IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 14, 2019, accepted August 25, 2019, date of publication August 29, 2019, date of current version September 13, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2938393

Research Hotspot Prediction and Regular Evolutionary Pattern Identification Based on NSFC Grants Using NMF and Semantic Retrieval

JINLI WANG[1], YONG FAN[2], LIBO FENG[3], ZHIWEN YE[3], AND HUI ZHANG[3]

[1]Faculty of Management and Economics, Kunming University of Science and Technology, Kunming 650093, China
[2]Institute of Humanities and Social Sciences, Kunming University of Science and Technology, Kunming 650093, China
[3]State Key Laboratory of Software Development Environment, Beihang University, Beijing 100191, China

Corresponding authors: Yong Fan (2748803741@qq.com) and Libo Feng (fenglibo@buaa.edu.cn)

This work was supported by the National Key Research and Development Program of China under Grant 2017YFB1400200.

ABSTRACT Analyzing the research hotspots and regular evolutionary patterns of R&D projects can help researchers find potential information. This paper considers the titles of the National Natural Science Foundation of China (NSFC) grants that have been awarded in the past 20 years as the research objects. First, we analyze the number and funding amounts of project grants for each department over the past 20 years. Second, we propose a topic discovery method that is based on nonnegative matrix factorization and further propose a keyword scoring method that is based on semantic retrieval, from which we can discover the regular evolutionary patterns of research hotspots. Finally, we conduct experiments on datasets on grants in the Department of Information Science. To identify research characteristics, trends and prospects, we explore the regular evolutionary patterns of research hotspots via experiments using three methods from multiple perspectives: word cloud, topic corresponded to keywords display and topic evolution display. The results of the experimental studies demonstrate that researchers can keep abreast of the main content and hotspots in the research field via hot spot discovery and the identification of regular evolutionary pattern of NSFC grants. This study can also help the government improve the allocation of scientific and technological resources and help decision makers make scientific decisions.

INDEX TERMS NSFC, nonnegative matrix factorization, hotspot prediction, regular evolutionary pattern, semantic association.

I. INTRODUCTION

The sustainable development of science and technology plays an important role in promoting collaborative innovation in a country and a region [1].Almost every country invests enormous human, financial and material resources in sustainable development and innovation in science and technology [2]–[5]. The National Natural Science Foundation of China (NSFC) is the most influential and wide-ranging national-level research project of China. It covers many of the basic research fields that attract the vast majority of the experts and scholars who have made substantial contributions to China's science and technology [6]. The projects of NSFC grants are important national science and technology resources. If we can make full use of them and explore the titles of the NSFC grants and the relationships among them, it will bring inspiration to researchers. However, there are currently few studies on scientific research projects, and project information mining is not sufficient. This paper considers the titles of NSFC grants over the past 20 years as the research object. First, we count the number of projects that have been funded in the past 20 years and, the total number and funding of the projects that were established by each of the eight departments, with the objective of obtaining a macroscopic understanding of the NSFC. Second, we propose two methods for dealing with NSFC datasets, namely, nonnegative matrix factorization (NMF) [7], [8] and the keyword retrieval (KR) [9] method based on semantic retrieval [10]. NMF is mainly used to mine hot words and to identify research hotspots. KR is mainly used to discover the evolutionary laws of research hotspots. The combination of the two methods can more accurately predict the research hotspots, development trends and

The associate editor coordinating the review of this article and approving it for publication was Lu An.

prospects in various fields [11]–[13]. Finally, we conducted experimental verification based on theoretical analysis.

We used four methods to analyze the results of the experiments: The first is segmentation information statistics, which can determine the complexities and granularities of the titles of grants projects. The second is word cloud displays which can visually display the hot words in research over the years. The third is topic display, which uses nonnegative matrix methods to analyze the distributions of hot topics and the corresponding keywords over the years. The fourth method is the evolution analysis of hotspots, which is based on semantic computing and uses the keyword scoring method, which is based on semantic retrieval, to obtain the scoring table of hot words in the past years to facilitate a more scientific and intuitive understanding of the proportion and evolution of hot words. Each of these four methods has its own advantages. The analysis and display of segmentation information statistics and word clouds can help readers and researchers focus on research hotspots and, predict research content and development trends.

Therefore, the main contributions of this research are as follows: First, we consider the titles of NSFC grants as the research objects and analyze the number and the funding amounts of NSFC grants over the past 20 years. Second, we extract and analyze the titles of NSFC grants, extract hot words from various research fields, and discover the evolution of research hotspots. Third, through this research, we can discover the key knowledge and the potential major opportunities that are hidden by using technical resources to provide decision support for technology developers and to provide predictions and references for project reports and researchers. Finally, the results that are presented in this paper can help researchers understand the research characteristics and development trends in this research field, and provide a reliable basis for project application and selection of topics, which will facilitate the smooth progress of applications, improve the quality of scientific research and innovation, and promote the sustainable development of science and technology.

The remainder of the paper is organized as follows: Section II reviews the literature on hot spot prediction and the sustainable development of science and technology and its applications. Section III describes the evolution of NSFC grants in terms of quantity and funding amount. Section IV presents the problem statement. Section V presents the theoretical basis and model, which mainly uses NMF and semantic retrieval methods to explore the titles of the NSFC grants and the relationships among them, which are also the theoretical basis for hot topic display and identification of the regular evolution patterns of hotspots. Section VI Section presents the flowchart and evaluation metrics of our methods. Section VII describes the process and the steps of the experiments and presents the results in the form of charts, texts and discussions. Finally, section VIII presents the conclusions of this paper and discusses future research directions.

II. RELATED WORKS

The sustainable development of science and technology has a huge impact on social progress [14]. Increasingly many scholars have identified changes in the development of science and technology via the study of papers and patents. Kim *et al.* [15] investigated sustainable technology development in the humanoid robot industry via research on international patents. Lee and Sohn [16] identified research trends in financial commerce via research on patents. Bai [17] realized the recommendation of scientific papers and research fields via the investigation of scientific papers. Norambuena *et al.* [18] proposed a sentiment analysis and opinion mining method that is based on scientific research papers. Waheed *et al.* [19] proposed a hybrid method for the recommendation of scientific research papers and promoted the full use of the papers. Through the analysis and mining of Korean patent information, Lee *et al.* [20] used topic modeling and Latent Dirichlet Allocation (LDA) to identify research opportunities and to predict research hotspots. Yuan *et al.* [21] proposed a novel dada perspective that was based on NSFC grants and identified international research collaboration partners for China. Nichols *et al.* [22] proposed a topic model approach for measuring interdisciplinarity of the national science foundation, Chen *et al.* [23] proposed a co-word analysis method to identify the intellectual structure from research grants. The above literatures take papers or research projects as research objects.

The analysis of scientific resources such as papers, patents, and projects involves the analysis and mining of short texts. Short text management is a key aspect of information management. It plays an important role in the rapid development process of artificial intelligence. Topic modeling is an important technology of short text management. For short text analysis, domestic and foreign scholars have conducted in-depth research and have proposed many practical algorithms and solutions. Via topic modeling, one can classify and manage massive semi-structured data and it plays a role in hotspot prediction and decision-making. Lee and Seung [24] proposed a nonnegative matrix factorization approach for large-scale text processing. Chen *et al.* [25] put forward a soft orthogonal nonnegative matrix factorization method with sparse representation for tracking the static and dynamic topics. Zhuang *et al.* [26] proposed a method for semantic feature learning for heterogeneous multitask classification via nonnegative matrix factorization.

Nonnegative matrix factorization is an effective method for mining shor-text topic models [27]. It maps the text to be processed into a high-dimensional matrix and reduces the dimensionality via matrix decomposition to realize the association between the topic and related words. Based on the standard nonnegative matrix factorization, methods such as sparse NMF [28], graph NMF [29], semi NMF [30], and orthogonal NMF [31] have been proposed, which can be used to apply NMF to text mining [32], image detection [8], signal processing [33], medical inspection [34] and other fields.

Pari [35] has investigated several social media platforms through thematic models and has identified emerging themes that help users discover new ideas and play an important role in public opinion analysis and tracking. Liu *et al.* [36] realized semisupervised community detection using nonnegative matrix factorization.

