ISSN: 2688-8653

Research on the Selection of Hot Topics and the Analysis of Evolution Trends in the Discipline of Information Science

Shali Cao, Jianhua Zhang, Na Xin and Yulong Lv

School of Anhui University of Finance and Economics, Bengbu, 233000, China.

Abstract

With the continuous development of the digital economy, the value of data has gradually become prominent and has become an important driving force for social development. The National Social Science Foundation of China is the highest-level scientific research fund project in the field of humanities and social sciences, with strong authority and representativeness. This project excavates the selected topics in the field of information science of the National Social Science Foundation of China, which can determine the hot topics and their changing trends. It is conducive to fully grasping the highest level and development trend of domestic scientific research data research, and is of great significance for improving the project management level and formulating scientific development plans in the field of information science, which will greatly reduce the problem of "difficult topic selection".

Keywords

Information science, Hot spot, Development trends.

1. Project background

With the advent of the era of artificial intelligence and big data, all aspects of human daily life are inseparable from the network and information technology, the text information of the real world is more presented as electronic version, and text mining has also become a research hotspot and learning focus in the field of information. The recognition and analysis of massive texts using computers has become a hot topic in research.

This project selects the 2012-2023 National Social Science Foundation project in the field of information science as the research object, uses python to carry out theme text mining, and explores the theme and evolution process of related projects in the field of information science of domestic social science funds through analysis methods such as theme analysis, cluster analysis and evolution analysis based on LDA algorithm (LDA, Linear Discriminant Analysis), and identifies hot topics for comparative analysis. In order to reveal the hot scientific research projects of domestic information science discipline more objectively and comprehensively, it will form an effective reference for promoting the innovation and development of domestic information science discipline.

2. Research status and development trends at home and abroad

Foreign research on text mining was carried out earlier, and the United States is one of the earliest countries to standardize literature and scientific research evaluation activities. In 1914, the United States established the Congressional Research Department CRS, and in 1975 issued the "Project Evaluation Standards". At the end of the 50s, H.P. Luhn conducted groundbreaking research in this field, proposing the idea of word frequency statistics for automatic classification. In 1960, Maron published his first paper on automatic classification, and numerous scholars such as K. Spark, G. Salton, and K.S. Jones also carried out fruitful research in this field. At present, foreign text mining research has entered the practical stage from the

ISSN: 2688-8653

experimental stage, and the project evaluation agencies in the United States mainly include Congress, non-official institutions and universities themselves. With the multi-dimensional penetration of scientific research activities and the increasingly integrated structure of knowledge system connotation and extension, the establishment and improvement of university evaluation system bear the brunt of it.

With the continuous development of science and technology, scientific and technological innovation has gradually become the main driving force for economic and social development. With the great attention of all walks of life, the investment funds continue to grow. At the advent of the fourth industrial revolution, China put forward the concept of "Made in China 2025", which put forward higher requirements for China's various scientific research work. The evaluation methods used in China are similar to those used in foreign countries, qualitative analysis is based on peer review, quantitative analysis is based on econometric methods, and there is still a certain gap with foreign countries in terms of index system, standardization, scientificity, etc., which is related to the fact that the development of literature and scientific research results evaluation in China is still in a perfect stage.

To sum up, through text mining technology to mine the information of the National Scientific Research Fund projects, can simply and clearly reflect the key text and key data of these projects to extract important information in them, China's National Natural Science Foundation projects in 2023 are divided into 14 large projects The number of projects has exceeded 1 million, of which the amount of key projects is huge, but in the text mining of these projects, China has a large gap compared with world-class countries, taking these key projects as an example, The probability of text mining in these key projects is less than 1%, and many developed countries in the world have begun to pay attention to text mining of scientific research projects, especially the use of concise and efficient Python language to text mining their own scientific research fund projects. Therefore, in the field of national scientific research fund projects, the use of Python language for text mining has great potential.

3. Research objectives

Systematically sorting out and analyzing the project establishment data of scientific research data topics in the field of information science of the National Social Science Foundation is conducive to fully grasping the highest level and development trend of scientific research data research in the field of domestic information science, and is of great significance for improving the project management level and formulating scientific development plans in this field. This project selects the 2012-2023 National Social Science Foundation of China in the field of information science as the research object, uses Python to carry out theme text mining, and explores the theme and evolution process of the domestic social science foundation in the field of information science through analysis methods such as theme analysis, cluster analysis and evolution analysis based on LDA algorithm, and identifies hot topics for comparative analysis, so as to reveal the hot projects in the field of information science of the domestic social science fund more objectively and comprehensively. It is an effective reference for promoting the innovation and development of domestic information science.

4. Research content

This topic follows the research idea of "literature review→ obtaining text data based on Python data crawling→ implementing theme extraction of short text data files based on LDA model→ forming high-frequency keyword co-occurrence in an orderly and orderly data→ and conducting cluster analysis and evolution analysis of project topics in the field of information science of China's social science foundation→ forming a research trend orientation in the field

ISSN: 2688-8653

of information science of social science foundations, and providing reference for subsequent scholars' research topics.

