Research on Core Technology Topic Identification Method Based on Chunk-LDAvis (Postprint)

Authors: Liu Ziqiang, Xu Haiyun, Yue Lixin, Fang Shu

Date: 2023-07-26T00:00:00+00:00

Abstract

[Purpose/Significance] Core technology theme identification based on extensive patent literature data facilitates the recognition of key technologies within a technical domain and the analysis of their development trajectories, constituting foundational intelligence work for technological innovation that holds significance for researchers, enterprises, and national-level stakeholders. [Method/Process] This study proposes a core technology theme identification methodology based on Chunk-LDAvis. Initially, topic identification is conducted using the classical LDA model. Subsequently, noun chunks are employed to annotate the initial LDA topic identification results, thereby constructing Chunk-LDA topic identification outputs and enhancing their interpretability. A topic network is then constructed through social network analysis methods to identify core technology themes. Finally, the LDAvis package in R is utilized to generate an interactive Chunk-LDAvis core technology theme correlation analysis visualization, revealing implicit connections among core technology themes to assist in their identification. [Results/Conclusion] An empirical study in the domain of nano-agriculture validates the accuracy and feasibility of the proposed methodology.

Full Text

Preamble

Vol. 63 No. 9 May 2019

Research on Core Technology Topic Identification Method Based on Chunk-LDAvis

Liu Ziqiang[1,2], Xu Haiyun[1,3], Yue Lixin[4], Fang Shu[1]

[1]Chengdu Library of Chinese Academy of Sciences, Chengdu 610041
[2]Department of Library, Information and Archives Management, School of Eco-

nomics and Management, University of Chinese Academy of Sciences, Beijing 100190
[3]Institute of Scientific and Technical Information of China (ISTIC), Beijing 100038
[4]School of Information Resource Management, Renmin University of China, Beijing 100872

Abstract

[**Purpose/Significance**] Core technology topic identification based on large-scale patent literature data helps identify key technologies in a technical field and analyze their development directions. This fundamental intelligence work for technological innovation holds significance for researchers, enterprises, and national-level decision-making. [**Method/Process**] This paper proposes a core technology topic identification method based on Chunk-LDAvis. First, topics are identified using the classic LDA model. Then, noun chunks are used to annotate the initial LDA topic identification results, constructing Chunk-LDA topic identification results to improve interpretability. Next, a topic network is constructed based on social network analysis methods to identify core technology topics. Finally, the LDAvis toolkit in R is used to draw interactive Chunk-LDAvis core technology topic association analysis maps, revealing hidden connections among core technology topics to assist in identification. [**Result/Conclusion**] An empirical study in the nano-agriculture field verifies the accuracy and feasibility of the proposed method.

Classification Number: G251.2
Keywords: Chunk-LDAvis; Patent Analysis; Topic Recognition; Core Technology Topics; Interactive Visualization

Introduction

With the rapid evolution of a new round of global technological revolution and industrial transformation, technological innovations continue to emerge, promoting the generation of new products, demands, and business formats. These innovations have become key drivers for sustained socio-economic development, influencing adjustments to economic patterns and industrial structures, and representing critical factors for development and national competitiveness. Currently, countries worldwide attach great importance to scientific and technological innovation, increasing investment in high-tech fields to seize opportunities in this new round of scientific and technological revolution. In recent years, China has consistently emphasized an innovation-driven development strategy, recognizing that technological innovation is the strategic support for improving social productivity and comprehensive national power, and must be placed at the core of national development.

In the era of big data, the volume of scientific and technological literature such as patents and papers is growing exponentially, making global, forward-looking,

and strategic scientific and technological intelligence services particularly important for supporting scientific and technological decision-making and innovation. The World Intellectual Property Organization notes that over 90% of scientific and technological information is reflected through patent information, making patent literature an important and reliable data source for analyzing technology development trends. Although global patent output is growing rapidly, scholars have found through evaluation of patent values in European countries after the 1950s that the value distribution of patent literature is uneven, with approximately 5%-10% of patents accounting for half of the total value. How to accurately and effectively capture core technologies from massive patent literature and predict their development trends has become an urgent problem in current scientific and technological intelligence work.