On hotspot prediction and evolution, many scholars have carried out related research. Liang *et al.* [37] proposed a fuzzy multilevel algorithm that uses particle swarm optimization (PSO) to optimize support vector regression machine (SVR) to realize the real-time dynamic evaluation of drilling risk and hotspot prediction. Li *et al.* [38] proposed a data locality optimization method that is based on data migration (DLO-Migrate) and a data locality optimization algorithm that is based on hotspot prediction. Bencatel *et al.* [39] aimed at reviewing the existing knowledge on mammalian terrestrial carnivores in Portugal to analyze research trends. Xiao *et al.* [40] proposed a prediction method for social hotspots that is based on dynamic tensor decomposition. Yang *et al.* [41] used urban and social media data fusion for crime hotspot prediction. Xia *et al.* [42] introduced an efficient approach that uses support vector machine (SVM) to predict hot spot residues in protein interfaces. Adepeju *et al.* [43] proposed a novel evaluation metric for sparse spatiotemporal point process hotspot predictions.

This paper used NMF as the topic mining method. To evaluate the quality of the topic, we compare three methods (baseline) with our approach: PCA, SVD and LDA. Principal component analysis (PCA) [44] is a method for studying the main components of a document, which is often used for matrix dimensionality reduction. Singular value decomposition (SVD) [45] can obtain a low rank approximation matrix via matrix decomposition, which could be applied for signal denoising and big data mining. Latent dirichlet allocation (LDA) [46] is a topic model for processing documents, which can be used to identify hidden topic information in a large document collection or corpus and can be applied for data dimensionality reduction and classification. We conduct experiments on datasets in terms of topic coherence and PMI score. The results show that the performance of our method is better than those of three methods.

III. CURRENT SCENARIO OF THE NSFC

This paper conducts a statistical analysis of each year's NSFC project grants from 1999 to 2018. Due to a lack of data, the statistical information in this paper may differ slightly from the actual data. The data are obtained from the NSFC website (http://www.nsfc.gov.cn).

We examine the project data of the eight departments of the NSFC. The numbers of NSFC grants are plotted in FIGURE 1.

According to FIGURE 1, the number of projects that were established by each faculty has increased linearly each year. Two special curves in the plot do not follow this trend: the curves for the Department of Life Sciences and the Department of Medical Sciences. The number of projects that were

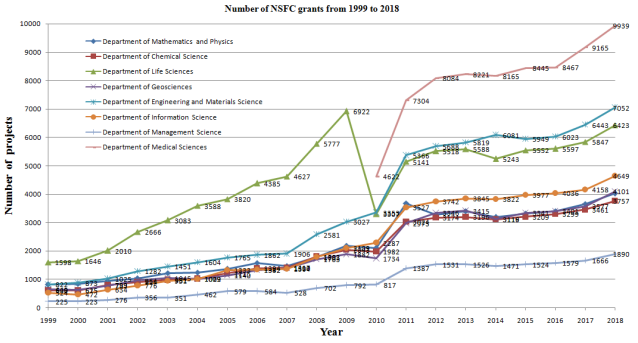

FIGURE 1. Number of projects that were established by the NSFC for each department.

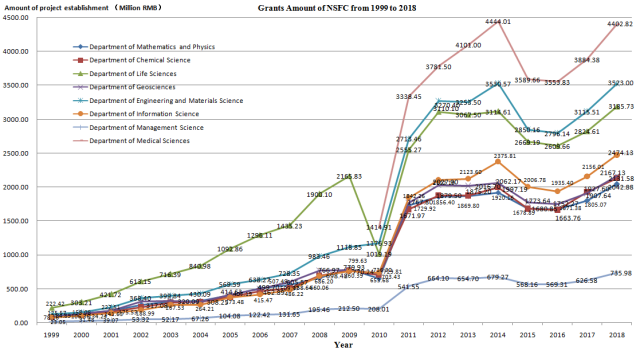

FIGURE 2. Amounts of grants that were established by the NSFC for each department.

established by the Department of Life Sciences fell suddenly in 2010 from 6,922 to 3,307 and the curve for the Department of Medical Sciences does not appear until 2010. This is because the Department of Medical Sciences was established by the NSFC in 2010, which separated it from the Department of Life Sciences. The separated Department of Medical Sciences has become a key area that is funded by the NSFC and the number of grants is increasing every year. By 2018, the number of grants had reached 9,039, which accounts for almost 25% of the total funding.

FIGURE 2 plots the changes in funding for each department from 1999 to 2018. In this figure, the abscissa axis represents the year and the ordinate axis represents the funding. The amounts of funding are plotted in FIGURE2.

According to FIGURE 2, the lowest amount of money has been invested in the Department of Management Science and the highest has been invested in Department of Medical Sciences. The grants funding for each NSFC department follows a linear growth trend. The Department of Information Science received funding of only 78.16 million RMB in 1999. By 2018, its funding had reached 2,474.13 million RMB. The average funding for each item increased from 149.1 thousand RMB in 1999 to 532.1 thousand RMB in 2018. However, grant funding for all faculties suddenly dropped by approximately 20% in 2015. According to investigations and interviews, this was mainly because indirect funds were

deducted from the total funding in 2015, which led to a significant reduction in funding for all NSFC projects.

IV. PROBLEM STATEMENT

In this section, we formalize the titles of each NSFC project and convert it into a vector form that the computer can handle.

The NSFC grant data consist of the person in charge, the title, the amount, the type, and the years of the study. We preprocess the project data and only retain the project title. The project title is a document and all items of the project in the current year constitute a set of documents. Let X={X_1,X_2, ... X_i, ... X_n} be the set of all documents. Given any document, we process the word segmentation. After removing the irrelevant adjectives and adverbs, the system will construct a dictionary of the document R. Then, the system will traverse each record X_i in document X, compare it to the dictionary, and construct matrix V according to formula (1). Let V be a set of corresponding word segmentation results. Matrix V is constructed as follows:

$$V_{ij} = \begin{cases} 0, & if\ X_{ij} \notin \mathrm{R} \\ 1, & if\ X_{ij} \in \mathrm{R} \end{cases} \tag{1}$$

All X_i are traversed to forming the word vector sparse matrix V of the document. The matrix is n-dimensional with m samples and each element in the matrix is nonnegative, namely, $V_{ij} >= 0$.

V. TOPIC MODELING AND KEYWORD SCORING

In this section, we will propose the model for analyzing the research hotspots and the evolution trends of NFSC grants. We propose an NMF method for solving the hot knowledge discovery problem of short-text information and expound its theoretical basis and algorithm. Then, a keyword retrieval scoring model that is based on semantic query is proposed, which can effectively identify the regular evolution of research hotspots.

A. NON-NEGATIVE MATRIX FACTORIZATION

This subsection introduces the basic theory and algorithms of standard nonnegative matrix factorization, which will lay the foundation for experiments.

Standard nonnegative matrix factorization is an effective tool for reducing the dimensionality of high-dimensional data which was proposed by Lee [22]. The strategy is illustrated in FIGURE 3. In this figure, V is a set of documents, each record in W represents a topic and each column in H is a subject word that is associated with it.