Based on the high-dimensional characteristics of short text data, this project mainly discusses the preprocessing of the project selection in the field of information science of the China Social Science Foundation, first uses the LDA algorithm to cut words, clarifies the high-frequency words and research hotspots of the fund topics, and proposes to modify each type of theme in combination with the background influence of potential political and cultural theories and the development of the environment of the times. Through the data visualization method, the topic selection tendency of the information science field projects of the China Social Science Foundation is studied, and the evolution law of the research development in the field of social sciences is obtained, which further provides a new perspective and reference for future social science research.

4.1. Research hot keyword analysis based on word frequency

Word Frequency Analysis is the statistics and analysis of the number of occurrences of important words in the body of the document, and is an important means of text mining. It is a traditional and representative content analysis method in bibliometrics, and the basic principle is to determine hot spots and their change trends through the change of frequency of words.

4.2. Research topic findings based on the LDA model

LDA was proposed by Blei, David M., Ng, Andrew Y., and Jordan in 2003 to speculate on the subject distribution of documents. It can give the topics in the academic literature set related to the discipline of information science in the form of probability distribution, so that after analyzing some documents to extract their topic distribution, it can carry out topic clustering or text classification according to the topic distribution.

4.3. Research on the evolution analysis of hot topics

Collect the projects in the field of information science in the database database of national social science projects from 2012 to 2023, extract research topics through LDA theme model, identify hot topics based on the theme life cycle, construct the evolution path of themes combined with time slices, and analyze the hot topics and evolution trends of information science research from the theoretical and application dimensions of information science research.

5. Research innovation points

5.1. Innovation in research perspective

In terms of research perspective, few scholars have conducted systematic research on the projects established by the National Social Science Foundation in the field of information science. Therefore, the topic of this project is relatively novel, the research perspective is unique, and it can further study the related research and application in the field of information science, which has certain academic value.

5.2. Innovation in research methods

In terms of research methods, most of the existing literature adopts content analysis method and quantitative description method, but this topic can extract information from unstructured text, identify the subject in the document, and mine the hidden information in the text, and deeply analyze the development status of information science data research in China by collecting and combing the data of relevant literature in the discipline of information science, and using the theme text analysis method based on LDA algorithm. The follow-up research on information science data, as well as the selection, application, organization and management of information science disciplines are of great significance. Compared with the traditional method

ISSN: 2688-8653

of studying relevant literature of national intelligence discipline, our text mining of relevant literature of national intelligence discipline based on Python improves the efficiency of research, and the use of Python programming for text data mining has the characteristics of fast speed and accurate mining, which can greatly improve the efficiency of the commonality, difference and trend research of related social science projects.

6. Key scientific questions to be solved

6.1. LDA model construction and implementation

LDA is an unsupervised machine learning technique that can be used to identify subject information latent in a large-scale document collection or corpus. It uses the bag of words approach, which treats each document as a word frequency vector, transforming textual information into digital information that is easy to model. After obtaining the literature themes of the 2012-2023 National Social Science Foundation in the field of information science, the obtained results were visually analyzed by social networks and overlay graphs.

6.2. Theme evolution visualization

Theme evolution is a change over time, including two aspects: (1) the theme changes over time intensity; (2) The subject content changes over time. How to track the subsequent development of research topics is a problem that we are concerned about and need to solve urgently. Over time, the content of the research topic will change, and the intensity will also go through a process from high tide to low tide. It is of practical significance to organize these large-scale documents at scale and obtain the evolution of topics in the project topic collection in chronological order to help scholars in the field of information science track topics.

7. Research protocol

7.1. Data Sources and Preprocessing

The projects in the field of information science of the National Social Science Foundation adopted in this project are from the National Social Science Project Database, and the topics of library, information and philology in the National Social Science Project Database from 2012 to 2023 are selected, with a total of 2535 projects (the title, keywords, and abstract of each project are downloaded to represent a project). (1) First, use the jieba Chinese word segmentation tool to preprocess the document, and divide each item into independent words; (2) In natural language processing, certain words and words are removed in advance to improve the efficiency of project processing.

7.2. Python data crawling mining method

Firstly, based on Python, the data crawling and combing of the National Social Science Foundation project in the field of information science in China from 2012 to 2023 are carried out, and the short text of the research topic is segmented, sorted, sorted and clustered by using the Jie Ba library to form the data basis of this topic.

7.3. LDA Algorithm Topic Extraction Method

LDA, proposed by Blei in 2003, is a discrete dataset modeling topic generation model used to estimate the distribution of document topics, grouping models based on the frequency of different words occurring together. It is a three-layer Bayesian network structure composed of a document layer, a topic layer and a word layer, the core of which is that a document contains multiple topics (Topics), and each word (Word) is generated by a fixed topic. The LDA topic model can be used to convert text information into data information, so that the topic of each document in the document set is given in the form of probability distribution. The LDA

ISSN: 2688-8653

algorithm is used to extract multiple clustering topics in the document to achieve topic clustering or text classification, in which each topic contains the words with the highest common frequency, which can reflect the corresponding research hotspots and orientations.

