic model needs to determine the number of potential topics in the text in order to reduce the dimension[9], but the LDA method cannot automatically determine the number of potential topics. The[10]proposed the Hierarchical Dirichlet Processes, which can automatically determines the number of topics by the non-parametric model.This method is inefficient for large-scale text analysis, and it is difficult to ensure accuracy; Blei[11] proposes to use the confusion calculation method to determine the optimal number of topics. The lower the confusion, the greater the sentence probability and the more generalized the language model is.But this method leads to the similarity between topics is too large . In order to optimize the topic extraction effect, this paper determines the optimal topic number using the Perplexity-Var method, which considers the similarity and confusion between topics. The method measures the structural stability of the topic and punish excessive over-the-top topics through the divergence of the topic (DJS). It minimize the number of topics while ensuring that the differences between themes are maximized. Finally, using the Heinrich's[12] parameter estimation method, set $\alpha = 50/K$, $\beta = 0.1$。

## 3. DATA

The formulation and management of science and technology policies are of great significance to the promotion of national science and technology system management and innovation. Science and technology management and policy as a discipline also have a guiding and advancing role in the development of science and technology and the reform of science and technology innovation system. This paper uses the Science and Technology Information Sharing System to search for “Science and Technology Management and Policy”.It has obtained 113 project data from Science and Technology Information Sharing System on Science and Technology Management and Policy from 2008 to 2017. We use the title, keywords, and summary information of the fund project as the data set to perform LDA analysis.

The case study aims to verify the feasibility and effectiveness of the scientific funding theme evolution method proposed in this paper. It can reflect the content and direction of the National Natural Science Council (NSFC) funding for science and technology management and policy disciplines in recent years, as well as the issues that need to be addressed in the areas of technology management and policy. The research results provide guidance or reference for researchers and related science and technology management departments. Before conducting the science fund topics analysis and forecasting, we first time-divided the data obtained during the year. The 2008-2017 data is divided into five time intervals by year. Then, in order to clean the data before the experiment, we perform pre-processing operations such as word segmentation for each time interval. In order to obtain the required corpus, the JIEBA word segmentation tool of Python Chinese word segmentation component is used to segment the original text data into word segmentation and part-of-speech tagging in this paper . The JIEBA word segmentation tool performs well in both speed and accuracy. At the same time, the JIEBA word segmentation component also supports custom thesaurus and stop word filtering operations. Text preprocessing is a process that needs to be repeated. In this process, the word segmentation custom dictionary needs to be expanded and feature selected until the processing result satisfies the requirements of the model.

## 4. RESULTS

When modeling the theme of the science fund, the number of optimal topics is determined by the degree of confusion. The smaller the confusion, the stronger the generalization ability of the model. This study calculates the Perplexity value for 2 to 20 subject numbers. As shown in the figure, the experiment shows that when the number of topics is divided into 5, the confusion is the smallest, so the optimal number of topics is set to 5.

The number of subject terms under each topic in this article is set to 10. Each vocabulary is output in descending order of probability. The boundaries of the technical field between themes are clear. The classification effect is ideal. The table 3.1 shows the number of topics extracted in five time intervals. The table 3.2 shows the distribution of the keywords in 2014-2015.