According to the basic strategy that is illustrated in FIGURE 3, we perform matrix decomposition on matrix V; the formula can be expressed as

$$V_{n\times m} \approx W_{n\times r} \times H_{r\times m} \tag{2}$$

where W is the basic matrix and H is the coefficient matrix in which r is smaller than n and m, namely, $r \ll n$ and $r \ll m$. By using the coefficient matrix instead of the original data matrix, dimensionality reduction of the original data matrix

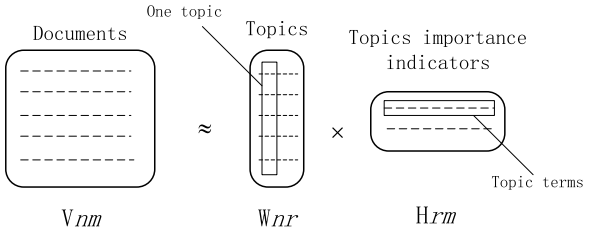

FIGURE 3. Nonnegative matrix factorization strategy.

can be realized. The iterative matrix can be expressed as

$$W_{ik} = W_{ik} \times \frac{(VH^T)_{ik}}{(WHH^T)_{ik}} \tag{3}$$

$$H_{kj} = H_{kj} \times \frac{(W^T V)_{kj}}{(W^T WH)_{kj}} \tag{4}$$

Consider the unconstrained optimization problem of minimizing the loss of the matrix:

$$\min D_E(V||WH) = \frac{1}{2}||V - WH||_2^2 \tag{5}$$

The gradient descent method is applied, which can be expressed as

$$w_{ik} \leftarrow w_{ik} - \mu_{ik} \frac{\partial D_E(V||WH)}{\partial w_{ik}} \tag{6}$$

$$h_{kj} \leftarrow h_{kj} - \eta_{kj} \frac{\partial D_E(V||WH)}{\partial h_{kj}} \tag{7}$$

where

$$\frac{\partial D_E(V||WH)}{\partial w_{ik}} = -[(V - WH)H^T]_{ik} \tag{8}$$

$$\frac{\partial D_E(V||WH)}{\partial h_{kj}} = -[W^T(V - WH)]_{kj} \tag{9}$$

The above equation is converted to a multiplication calculation according to the gradient descent method:

$$\mu_{ik} = \frac{w_{ik}}{[WHH^T]_{ik}} \tag{10}$$

$$\eta_{kj} = \frac{h_{kj}}{[W^T WH]_{kj}} \tag{11}$$

Then, the gradient descent algorithm is reexpressed as a multiplication algorithm.

$$W_{ik} = W_{ik} \times \frac{(VH^T)_{ik}}{(WHH^T)_{ik}} \tag{12}$$

$$H_{kj} = H_{kj} \times \frac{(W^T V)_{kj}}{(W^T WH)_{kj}} \tag{13}$$

The condition for convergence is the stability of both W and H. The stabilities of W and H are determined as follows. Assuming the number of iterations is t, the state of the W and H matrices at time t are W^t and H^t, the states at time $t+1$ are expressed as W^{t+1} and H^{t+1}. For any small real number ε, if the following are satisfied,

$$||W^{t+1} - W^t|| < \varepsilon \tag{14}$$

$$||H^{t+1} - H^t|| < \varepsilon \tag{15}$$

then it is concluded that W and H have converged.

We can also use the Euclidean distance to judge whether W and H have converged.

Assume that at time t, the Euclidean distance between H and W is $D_E(W^t, H^t)$ and at time $t+1$, the Euclidean distance between H and W is $D_E(W^{t+1}, H^{t+1})$. Then, the converge condition for matrices W and H is

$$||D_E(W^{t+1}, H^{t+1}) - D_E(W^t, H^t)|| < \varepsilon \tag{16}$$

where ε is an arbitrarily small real number.

According to the results of the previous subsection, the algorithm for standard NMF is presented as Algorithm 1.

Algorithm 1 Standard NMF

```
Input:document set matrix V
Output:topic matrix W, implicit keyword matrix H
Begin
    While not converged do
        Update w with formula(12);
        Update h with formula(13);
      i = i +1;
      End
End
```

B. KEYWORD SCORING METHOD BASED ON SEMANTIC RETRIEVAL

The topic model, which is based on NMF, can identify hot knowledge in the research field and can also identify the relationships between the research field and the keywords; however, it cannot clearly monitor the evolution of hot knowledge in the research field. Therefore, we propose a keyword scoring method that is based on semantic retrieval. The method can obtain the number of occurrences of vocabulary that is related to the meaning of a word via keyword retrieval, calculate the proportion of appearances of the word each year, and visually examine the regular evolutionary pattern of the research hotspot.

In this subsection, we propose a method for associating keywords and documents that is based on semantic retrieval, which uses semantic calculations to determine how often each keyword appears in the entire document. Keywords typically consist of hot words from the annual project research. By querying the vocabulary frequency of the keyword list in each document, it is possible to effectively evaluate the research on hot words in that year. The semantic retrieval method utilizes the *lucene* retrieval scoring formula. The steps are as follows: First, by querying the keywords, we will obtain the scores between the keywords and each item to identify the relationship between the keywords and the hotspots in the research field. Second, we can obtain the proportion of the keywords for each year. Finally, by normalizing each research hotspot according to the year, we can obtain the evolution law of the research hotspot.

The *lucene* scoring formula is as follows:

$$score(q, d) = coord(q, d) \times queryNorm(q) \times \sum_{tinq} tf(tind) \times idf(t)^2 \times t.getBoost() \times norm(t, d) \tag{17}$$

where t denotes the term; *coord(q,d)* indicates that the more search terms are included in a document, the higher the score for this document; and *queryNorm(q)* represents the variance of each query item. This value does not affect the ordering; it only causes the scores between queries to be comparable; *tf(t in d)* represents the frequency with which term t appears in document d; *idf(t)* indicates the documents in which term t has appeared; and *norm(t, d)* is a normalization factor.

According to the characteristics of our experimental entries, the formula can be reexpressed as follows:

$$score(q, d) = \cos(\theta) = \frac{1}{\sqrt{\sum_{tinq} idf(t)^2}} \times \sum_{tinq} tf(t, d) \times idf(t)^2 \times \frac{1}{\sqrt{num\ of\ terms\ in\ field\ f}} \tag{18}$$

Based on the keyword query results, we define the normalization formula. The formula is expressed as the number of times a keyword appears in a year divided by the total number of occurrences of the keyword, as follows:

$$p(keywords, year) = \frac{number(i, year)}{\sum_{i=1}^{N} number(i, year)} \tag{19}$$

Finally, we can derive the evolution trend of a keyword according to the year.

VI. FLOWCHART AND EVALUATION METRICS

In this section, we present an experimental flow that is based on NSFC text processing and mining and the evaluate metrics.

A. FLOWCHART FOR OUR METHOD

Based on the data on the NSFC projects that have been funded over the years, we will study the titles of the projects as the research objects, extract relevant statistical information, and conduct various experimental demonstrations to investigate the continuous research scenario in the hotspots of science and technology. A variety of technical knowledge is used in the process of short-text mining of project topics, e.g., word segmentation, topic modeling, keyword analysis, nonnegative matrix factorization, text mining, semantic search and application analysis. FIGURE 4 illustrates the procedure of our proposed method.

According to FIGURE 4, our method is divided into 10 steps:

- Step1. Obtain the NSFC grants that have been awarded over the years as a dataset, which can be obtained from the NSFC website via crawler technology.
- Step2. Preprocess the items of NSFC grants and remove irrelevant columns, such as those that correspond to

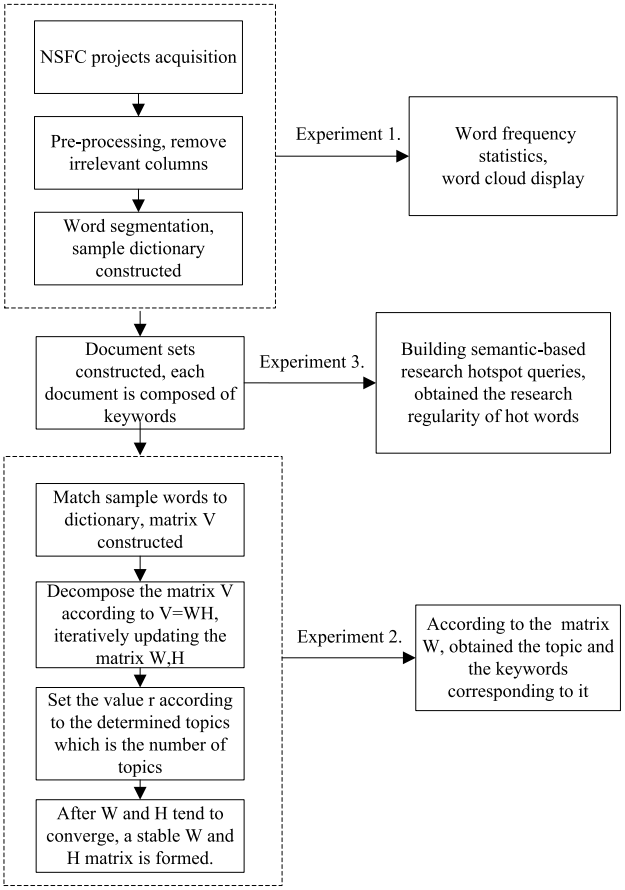

FIGURE 4. Procedure of short text mining of NSFC project titles.

the project time, the project leader, the project funding amount, and the research period.