Therefore, numerous scholars have conducted technology identification and prediction research based on patent literature data, achieving many results. For example, core technology identification and prediction analysis methods based on patent citation analysis, patent keyword analysis, and visual analysis have provided certain assistance for scientific and technological innovation in various countries. However, as intelligence needs continue to deepen, corresponding core technology identification and prediction methods need further improvement. Building on current research, this paper proposes a core technology topic identification method based on Chunk-LDAvis to address deepening scientific and technological innovation intelligence needs, providing references for scientific and technological intelligence work at different levels including researchers, enterprises, and nations.

1 Related Research

1.1 Core Technology Topic Identification Based on Citation Features

With the rapid evolution of global technological revolution and industrial transformation, research on core technology identification based on scientific and technological literature has received significant attention from scholars, enterprises, and governments worldwide. Scholars have conducted extensive research on how to efficiently and accurately identify core technologies, hot technologies, and their development trends using scientific and technological literature data. These efforts can be broadly divided into two directions: analyzing citation features such as co-citation, bibliographic coupling, and direct citation of patent literature; and analyzing text content features such as titles and abstracts.

Among them, methods based on patent citation features for core technology topic identification have attracted early scholarly attention. For example, O. Kwon et al. identified core technologies by constructing patent citation coupling networks and co-citation networks and comprehensively analyzing patent distribution, verifying the method' s effectiveness through empirical studies in three fields. C. Choi et al. proposed a core technology identification method based on main path analysis algorithms: first constructing a patent citation network, then

using main path discovery algorithms to extract the main path of patent technology development, and finally analyzing technology evolution 脉络 to identify key technologies and their development trends. C. W. Hsu et al. used patent clustering methods to establish mutual citation matrices between related technologies in the biohydrogen production field, mapping technology development and identifying representative technology fields. Zhang Xin et al. improved the original PageRank algorithm by combining citation counts and patent age, applying it to the OLED field to identify core patents. Kang Chuanbo et al. constructed citation networks of patent literature based on mutual citation relationships, then identified core patent topics based on individual value and network value indicators.

Core technology identification methods based on citation features can effectively identify core technologies. However, due to citation time lag (the time required for a document to be cited after publication, and for citing documents to be published), many scholars question the timeliness and accuracy of citation-based core technology identification. They attempt to delve deeper into patent literature content, conducting co-occurrence and clustering analysis based on text content features (patent titles, abstracts, etc.) to identify more interpretable and accurate core technology topics.

1.2 Core Technology Topic Identification Based on Content Features

With the development of natural language processing technologies (text clustering, LDA topic models, and community detection, etc.), methods for identifying core technology topics based on patent title, abstract, and other content features have gradually gained scholarly attention. For example, Y. G. Lee et al. proposed a "technology cluster analysis" method for selecting core strategic research areas and applied it to national R&D projects in the nanotechnology field. The specific approach involved keyword extraction, patent document clustering, and analyzing the hierarchical distribution relationship of keywords in patent document clusters to identify core technologies, using this method to predict three major core technology clusters in South Korea' s nanotechnology field. Luan Chunjuan et al. used the Derwent Innovation Index as a data source, extracted "Derwent Manual Codes" (DMC), and drew co-occurrence networks for visual analysis to identify core technology fields, conducting an empirical study in the aerospace field. Fan Yu et al. proposed a topic model and clustering algorithm suitable for patent information clustering, combining the Latent Dirichlet Allocation (LDA) topic model with the OPTICS algorithm for core patent topic analysis. Li Jiajia et al. used social network analysis methods to compare and analyze patent classification code co-occurrence networks of different countries including China, the United States, and Europe, identifying core patent fields. Yi Huifang et al. combined the LDA model with strategic coordinate mapping for patent technology topic analysis, identifying core technology topics and their structural characteristics, which is of great significance for objectively and reasonably tracking technology frontiers and improving R&D efficiency.