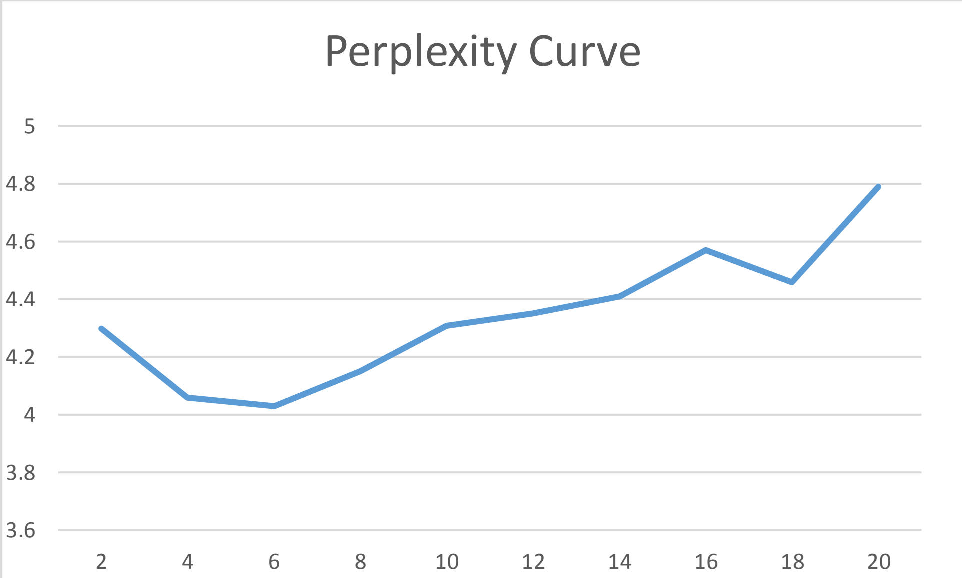

Fig. 1 Number of Topic(T)-Perplexity

Table 1 Subject recognition result

| 2008-2009 | 2010-2011 | 2012-2013 | 2014-2015 | 2016-2017 |
|---|---|---|---|---|
| Region | Policy | Innovation | Alliance | The internet |
| Manufacturing | Innovation | Industry | Innovation | Cooperation |
| Innovation | Industry | Technology | Science | Patent |
| Input | Technology | Policy | Industry | Research |
| Industry | Network | Path | Patent | Innovation |

Table 3.2 Distribution of topic words in 2014-2015

| Topic0 | Topic1 | Topic4 | Topic5 | Topic6 |
|---|---|---|---|---|
| Industry | Patent | Technology | Innovation | alliance |
| Integrated | Innovation | Science | The internet | strategy |
| Region | Value | Information | Research | Innovation |
| Strategy | Examination | Research | Independence | Conflict |
| GreenClean | Policy | University | Creativity | Competitiveness |
| Energy | System | Funding | Creative | Country |
| Innovation | Application | Policy | mechanism | Industry |
| West | Technical Standard | Cost | management | Sovereignty |

In 2008-2009, the fund in the subject of science and technology management and policy focused on "manufacturing", "regional development" and "innovation investment" . In 2010-2011, the funded hot topics focused on "technical innovation", "scientometrics" and "technology policy"; The hot topics in 2012-2013 are "Industrial Independent Innovation", "Innovation System", "Contract", "Policy Research and Industrial Technology Innovation", "Industrial Transformation", "Innovation Network" ; In 2014-2015, the funding focus is more on the "innovation network" and "patent system", and it is developed and extended to "industry alliance and green industry", "independent innovation

and policy research”, “innovation network analysis”, “technology funding policy”, “national innovation capability and strategic alliance”. ";In 2016-2017,the hot topics focus on the "innovation cooperation network and innovation ecosystem" "patent system" "production,study and research" .
It can be seen from the results that the National Natural Science Foundation's funding for science and technology policy and innovation disciplines is as focused as ever. The words “innovation”, “patents”, “industry” and “scientific and technological policies” are all appearing at high frequencies at all stages. But as time goes by, hot topics are also evolving and changing. With the changes and deepening of the direction of science funded projects, the hot words of “innovation network”, “integration”, “collaborative innovation” and “production, study and research” began to appear. It can be concluded that the National Natural Science Foundation of China has a large proportion of topics related to “innovation” in science and technology management and policy-funded projects. Innovation policy have always been the main theme. In recent years there has been a trend towards in-depth research on “innovation networks” and “innovative ecology”. At the same time, the trend pay attention to the methods and methods of "scientometrics" and attach great importance to the "production, study and research" alliance and the development and application of high-tech enterprises.