- Step3. Process the items in Python language and use the Jieba package of the Python system for word segmentation. Then, remove the adverbs, auxiliary words, and adjectives from the document, retain the nouns, and collect all the nouns to form dictionary.
- Step4. The frequency of each noun is counted and the word cloud map is displayed according to the number of keywords. The word cloud map display can visualize the distribution of research hotspots; however, it is not accurate.
- Step5. Compare the vocabulary in each document with the dictionary. If a word appears, it is marked as 1; otherwise, it is recorded as 0. Via this approach, a high-dimensional sparse matrix V is constructed.
- Step6. Matrix decomposition is conducted according to $U = W * H$, where W is the set of topics and H is the set of keywords that match it. The iterative formula is based on equations (12) and (13) of 4.3.
- Step7. Determine the value of R, which is the topic's value. According to the requirements, the value of R can be set as $R = [20,40,60,80,100]$.
- Step8. If W and H converge, stable W and H matrices are formed. The convergence conditions for W and H are expressed in equations (14) and (15).
- Step9. After W and H have stabilized, matrix W is the R topics and matrix H represents the keywords that correspond to these topics.
- Step10. Construct a semantic-based research hotspot query. According to the *lucene* scoring formula (18) and the normalization formula (19), the score values of the keywords and the distributions of the words according to the year are obtained and the hotspot field evolution of the NSFC is identified.

B. EVALUATION METRICS

To evaluate the performance our model, first, we use two metrics to evaluate the topic quality: topic coherence and PMI. Second, we conduct the following experiments on our datasets:

(1) Topic coherence

Topic coherence is a famous metric for measuring the quality of a topic, which was proposed by Mimno *et al.* [47]. It is expressed as follows:

$$C(t; V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})} \tag{20}$$

where $D(v)$ is the document frequency of word type v and $D(v, v^{'})$ is the co-occurrence frequency of word types v and $v^{'}$; $V(t) = (v_1^{(t)}, v_2^{(t)}, \ldots v_M^{(t)})$ is a list of the M most likely words for topic t; and $v_m^{(t)}$, $v_l^{(t)}$ represent the mth and lth terms, respectively, topic t.

(2) The point wise mutual information (PMI) score [48] is another highly popular metric for measuring the quality of a topic model for a comprehensive assessment. The PMI score for each topic is the median PMI for all pairs of words for a topic. It is expressed as follows:

$$PMI(w_i, w_j) = \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \tag{21}$$

where w_i, w_j are the top k lists of words for the topic, $p(w_i, w_j)$ denotes the probability that words w_i and w_j appear in the same document, and the $p(w_i)$ and $p(w_j)$ denote the probabilities that the mth and lth terms, respectively, occur in the document.

VII. EXPERIMENTS

In this section, based on the previous theoretical analysis, the Department of Informatics grant data of the NSFC is regarded as the research object. We conduct experiments using three methods, namely word cloud, topic display and semantic, from various perspectives and examine the evolution of research hotspots over the past 20 years.

A. DATASETS

Our experiments are conducted on twenty real-world short text datasets. Our experimental datasets consist of

TABLE 1. Segmentation results for the project titles of the department of information sciences.

Year	Totalline	Average length	Dictionary length	Line_maxlength
2000	423	**2.17**	95	5
2002	729	**2.48**	146	9
2004	960	**2.76**	188	9
2006	1287	**2.92**	260	8
2008	1729	**3.12**	348	10
2010	2236	**3.37**	437	10
2012	3484	**3.66**	639	11
2014	3792	**3.81**	695	12
2016	4001	**3.86**	724	10
2018	4604	**4.05**	820	10

20 databases, which are the lists of the NSFC grants that are awarded each year. We named them NSFC 1999—NSFC2018.

The NSFC includes eight departments: the Department of Mathematics and Physics, the Department of Chemistry, the Department of Life Sciences, the Department of Engineering and Materials, the Department of Earth Sciences, the Department of Information Sciences, the Department of Management, and the Department of Medicine. Each data item of a project contains information such as the title, number, principal investigator, time, grant amount and institution. To present the experimental results more clearly, we select the data of the Department of Information Science for detailed analysis and experimentation.

B. DATASETS SEGMENTATION INFORMATION

This subsection conducts a segmentation statistical experiment on the titles of the projects in the Department of Information Science from 1999 to 2018. The statistics include four main fields: the total line, the average length, the dictionary length and line_maxlength. The total line refers to the number of projects that were funded in the corresponding year, which is mapped to the word vector space as the number of documents in the document collection. The average length is the average number of keywords in each document after the word segmentation, which can be defined as follows:

$$average\ length = \frac{\sum_{i=1}^{total\ line} number\ of\ keywords}{total\ line} \tag{22}$$

The dictionary length refers to the total number of dictionaries. Line_maxlength is the maximum length of the keywords in the projects, which corresponds to the maximum length of each document.

The experimental results are presented in Table 1.

This table presents the relevant information after the word segmentation and keyword statistics from the Department of Information Science's projects from 2000 to 2018. According to Table 1, both the totalline and the dictionary length increase gradually, which is related to the number of projects that are established each year. The average length also exhibits a gradual upward trend; hence, the applicants are more comprehensive and precise in the process of writing and the keywords in the project are being used with increasing frequency. At the same time, the results also demonstrate that scholars are becoming increasingly in-depth and proficient in their understanding and mastery of their research fields.

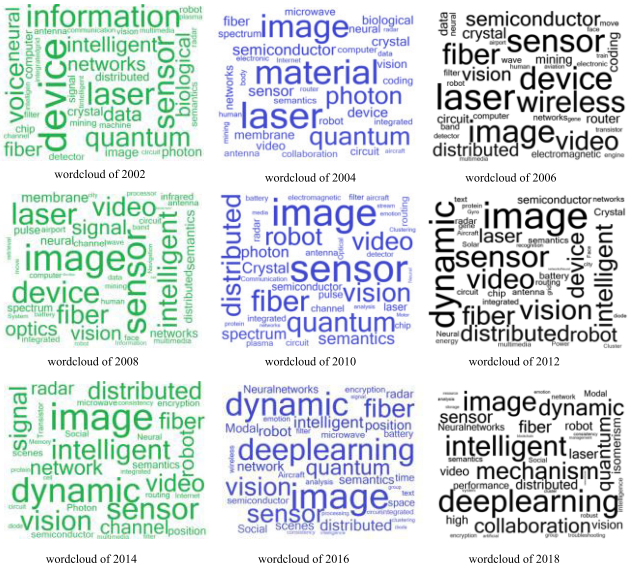

FIGURE 5. Word clouds that display changes in hot words.

C. HOT WORDS DISPLAY BASED ON WORD CLOUD

We can observe the regular evolution of research hotspots from the word cloud. This section presents hot words from 1999 to 2018 in the form of word clouds. The experimental process can be summarized into three steps: the first step is word segmentation; the second step is the calculation of statistics for the nouns, the removal of words with frequencies of less than 3, and the removal of some nouns with no special meaning; and the third step is the generation of the word cloud based on the statistical results, in which the larger the font, the more times the corresponding word appears. To reduce the size of the experimental display, this article shows it for every two years.

According to FIGURE 5, the evolution of scientific research hotspots in the past 20 years can be clearly identified. The research hotspots focus on hardware devices prior to 2010, e.g., sensors, lasers, fiber optics, semiconductors, crystals, circuits, and transistors. However, after 2010, with the development of technology, neural networks, social networks, sentiment analysis, big data, Internet of Things and other technologies have become research hotspots.