Although co-occurrence and clustering analysis based on patent literature text content (keywords, classification codes, etc.) has certain advantages over citation-based methods (no citation time lag), it also has limitations. For instance, keywords lack semantic relationships, cannot reflect associations between words, and cannot effectively reveal relationships between technology topics.

1.3 Improvement and Application of LDA Model

The LDA model was first proposed by D. M. Blei et al. in 2003 and can express semantic hierarchical relationships between words at the statistical probability level. In 2006, D. M. Blei et al. proposed the dynamic topic model, enabling LDA to process document datasets with timestamps for dynamic topic identification and tracking. However, the classic LDA model has certain limitations: for example, each topic in LDA identification results is a set of words that is not easy to interpret; after topic identification, the associations between topic-topic and topic-word are difficult to measure.

Addressing these two limitations, relevant scholars have conducted improvement research and achieved numerous results, such as the TNG (Topical N-Grams) model, PhraseLDA model, and LDAvis model. Among them, the TNG and PhraseLDA models use phrases to represent topics, offering better semantic expressiveness; the LDAvis model can map LDA topic identification results into two-dimensional space based on multidimensional scaling algorithms, thereby revealing associations between topic-topic and topic-word. In recent years, scholars in library and information science have used LDA models and their improved algorithms for scientific research topic identification, applying them to scientific and technological intelligence analysis based on text data. For example, Fan Yunman et al. used the TNG model for emerging topic detection research. Zhang Qin et al. used the PhraseLDA model for topic phrase mining method research, with results showing high-quality topic phrases mined from multiple datasets.

Based on the above analysis, this study draws on the TNG and PhraseLDA models, using noun chunks (Chunk) to represent topics (which contain higher semantic information content than phrases), and then uses the LDAvis model to reveal associations between topic-topic and topic-word, thereby constructing the Chunk-LDAvis model and applying it to core technology topic identification research. Using Chunk-LDAvis for core technology topic identification can represent each core technology topic as a set of noun chunks on one hand, improving interpretability, and on the other hand reveal interconnections between core technology topics and topic words.

2 Core Technology Topic Identification Based on Chunk-LDAvis

Through summarizing related research, core technology topic identification research has two interconnected improvement directions: (1) enhancing semantic information of technology topics to increase the information content of content features; (2) identifying content-dimensional associations between technology topics and using these relationships to identify core technology topics. The former is the foundation, using topic models, semantic analysis, and other methods to more effectively (compared to keywords and citation links) summarize and generalize content features of patent texts. The latter is the deepening, adding semantic-dimensional associations between core technology topics rather than simple co-occurrence associations, and identifying core technology topics based on these semantic-dimensional relationships.

Based on the above analysis, this paper proposes a core technology topic identification framework based on Chunk-LDAvis, mainly including four systematic processes: data collection and processing, semantically enhanced topic identification, core technology topic determination, and association visualization analysis. The main idea is shown in Figure 1 [Figure 1: see original paper]:

Step 1: Data Collection and Processing. Determine the database according to the target, construct search queries to obtain patent literature in the corresponding technical field. Then perform data processing, including patent literature format conversion. Since the research purpose is core patent topic analysis, key information such as titles, abstracts, and time needs to be extracted and saved locally for subsequent research.

Step 2: Semantically Enhanced LDA Topic Identification. First, topic identification is performed based on the classic LDA model. Then, using part-of-speech tagging, syntactic analysis, and grammatical analysis, subject noun chunks (Subject Noun Trunk, i.e., noun phrases representing the subject) and object noun chunks (Object Noun Trunk, i.e., noun phrases representing the object) are extracted from the supporting documents of each topic. These chunks are used to annotate the initial LDA topic identification results, constructing Chunk-LDA topic identification results to enhance the semantic function and improve interpretability.

Step 3: Core Technology Topic Determination. Time windows are divided, topic networks are constructed, and core patent topics are identified based on social network analysis methods.