## 5. CONCLUSION AND DISCUSSION

In this paper, combined with the LDA theme model, a new method is proposed to study the main research topics the evolution path of the past decade of science and technology management and policy in the National Natural Science Foundation of China Science Foundation funded projects . Forecast the hot topics that may be funded in the future based on the scientific subject evolution path. The research results show that the research theme of the National Natural Science Foundation of Science and Technology Management and Policy Funding has always been related to the theme of "innovation". However, the research focus is constantly being adjusted and deepened, and the current situation is constantly expanding. By constructing a theme evolution path, it can reflect the changes in funding hot topics well. In recent years, funding for research on “innovation ecosystems”, “innovation capabilities and innovation performance evaluation” and “production, study and research” has increased. The results of this study will help the Natural Science Foundation to effectively evaluate the novelty and importance of the content of the application project in the management process. It helps to sort out the development of basic science disciplines. The results of the study will help optimize the allocation of funding structures for the China Science Foundation project and promote research in emerging disciplines. It can help to give full play to the natural science fund's leading role in prospering basic research and promoting independent innovation.
Through the case analysis, we can find that the method has better clustering effect. However, the cleaning effect of the text has a greater impact on the results. This study expands the stop words and builds a custom dictionary. However, due to its limitations on domain knowledge, further optimization is needed in data cleaning to further improve accuracy; it has certain subjectivity in the judgment of subject names. These shortcomings will be improved in conjunction with the latest advances in natural language processing. Strive to achieve automatic and accurate access to the topic

name. At the same time, it is necessary to reduce the complexity of the LDA model algorithm to achieve the research work of larger data sets.

## REFERENCES

[1] JiangHui Li,Shuo Wang. Let the Natural Science Foundation Become a Beacon to Lead Basic Research--An Interview with the National People's Congress and the Director of the National Natural Science Foundation of China, Academician Yang Wei [N]. People's Political Consultative Conference Newspaper,2013-03-10.
[2] ShuPing Qiu , MingZhi Wang.The Analysis of the Digital Library Research Paper in China from the Years of 1999 to 2008[J]. Journal of Intelligence, 2010, 29(2):1-5.
[3] Nuo Xu,JinTang Deng. Visual Analysis of China's Anti-competitive Intelligence Research Papers——Based on Citation Analysis from 1998 to 2010 [J].Information Science, 2012(5):725-730.
[4] XiaoHua Wang,Ning Xu.Text topic clustering and topic discovery in co-word analysis [J]. Information Science,2011(11):1621-1624.
[5] CaiDong Zhang. Research and Applications of Text Topic Evolution Model Based on LDA and HMM [D]. Xiamen University, 2013.
[6] Wei Chen,ChaoRan Lin.Analysis of the Evolutionary Trend of Technical Topics in Patents Based on LDA and HMM: Taking Marine Diesel Engine Technology as an Example [J]. Journal of the China Society for Scientific and Technical Information, July 2018, 37(7): 732-741.
[7] Hehang X , Xinjian G , Guohai C , et al. The patent mining analysis method based on Chinese word segmentation[J]. Science Research Management, 2011. 32(7)：138-142.
[8] Blei DM,NgAY,Jordan MI.Latent Dirichletal location[J]. Journal of Machine Learning Research,2003,3:993-1022.
[9] Peng G , Yuefen W . Topic Mining in Scientific Literature Based on LDA Topic Model and Life Cycle Theory[J]. Journal of the China Society for Scientific and Technical Information, 2015. 34(3):286-299.
[10] Teh YW,Jordan MI,Beal MJ,etal.Sharing clusters among related groups: hierarchical Dirichlet processes [C]//Proceedings of the Neural Information Processing Systems Conference, 2005: 1385-1392.
[11] BleiD. Probabilistic topic models[J].Communications of the ACM, 2012, 55(4):77-84.]
[12] Heinrich G. Parameter estimation for text analysis[R/OL]. http: // rakaposhi. eas.asu. edu/f12-cse571-mailarchive/pdflwcW7WccCL.pdf.