According to the figure, new words emerge every year. For example, keywords such as protein, face recognition, and gyro appeared in 2012; consistency and social networks emerged in 2014; neural network, group intelligence and deep learning emerged in 2016; and blockchain, sentiment analysis, encryption and social networking emerged in 2018. This phenomenon indicates that researchers have been constantly

TABLE 2. Hot topics and related terms for 2006 of the NSFC.

Topic#	Terms	Label
1	device, organic, passive, field effect, energy, multiplication frequency, plasma, technology, flexibility, metal, surface, variation, material	Device
2	space-time, vector, link, multiuser, system, figure, resonant cavity, ontology, fiber, network	Signal processing
3	antenna, cellular network, medium, plasma, popular, polarization, reconfiguration, particle, wide band	Mobile communication
4	information processing, multimedia, link, control algorithm, coordination control, network, sensor, strategy, location, business	Wireless sensing
5	password, function, system, microstructure, network, agreement, theory, algorithm, feature, information	Cryptography
6	image, direction, reconfiguration, human face, scene, probability, wave filter, manifold, vector, time domain, semantics	Graphic image
7	cell, information processing, artificial, wavelet transform, stability, visualization, model, agreement	Neural network
8	sequence, gene, informatics, visualization, reconfiguration, human face, probability, image, function, automata	Bioinformatics
9	route, strategy, satellite, wide band, agreement, architecture, network, function, wireless, environment, characteristic	Network technique
10	data base, control algorithm, coordination control, program, frequency, strategy, intellectualization, noise ,linear, stability	Distributed

exploring new research content and constantly innovating, which is closely related to social development, technological advancement, market demands and people's interests.

D. NMF-BASED TOPIC DISPLAY AND PERFORMANCE EVALUATION

In this subsection, we will conduct experiments on all datasets of NSFC1999-2018 using NMF. Furthermore, we set the values of K to 20, 40, 60, 80, and 100 and after running the program, we can identify multiple topics that reflect the convergence of the winning projects accurately. An important feature of topic processing is that these words do not necessarily appear in titles; instead they are combinations of problems that link topics together via similarity calculations. This subsection presents the main topic distributions and the corresponding keyword items from 2006 to 2018, which are displayed every four years, and selects 10 representative topics to facilitate understanding of the evolution of hot words.

In 2006, most topics are related to hardware communication, such as signal processing, mobile communication, sensor networks, and distributed processing, which is related to the research background of information science at that time. This period corresponded to the early stage of development of the third generation of mobile communications. Therefore, China's basic research on mobile communications has increased substantially, thereby effectively promoting the marketization of technology. The experimental results indirectly demonstrate that the state's scientific research input and policy support play important roles in promoting the transformation of science and technology into industrialization.

TABLE 3. Hot topics and related terms for 2010 of the NSFC.

Topic#	Terms	Label
1	data, process, dispatch, high-dimension, magnanimity, cluster, pattern, uncertainty, structured, agreement, industry, space	Industrial process
2	algorithm, problem, target, key, mechanism, quantum, base, video, dispatch, high-performance, plan, wave filtering	Algorithm research
3	network, wireless, code, relay, route, isomerism ,distributed, coordination, dynamics, agreement, dispatch	Network technique
4	structure, quantum, material, semiconductor, device, metal, surface, optics, oxide, plasma, sequence	Hardware
5	environment, signal, mobile robot, coordination, location, formation, networking, vision, high-mobility, scene	Robot
6	image, feature, vision, goal, video, sequence, semantic, quality, space, classification, spectrum, content	Graphic image
7	sensor, wireless, mechanism, energy, route, network security, agreement, multiuser, coordination, spectrum, node	Wireless sensor network
8	technology, code, software, spectrum, antenna, signal, goal, distributed, wavelength, video, reconfiguration, semantic	Mobile communication
9	information, vision, location, protein, nerve, human body, high-dimensional, pattern, quality, modal	Cross domain
10	intelligence, multimedia, electronic, information science, computer, signal processing, figure, optical communication, safety	Multimedia

TABLE 4. Hot topics and related terms for 2014 of the NSFC.

Topic#	Terms	Label
1	social contact, media, theme, community, individuation, journalism, relation, granularity, network, map, user, agreement	Social network
2	distributed, location, energy, optical fiber, multimedia, channel, gas, spectrum, resources, networking, agreement	Wireless sensor network
3	video, code, intelligence, feature, high-performance, robot, coordination, processor, channel, networking, granularity	Robot
4	distributed, uniformity, time-lag, networking, wave filtering, coordination, performance, robust, robot, controller	Intelligent system
5	image, feature, spectrum, vision, classification, semantic, resolution power, space, depth, modality, feature extraction, stereoscopic	Graphic image
6	Intelligence, group, uniformity, coordination control, route, colonycity, decision, cluster, stability	Swarm intelligence
7	event, modality, biology, networking, society, social contact, media, public sentiment, emotion, news, trend, hot spot, theme	Public opinion analysis
8	data, magnanimity, high-dimension, high throughput, query processing, quality, society, dimensionality reduction, feature selection, visualization	Big data
9	model, probability, Bayes, society, group, Markov, life, visualization, outline, theme	Stochastic theory
10	depth, pedestrian, machine, sign, multi precision, vehicle, viewpoint, multimedia, news, quality	Vehicular network

According to the hot topics in 2010, the research hotspots are scattered; however, they promote one another and cross-domain research topics appeared. For example, the third topic is network technology, and the fifth, sixth, and eighth topics are robots, graphic images, and mobile communications, respectively. We conclude that the rapid development of computer network technology will promote the transformation of technologies such as robotics, graphic images and mobile

TABLE 5. Hot topics and related terms for 2018 of the NSFC.

Topic #	Terms	Label
1	depth, prior, multitask, face recognition, cluster, expression, resolving power, individuation, multisource, neural network	Machine learning
2	distributed, networking, uniformity, block chain, storage system, safety control, network structure, resource management, energy	Blockchain
3	social contact, user, media, emotion, hot spot, cross-domain, interest, trust, text, individuation, resource allocation, resource	Social networks
4	decision, multitask, granularity, task scheduling, group intelligence, mobile robot, uncertainty, humanoid, hierarchical, sort	Robot
5	power, radio frequency, figure, linear, high efficiency, microwave, enhancement type, power supply, broadband, reliability, low-power consumption	Mobile communicat ion
6	semantic, visual angle, frame, deep level, probability, cluster, vector, wisdom, feature analysis, index	Semantic computation
7	environment, traffic, noise, industry, electromagnetism, parameter, deploy, channel, voice, route, carrier, atlas	Intelligent transportatio n
8	image, resolving power, medical science, recover, visual angle, feature extraction, text, saliency, region, pixel, color	Image processing
9	software, framework, intellectualization, data center, radio frequency, network system, control technology, integrated circuit, flow	New network
10	video, panorama, pixel, city, saliency, reasoning method, screen, streaming media, image enhancement, deformation, measure	Virtual reality

communication due to the background of the development of science and technology at that time. For the ninth topic, words such as protein, nerves, and human body emerged. These medical terms appear in the field of information; hence, they are no longer a specialized terms in the medical field. Analyzing and solving problems in the medical field via computer methods such as machine learning is a trend.

According to the hot topics in 2014, traditional research hotspots continue to be studied, such as wireless sensor networks and graphic images. At the same time, the research on new media topics has achieved blowout status, such as social networking, group intelligence, public opinion analysis, big data, vehicle networks and other research hotspots. These research hotspots are not only extensions of previous research hotspots; new vocabulary and new knowledge have also emerged, which are the key technologies for social development and applications.

The latest year for the NSFC is 2018. According to Table 7, research hotspots in new areas emerge in this year, such as machine learning, blockchain, intelligent transportation, and new networks. Once again, scientific research has followed the requirements and pace of the times. Researchers have been constantly innovating and making breakthroughs. The new research hotspots will promote research and development in the next few years, which will give China's scientific research a brighter future.