Step 4: Core Technology Topic Visualization Analysis Based on Chunk-LDAvis. Using Web front-end technology, interactive Chunk-LDAvis core technology topic association analysis maps are drawn. Then a Web database is built for online testing, improving from two aspects: semantic enhancement and interpretability of core technology topic identification results, thereby effectively identifying and analyzing core technology topics.

The main steps are detailed below.

2.1 Semantically Enhanced LDA Topic Identification

(1) Initial LDA Topic Identification. In recent years, numerous topic models have been proposed, such as Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (pLSA), and LDA models. Compared with LSA and pLSA models, the LDA model can not only predict topic distributions of documents in the training set but also effectively predict topic distributions of documents and words not in the training set. Therefore, the LDA model has gradually become one of the most effective tools for analyzing large-scale unstructured document collections.

Specifically, LDA is a three-layer (word, topic, and document) Bayesian probability model (see Figure 2 [Figure 2: see original paper]). The LDA model assumes that documents are composed of several latent topics, and topics are composed of all words in the vocabulary. The joint distribution probability of the LDA topic model is shown in formula (1):

$$P(\theta, z, w) = P(\theta|w) \prod P(z_n|\theta) P(w_n|z_n, \beta) \quad \text{formula (1)}$$

Where M is the number of documents, K is the number of topics, N represents the number of words in the m-th document, is sampled from a Dirichlet distribution with parameter α, z represents topics, w represents topic words, and ϕ is a Dirichlet distribution with parameter β.

The LDA model generation process can be summarized as follows: 1) Sample topic-word distribution *ϕk for each topic from a Dirichlet distribution with parameter β, i.e.,* ϕk ~ Dir(β), k [1, K]. 2) Sample document-topic distribution _m for each document from a Dirichlet distribution with parameter α, i.e., _m ~ Dir(α), m [1, M]. 3) For the n-th word (n [1, N_m]) in document m: Sample a topic z_m,n from a multinomial distribution with parameter _m, i.e., z_m,n ~ Mult(_m). 4) Sample a specific word w_m,n from a multinomial distribution with parameter *ϕ{z_m},n, i.e., w_m,n ~ Mult(ϕ{z_m},n).*

The key to using LDA for modeling document data is to infer the hyperparameters α and β, i.e., to calculate the implicit parameters of each document-topic distribution _m and topic-word distribution *ϕ{z_m},n. Current parameter estimation methods for LDA models include Maximum A Posteriori (MAP), Variational Bayes (VB), Collapsed Variational Bayesian Inference (CVB), and Gibbs Sampling (GS). This study uses the LDA*{Gibbs} model from the topicmodels package in R to estimate LDA model parameters.

In R, there are mainly two packages providing LDA models: LDA and topicmodels. The former provides classic LDA based on Gibbs sampling, MMSB (the Mixed Membership Stochastic Block Model), RTM (Relational Topic Model), and sLDA (supervised LDA), RTM based on VEM (Variational

Expectation-Maximization). The latter provides LDA_{VEM}, LDA_{Gibbs}, and CTM_{VEM} (correlated topics model).

(2) Semantic Chunk Annotation. After initial LDA topic identification, for the supporting documents of each topic, Python is used with part-of-speech tagging, syntactic analysis, and grammatical analysis to extract noun chunks representing subjects and objects in the supporting documents of a certain topic. This can be divided into three steps: TAG, CHUNK, and ROLE.

First, TAG: According to the role of each word in the sentence, part-of-speech tags are assigned, mainly including verbs (VB), nouns (NN), pronouns (PR + DT), adjectives (JJ), adverbs (RB), prepositions (IN), conjunctions (CC), and interjections (UH).