7.4. Prediction methods for future hot keywords

According to the statistical results of the hot keywords of "all-discipline journals" appearing nearly 10 years earlier than the hot keywords of "information science journals", the references, G30 (scientific research theory), G31 (scientific research work), G32 (scientific research projects in various countries around the world), G35 (information science, information work), G20 (information and communication theory), G23 (publishing business), G25 (library science, library business), G27 (archival science, archival business), TP39 (computing technology, Computer Technology) 9 classes to establish a relatively specific "reference to a whole discipline journal" to predict the scope of statistical experiments. First, these 9 categories were used to search for relevant journals, and the relevant journals retrieved were called "reference all-discipline journals". Count all keywords of "reference to all-discipline journals", and finally select hot keywords from these keywords, the statistical time is 2000-2020, the top 300 keywords that appear every year as hot keywords, a total of 833 keywords are selected, and 4516745 times. Among them, 258 keywords appeared after 2013. The distribution of 258 keywords appearing in "reference to all-discipline journals" after 2011 in "reference all-discipline journals" and "information science journals" was counted, and the distribution of 258 keywords in "reference all-discipline journals" and "information science journals" was compared. The 258 keywords appeared 624 873 times in "References to All-Discipline Journals" and 174 keywords appeared in "Information Science Journals" with 630 occurrences.

8. Conclusion

This paper statistically analyzes the keywords of domestic information science journal papers published from 2012 to 2023, obtains 60 hot keywords, and divides the 60 hot keywords into three stages: 2015 and before, 2015-2020, 2020-2023. The hot keywords of the three stages represent the research focus and basic evolution trajectory of intelligence journal papers. Through the comparative analysis with the hot keywords of the whole discipline, it is found that the hot keywords of the information science journal papers also have unique hot keywords that are different from the keywords of the whole discipline, and some hot keywords of the information science journal papers lag behind the keywords of the whole discipline by 10 years. This characteristic reflects the law of the field characteristics and overall characteristics of the field discipline and the whole discipline, and also reflects the characteristics of the hot keywords of the information science journal articles closely following the technological development of the whole discipline and the continuous introduction and innovative development of the social demand.

Acknowledgements

This work is supported by Anhui University of Finance and Economics Undergraduate Research innovation fund project fund, Project number: XSKY23027ZD.

References

[1] Scientific research evaluation and its reference in American universities[J].Management Observation,Lu Yiyi,Guo Shengwei,2016(21).

[2] Gross PF. A critical review of some basic considerations in post-secondary education evaluation[J]. Policy Sciences.1973.4（02）：171-195.)

ISSN: 2688-8653

[3] China Science Foundation,Feng Yueqiang,Qi Wei,2007(1):14-16.)

[4] Research Project Funding Management:A Comparative Study of China and the UK[J].Research Management,Gu Quan,2012,33(1):120-126.)

[5] ZHOU Peng,ZHANG Min,GUO Shengwei. Analysis of the historical evolution, current situation and countermeasures of scientific research evaluation in China[J].Management Observation,2016(32):173-176.)

[6] GONCALVES ANTONELLA HONORIA IMANE. Extraction and prediction of hot topics in scientific management based on text mining and potential Dirichlet assignment[D].Harbin Institute of Technology,2018.)

[7] LI Chang,YI Huifang,WU Hong,JI Fangyan. Theme analysis of patent technology of unmanned car:Based on WI-LDA theme model[J].Journal of Intelligence,2018,37(12):50-55.)

[8] TAN Xu, ZHUANG Muni, MAO Tatian, ZHANG Qian. Analysis of large-scale network public opinion emotion evolution based on LDA-ARMA hybrid model[J].Journal of Intelligence,2020,39(10):121-129.)

[9] ZHAO Kai, WANG Hongyuan. Research on LDA optimal topic number selection method:A case study of CNKI literature[J].Statistics and Decision,2020,36(16)

[10] WANG Xiwei, ZHANG Liu, HUANG Bo, WEI Yanan. Construction and empirical research on the theme map of microblog users based on LDA:A case study of "Egyptian Aviation Difficulty"[J].Data Analysis and Knowledge Discovery,2020,4(10):47-57.)

[11] WEI Xueyong. Bayesian quantile regression model and its application[D].Guizhou Minzu University,2022.)

[12] DING Yue, WANG Xueming. Naive Bayes classification algorithm based on improved feature weighting[J].Application Research of Computers,2019,36(12)

[13] Si Dan, Liu Dongsu. Research on text classification method based on weighted Word2vec[J].Information Science,2019,37(11):38-42.)

[14] Du Zengwen. Research on microblog theme detection model based on Dirichlet regression[D].University of Chinese Academy of Sciences (School of Artificial Intelligence), University of Chinese Academy of Sciences,2020.)