According to the above topic changes, various research hotspots are still being studied; however, the research keywords have changed substantially. For example, the previous vocabulary that was related to network technology topics included mainly routing, strategy, satellite, and broadband. By 2018, the network-related vocabulary had evolved into terms such as intelligence, data center, and control technology. Hence, the research in the field is more detailed and promotes the development of the field. The research hotspots have been constantly changing through research on topics over the years. China's scientific research has consistently yielded innovations and breakthroughs; this is closely related to social development, scientific and technological progress, the market and people's needs.

To evaluate the performance of our method, we conduct experiments and compare the results on our datasets with those of three methods (PCA, SVD, LDA) in terms of topic coherence and PMI score. The experimental results are presented in Table 6 and Table 7.

According to Table 6, NMF outperforms PCA, SVD and LDA. We compared the topic coherence values of the top-n terms for the topic under K = 40, 60, 80 and 100. For K = 40 and n = 5, LDA yields a satisfactory result. In all other cases, NMF outperforms the other three methods. As K increases, the topic coherence value of NMF increases.

According to Table 7, NMF outperforms PCA, SVD and LDA in terms of PMI score for all values of K and n.

E. HOTSPOTS EVULUTION ANALYSIS BASED ON SEMANTIC COMPUTING

According to the project establishment in previous years, we selected 20 hot words as the query keywords. Via the *lucene* scoring process, the annual score of each word is obtained and normalized according to the year. The results are presented in Table 8.

The evolution of each hotspot is observed in Table 8. Various research fields have always been the focus of researchers, such as neural networks and sensors, in which the proportion of each year is balanced according to the normalized score table. Other areas are emerging hotspots, such as swarm intelligence, sentiment analysis, deep learning and blockchain.

To represent the evolutionary trends the research hotspots, we calculate the attention for each topic. The specific formula is as follows.

$$Attention(year, k) = \frac{\sum_{n=1}^{k} Ntopic_i^k}{N^{(year)}} \tag{23}$$

where k represents the top-k words in a topic, i denotes the ith topic, $N^{(year)}$ denotes the total number of the hot words in the specified year, and $Ntopic_i^k$ represents the number of top-k words in $topic_i$.

Then, we obtained the hot topic evolution based on the experimental results, which is plotted in FIGURE 6.

In FIGURE 6, the abscissa corresponds to the year and the ordinate to the degree of attention. According to FIGURE 6, the topic of wireless sensor networks has been a research hotspot since 2002 and it has increased in popularity over the years. Topic robot has been a research hotspot since 1999 and continues to increase in popularity. Topic integrated circuit has been a research hotspot since 1999; however, afterwards,

TABLE 6. Top-n term topic coherence comparison with various values of K and n on four baselines.

Baseline	K=40				K=60				K=80				K=100			
	Top5	Top10	Top15	Top20	Top5	Top10	Top15	Top20	Top5	Top10	Top15	Top20	Top5	Top10	Top15	Top20
PCA	-33.22	-156.1	-375.8	-696.2	-34.00	-158.7	-378.1	-695.9	-34.33	-157.7	-376.2	-689.0	-34.06	-157.0.	371.6	-679.4
SVD	-32.84	-154.7	-367.2	-685.1	-34.21	-157.9	-373.1	-688.0	-34.24	-158.6	-371.8	-682.5	-33.78	--155.7	-367.8	-673.7
LDA	**-30.68**	-151.6	-363.7	-658.8	-31.65	-156.8	-381.7	-689.0	-34.48	-165.9	-387.0	-699.0	-34.68	-167.3	-395.3	-709.1
NMF	-32.06	**-145.0**	**-337.0**	**-601.3**	**-29.53**	**-134.2**	**-311.6**	**-554.6**	**-28.60**	**-127.9**	**-292.6**	**-519.9**	**-27.53**	**-121.9**	**-279.2**	**-496.0**

TABLE 7. Top-n term PMI value comparison with various values of K and n on four baselines.

Baseline	K=40				K=60				K=80				K=100			
	Top5	Top10	Top15	Top20	Top5	Top10	Top15	Top20	Top5	Top10	Top15	Top20	Top5	Top10	Top15	Top20
PCA	0.046	0.097	0.205	0.304	0.251	0.285	0.349	0.435	0.461	0.478	0.537	0.609	0.675	0.673	0.726	0.788
SVD	0.159	0.168	0.270	0.382	0.300	0.317	0.389	0.496	0.494	0.512	0.571	0.648	0.729	0.720	0.752	0.815
LDA	0.852	1.065	1.304	1.523	0.581	0.795	1.061	1.355	0.376	0.664	0.930	1.203	0.175	0.533	0.807	1.091
NMF	**1.588**	**1.032**	**2.034**	**2.268**	**2.256**	**2.403**	**2.617**	**2.733**	**2.607**	**2.856**	**2.983**	**3.128**	**2.924**	**3.083**	**3.262**	**3.373**

TABLE 8. Semantic-based specified keyword query normalization scores.

Key word	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
sensor	0.0065	0.0058	0.0073	0.0138	0.0189	0.0196	0.029	0.0297	0.0377	0.0486	0.0602	0.0703	0.0863	0.0819	0.0827	0.0819	0.0725	0.0754	0.0841	0.0877
intelligent	0.0106	0.0099	0.0185	0.0177	0.0256	0.0234	0.027	0.027	0.0192	0.0284	0.049	0.0426	0.061	0.0582	0.0674	0.0667	0.0873	0.105	0.0965	0.159
image	0.0038	0.0	0.0022	0.0059	0.0124	0.0129	0.0178	0.021	0.0248	0.042	0.0479	0.0511	0.064	0.0942	0.0764	0.0947	0.1039	0.1055	0.1103	0.1093
neuralnetworks	0.0177	0.0227	0.0152	0.0328	0.0126	0.0278	0.0202	0.0278	0.0227	0.0455	0.0278	0.0177	0.0404	0.053	0.0328	0.0682	0.1086	0.1086	0.1035	0.1944
robot	0.0096	0.0055	0.011	0.0096	0.0193	0.0193	0.0344	0.0234	0.0248	0.0372	0.0399	0.0496	0.0744	0.0771	0.0661	0.0785	0.0978	0.0964	0.1129	0.1129
integratedcircuit	0.0311	0.0178	0.0	0.0178	0.0311	0.04	0.0533	0.0178	0.04	0.0444	0.04	0.0667	0.0578	0.0978	0.0578	0.0844	0.08	0.08	0.0711	0.0711
microwave	0.0156	0.013	0.0286	0.0156	0.0364	0.0364	0.0286	0.0286	0.026	0.0442	0.039	0.0494	0.0364	0.0468	0.0779	0.0701	0.0909	0.0831	0.1169	0.1169
distributed	0.006	0.0048	0.0084	0.012	0.0108	0.0157	0.0301	0.0277	0.0217	0.0217	0.0482	0.0494	0.0554	0.088	0.0675	0.1024	0.0976	0.0988	0.094	0.1398
multimedia	0.0267	0.0333	0.0267	0.0267	0.0267	0.0	0.0267	0.0267	0.0267	0.0533	0.0667	0.0733	0.1067	0.1	0.0533	0.12	0.0533	0.0467	0.0533	0.0533
vision	0.0053	0.0063	0.0127	0.0053	0.0074	0.0179	0.0158	0.0253	0.02	0.0327	0.0485	0.0432	0.0675	0.1118	0.0759	0.0939	0.1002	0.1023	0.1129	0.0949
semantics	0.0086	0.0	0.0138	0.0121	0.0138	0.0173	0.0207	0.038	0.0328	0.0397	0.0397	0.0518	0.076	0.0881	0.0984	0.0812	0.0829	0.0864	0.1123	0.0864
encryption	0.0	0.0	0.012	0.0	0.033	0.033	0.024	0.042	0.021	0.042	0.0721	0.0511	0.0811	0.1051	0.0511	0.0991	0.0811	0.0841	0.0871	0.0811
protein	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0391	0.0	0.0391	0.0781	0.0859	0.1016	0.1094	0.1328	0.1016	0.0938	0.125	0.0938
cell	0.0	0.0	0.0	0.0	0.0192	0.0	0.0481	0.0288	0.0288	0.0	0.0337	0.0529	0.0913	0.0625	0.0913	0.0769	0.0913	0.0913	0.0865	0.1971
clustering	0.0	0.0	0.0	0.0	0.0248	0.0	0.0	0.0	0.0311	0.0435	0.0559	0.0497	0.1056	0.0807	0.0745	0.0994	0.1118	0.1056	0.0621	0.1553
swarm	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0376	0.0	0.0301	0.0376	0.0376	0.1053	0.0902	0.0977	0.1203	0.1128	0.1429	0.1053	0.0827
emotion	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0253	0.0443	0.038	0.0506	0.0443	0.1266	0.0633	0.0949	0.1266	0.1392	0.1456	0.1013
consensus	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.034	0.034	0.0408	0.0748	0.1088	0.1361	0.1293	0.1156	0.1156	0.0952	0.1156
blockchain	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1053	0.3684	0.5263
deeplearning	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.005	0.005	0.005	0.0138	0.0214	0.0314	0.0666	0.1583	0.1583	0.2111	0.3241

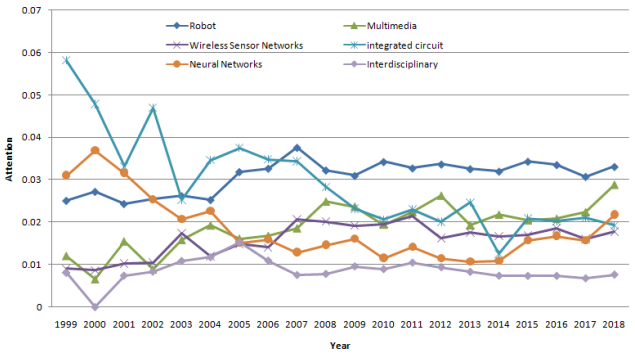

FIGURE 6. Evolutionary trends in several topics.

the attention gradually decreased. Topic of interdisciplinary first appeared in the Department of Information Science in 2000 and exhibited a growing trend in subsequent years; hence, interdisciplinary research is gradually becoming a research hotspot.

VIII. CONCLUSION AND FUTURE WORK

Based on an analysis of the numbers and the funding amounts of NSFC projects in the past 20 years, this paper explores and mines NSFC project titles via NMF and the semantic search-based keyword scoring method. These methods can predict research hotspots in various fields and discover the evolution of NSFC research hotspots. Furthermore, we conduct our experiments on 20 datasets and present the results for the NSFC Department of Information Sciences grants from various perspectives, from which we can see clearly identify the research hotspots and regular evolutionary patterns. The experimental process can be summarized into the following four steps: First, according to the segmentation statistics of the titles of projects, the scientific researchers' consideration of project titles is becoming increasingly comprehensive and meticulous. Second, we obtained a word cloud map that is based on the frequency of noun occurrences by segmenting the project titles and extracting key nouns as the research vocabulary. Third, we extract the topics and related keywords

of the research field using NMF, which exhibits superior performance and higher accuracy in predicting the evolution of hotspots. Finally, we can derive the distribution of hot words in the past years by analyzing the scoring method based on semantic search keywords, which can more accurately predict the evolution of research hotspots.

However, several limitations are encountered in this study: There are eight departments in the NSFC, and the number of projects is growing rapidly each year. There are increasingly more disciplines and integrations, which renders it difficult to process topic predictions. Our method may not be the optimal method; therefore, we will explore more precise methods in future research. In addition, this article only considered the Department of Information Science for prediction and did not cover all departments. In future research, we will explore other more effective and accurate methods and models for tracking and predicting the research hotspots of science and technology projects to better serve researchers and to promote the sustainable development of national science and technology innovation.

REFERENCES

[1] Z. Engin and P. Treleaven, "Algorithmic government: Automating public services and supporting civil servants in using data science technologies," *Comput. J.*, vol. 62, pp. 448–460, Mar. 2019.
[2] F. Teodoridis, "Understanding team knowledge production: The interrelated roles of technology and expertise," *Manage. Sci.*, vol. 64, no. 8, pp. 3625–3648, 2017.
[3] S. M. Flipse and S. Puylaert, "Organizing a collaborative development of technological design requirements using a constructive dialogue on value profiles: A case in automated vehicle development," *Sci. Eng. Ethics*, vol. 24, pp. 49–72, Feb. 2018.
[4] J. Krzyszt, R. Blake, H. Pascal, M. Gengchena, D. Stephaniec, F. Maartenc, M. Patrick, and L. Trevor, "On the prospects of blockchain and distributed ledger technologies for open science and academic publishing," *Semantic Web*, vol. 9, no. 5, pp. 545–555, 2018.
[5] L. Feng, H. Zhang, Y. Chen, and L. Lou, "Scalable dynamic multi-agent practical byzantine fault-tolerant consensus in permissioned blockchain," *Appl. Sci.*, vol. 8, no. 10, p. 1919, 2018.
[6] J. An, K. Kim, L. Mortara, and S. Lee, "Deriving technology intelligence from patents: Preposition-based semantic analysis," *J. Inform.*, vol. 12, pp. 217–236, Feb. 2018.
[7] Y. Hao, L. Song, M. Wang, L. Cui, and H. Wang, "Underdetermined source separation of bearing faults based on optimized intrinsic characteristic-scale decomposition and local non-negative matrix factorization," *IEEE Access*, vol. 7, pp. 11427–11435, 2019.
[8] R. Rad and M. Jamzad, "A multi-view-group non-negative matrix factorization approach for automatic image annotation," *Multimedia Tools Appl.*, vol. 77, pp. 17109–17129, Jul. 2018.
[9] A. S. Raamkumar, S. Foo, and N. Pang, "Using author-specified keywords in building an initial reading list of research papers in scientific paper retrieval and recommender systems," *Inf. Process. Manage.*, vol. 53, no. 3, pp. 577–594, 2017.
[10] H. Kamper, G. Shakhnarovich, and K. Livescu, "Semantic speech retrieval with a visually grounded model of untranscribed speech," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 1, pp. 89–98, Jan. 2019.
[11] Q. Zhao, Z. Li, Z. Zhao, and J. Ma, "Industrial policy and innovation capability of strategic emerging industries: Empirical evidence from chinese new energy vehicle industry," *Sustainability*, vol. 11, no. 10, p. 2785, 2019.
[12] A. J. C. Trappey, P. P. J. Chen, C. V. Trappey, and L. Ma, "A machine learning approach for solar power technology review and patent evolution analysis," *Appl. Sci.*, vol. 9, no. 7, p. 1478, 2019.
[13] H. Lu and H. You, "Roadmap modeling and assessment approach for defense technology system of systems," *Appl. Sci.*, vol. 8, no. 6, p. 908, 2018.
[14] A. Patelli, G. Cimini, E. Pugliese, and A. Gabrielli, "The scientific influence of nations on global scientific and technological development," *J. Inform.*, vol. 11, pp. 1229–1237, Nov. 2017.
[15] J. Kim, J. Lee, G. Kim, S. Park, and D. Jang, "A hybrid method of analyzing patents for sustainable technology management in humanoid robot industry," *Sustainability*, vol. 8, no. 5, p. 474, 2016.
[16] W. S. Lee and S. Y. Sohn, "Identifying emerging trends of financial business method patents," *Sustainability*, vol. 9, no. 9, p. 1670, 2017.
[17] X. Bai, M. Wang, I. Lee, Z. Yang, X. Kong, and F. Xia, "Scientific paper recommendation: A survey," *IEEE Access*, vol. 7, pp. 9324–9339, 2019.
[18] K. Norambuena, B. Lettura, E. F. Villegas, and C. Meneses, "Sentiment analysis and opinion mining applied to scientific paper reviews," *Intell. Data Anal.*, vol. 23, no. 1, pp. 191–214, 2019.
[19] W. Waheed, M. Imran, B. Raza, A. K. Malik, and H. A. Khattak, "A hybrid approach toward research paper recommendation using centrality measures and author ranking," *IEEE Access*, vol. 7, pp. 33145–33158, 2019.
[20] J. Lee, J.-H. Kang, S. Jun, H. Lim, D. Jang, and S. Park, "Ensemble modeling for sustainable technology transfer," *Sustainability*, vol. 10, no. 7, p. 2248, 2018.
[21] L. Yuan, Y. Hao, M. Li, C. Bao, J. Li, and D. Wu, "Who are the international research collaboration partners for China? A novel data perspective based on NSFC grants," *Scientometrics*, vol. 116, pp. 401–422, Jul. 2018.
[22] L. G. Nichols, "A topic model approach to measuring interdisciplinarity at the national science foundation," *Scientometrics*, vol. 100, pp. 741–754, Sep. 2014.
[23] X. Chen, J. Li, X. Sun, and D. Wu, "Early identification of intellectual structure based on co-word analysis from research grants," in *Scientometrics*. Dordrecht, The Netherlands: Springer, 2019. doi: 10.1007/s11192-019-03187-9.
[24] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, Oct. 1999.
[25] Y. Chen, H. Zhang, R. Liu, and Z. Ye, "Soft orthogonal non-negative matrix factorization with sparse representation: Static and dynamic," *Neurocomputing*, vol. 310, pp. 148–164, Oct. 2018.
[26] F. Zhuang, X. Li, X. Jin, D. Zhang, L. Qiu, and Q. He, "Semantic feature learning for heterogeneous multitask classification via non-negative matrix factorization," *IEEE Trans. Cybern.*, vol. 48, no. 8, pp. 2284–2293, Aug. 2018.
[27] M. H. Aghdam, M. AnaLoui, and P. Kabiri, "Collaborative filtering using non-negative matrix factorisation," *J. Inf. Sci.*, vol. 43, pp. 567–579, 2017.
[28] J. Woo, J. L. Prince, M. Stone, F. Xing, A. D. Gomez, J. R. Green, C. J. Hartnick, T. J. Brady, T. G. Reese, Van J. Wedeen, and G. El Fakhri, "A sparse non-negative matrix factorization framework for identifying functional units of tongue behavior from MRI," *IEEE Trans. Med. Imag.*, vol. 38, no. 3, pp. 730–740, Mar. 2019.
[29] Q. Xiao, J. Luo, C. Liang, J. Cai, and P. Ding, "A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations," *Bioinformatics*, vol. 34, no. 2, pp. 239–248, 2017.
[30] Y. Chen, H. Zhang, X. Zhang, and R. Liu, "Regularized semi-non-negative matrix factorization for hashing," *IEEE Trans. Multimedia*, vol. 20, no. 7, pp. 1823–1836, Jul. 2018.
[31] D. Tolić, N. Antulov-Fantulin, and I. Kopriva, "A nonlinear orthogonal non-negative matrix factorization approach to subspace clustering," *Pattern Recognit.*, vol. 82, pp. 40–55, Oct. 2018.
[32] J. Bobadilla, R. Bojorque, A. H. Esteban, and R. Hurtado, "Recommender systems clustering using Bayesian non negative matrix factorization," *IEEE Access*, vol. 6, pp. 3549–3564, 2018.
[33] N. Dia, J. Fontecave-Jallon, P.-Y. Gumery, and B. Rivet, "Denoising phonocardiogram signals with non-negative matrix factorization informed by synchronous electrocardiogram," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Rome, Italy, Sep. 2018, pp. 51–55.
[34] R. Stoean and M. A. Atencia-Ruiz, "Non-negative matrix factorization for medical imaging," in *Proc. 26th Eur. Symp. Artif. Neural Netw. (ESANN)*, Bruges, Belgium, 2018, pp. 25–27.
[35] R. Pokharel, P. D. Haghighi, P. P. Jayaraman, and D. Georgakopoulos, "Analysing emerging topics across multiple social media platforms," in *Proc. Australas. Comput. Sci. Week Multi Conf.*, Sydney, NSW, Australia, 2019, p. 16.
[36] X. Liu, W. Wang, D. He, P. Jiao, D. Jin, and C. V. Cannistraci, "Semi-supervised community detection based on non-negative matrix factorization with node popularity," *Inf. Sci.*, vol. 381, pp. 304–321, Mar. 2017.
[37] H. Liang, J. Zou, Z. Li, M. J. Khan, and Y. Lu, "Dynamic evaluation of drilling leakage risk based on fuzzy theory and PSO-SVR algorithm," *Future Gener. Comput. Syst.*, vol. 95, pp. 454–466, Jun. 2019.

[38] C. Li, J. Zhang, T. Ma, H. Tang, L. Zhang, and Y. Luo, "Data locality optimization based on data migration and hotspots prediction in geo-distributed cloud environment," *Knowl.-Based Syst.*, vol. 165, pp. 321–334, Feb. 2019.
[39] J. Bencatel, C. C. Ferreira, A. M. Barbosa, L. M. Rosalino, and F. Álvares, "Research trends and geographical distribution of mammalian carnivores in Portugal (SW Europe)," *PLoS ONE*, vol. 13, p. 11, Nov. 2018.
[40] Y. Xiao, X. Li, S. Yang, and Y. Liu, "Who will retweet? A prediction method for social hotspots based on dynamic tensor decomposition," *Sci. China Inf. Sci.*, vol. 61, p. 9, Sep. 2018.
[41] D. Yang, T. Heaney, A. Tonon, L. Wang, and P. Cudré-Mauroux, "CrimeTelescope: Crime hotspot prediction based on urban and social media data fusion," *World Wide Web*, vol. 21, pp. 1323–1347, Sep. 2018.
[42] J.-F. Xia, X.-M. Zhao, J. Song, and D.-S. Huang, "APIS: Accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility," *Bioinformatics*, vol. 11, p. 174, Apr. 2010.
[43] M. Adepeju, G. Rosser, and T. Cheng, "Novel evaluation metrics for sparse spatio-temporal point process hotspot predictions—A crime case study," *Int. J. Geograph. Inf. Sci.*, vol. 30, no. 11, pp. 2133–2154, 2016.
[44] T. A. Abeo, X.-J. Shen, E. D. Ganaa, Q. Zhu, B.-K. Bao, and Z.-J. Zha, "Manifold alignment via global and local structures preserving PCA framework," *IEEE Access*, vol. 7, pp. 38123–38134, 2019.
[45] Y. Yang and J. Rao, "Robust and efficient harmonics denoising in large dataset based on random SVD and soft thresholding," *IEEE Access*, vol. 7, pp. 77607–77617, 2019.
[46] F. Gurcan and N. E. Cagiltay, "Big data software engineering: Analysis of knowledge domains and skill sets using LDA-based topic modeling," *IEEE Access*, vol. 7, pp. 82541–82552, 2019.
[47] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2011, pp. 262–272.
[48] D. Newman, S. Karimi, and L. Cavedon, "External evaluation of topic models," in *Proc. ADCS*, Sydney, NSW, Australia, 2009, pp. 11–18.

JINLI WANG received the M.S. degree from Dali University, in July 2018. She is currently pursuing the Ph.D. degree with the School of Management and Economics, Kunming University of Science and Technology. Her research interests include science and technology policy and management, machine learning, science and technology information management.

YONG FAN received the Ph.D. degree from Sun Yat-sen University, Guangzhou. He was a Postdoctoral Fellow with the Renmin University of China. He is currently a Full Professor with the Faculty of Marxism, Kunming University of Science and Technology, Kunming, China. His research interests include the philosophy of development, science and technology policy and management, science and education management and knowledge innovation, and engineering ethics research.

LIBO FENG received the M.S. degree from the Beijing University of Posts and Telecommunication, in 2008. He is currently pursuing the Ph.D. degree with the School of Computer Science, Beihang University. His research interests include big data analysis and processing, distributed systems, machine learning, and blockchain.

ZHIWEN YE received the bachelor's degree from Zhengzhou University, in 2017. He is currently pursuing the master's degree in computer science and technology with the State Key Laboratory of Software Development Environment, Beihang University. His current research interests include machine learning, data mining, search engine, and web information retrieval.

HUI ZHANG received the Ph.D. degree from Beihang University, in 2010. He was a Visiting Scholar with the Argonne National Laboratory, USA, in 2007. He is currently a Professor with the School of Computer Science, Beihang University, China. His current research interests include big data analysis and processing, distributed systems, machine learning, and blockchain.