CHUNK: Chunk labels are assigned to groups of words that belong together (i.e., phrases), such as noun phrases (NP, e.g., "the red coat") and verb phrases (VP, e.g., "is doing"), as shown in Table 1 :

Table 1 Chunk Labels and Their Meanings

Chunk Label	Pattern	Example
NP	DT + RB + JJ + NN + PR	the strange bird
PP	TO + IN	in between
VP	RB + MD + VB	was looking
ADJP	CC + RB + JJ	warm and cozy
SBAR	-	whether or not
INTJ	-	hello

ROLE: Semantic role labels describe relationships between different chunks, clarifying the function of chunks in sentences. The most common roles in sentences are SBV (subject noun phrase) and OBJ (object noun phrase). The subject of a sentence is the person, thing, place, or idea that does something. The object of a sentence is the person/thing affected by the action, as shown in Table 2 :

Table 2 Chunk Semantic Role Labels and Their Meanings

Semantic Role Label	Pattern	Example
SBV	NP	the boys sat on the Chair
OBJ	NP + SBAR	the boys sat on the Chair

To intuitively explain the above process, take the sentence "Phrase-LDAvis model is helpful to detect the core technology topic" as an example for semantic chunk extraction testing. The results are shown in Figure 3 [Figure 3: see original paper], which can annotate the part-of-speech, chunk, and semantic role of each word, ultimately obtaining the noun chunk "Phrase-LDA model"

representing the subject component and the noun chunk "the core technology topic" representing the object component.

(3) Chunk-LDA Annotation. Based on the semantic chunk extraction results from step (2), the topic words (single topic words) from the initial LDA topic identification results in step (1) are annotated with chunks to achieve chunk annotation of LDA topic identification results. For example, annotating the topic word "technology" in the sentence "Phrase-LDAvis model is helpful to detect the core technology topic" with chunks: technology → the core technology topic, resulting in semantically enhanced Chunk-LDA and improving the readability (semantic function) of topic identification results.

Since a topic word may correspond to several subject and object noun chunks, the key problem in this step is how to determine the chunk corresponding to the topic word. The solution adopted in this paper is: first, determine the corresponding generated documents based on the topic corresponding to the topic word (topic word-topic-topic document), then extract the semantic chunks of these corresponding documents and sort them by frequency, and finally select the chunk with the highest frequency corresponding to the topic word as a clue to complete Chunk-LDA construction.

2.2 Core Technology Topic Identification Based on SNA

There are explicit or implicit connections between technology topics contained in patent literature, and these connections can reveal the importance and core value of a certain technology topic. For example, the more connections technology topic T has with other topics, the higher the centrality of topic T. Current LDA topic identification methods can identify topics in large amounts of text but cannot analyze which topics are core topics. Therefore, this study attempts to further process LDA topic identification results using Social Network Analysis (SNA) methods: on the basis of LDA topic identification results, construct an LDA topic social network graph, and determine core technology topics through centrality indicators. The centrality calculation method is shown in formula (2):

$$C_i(T) = \sum A_{ij} C_j \quad \text{formula (2)}$$

Where $C_i(T)$ is the centrality of topic T_i, calculated using Bonacich' s Centrality, i.e., eigenvector centrality; A_{ij} is the adjacency matrix of the network, λ is a constant, and C_j is the neighbor of node C_i.

For example, several technology topic sets identified based on the LDA model are marked as $T = \{topic1, topic2, topic3, ..., topicn\}$. Then, based on the igraph package in R, the topic network G is constructed, and the centrality values $C_i(T)$ of each node are calculated. The size of the centrality values is represented by the size of topic nodes, as shown in Figure 4 [Figure 4: see original paper].

The specific processing tool is the igraph package in R for constructing topic network G, with visualization layout settings code shown in Figure 5 [Figure 5: see original paper].

2.3 Core Technology Topic Visualization Analysis Based on Chunk-LDAvis

Scientific and technological intelligence analysis should be user-oriented, but current core technology topic identification research results are mainly displayed as local static graphs, making it difficult to analyze core technology topic content at multiple levels and granularities. Users can often only view content provided by intelligence analysts. With the development of information technology, interactive visualization technology can compensate for this deficiency to a certain extent, i.e., through interactive visualization results, scientific and technological intelligence results can be displayed at multiple levels to meet users' personalized needs.

Moreover, although the previous step identified core technology topics, the specific relationships and content (subordinate words of topics) between core technology topics and other topics cannot be clearly obtained. Therefore, further analysis is needed. This study uses Multidimensional Scaling (MDS) to construct a low-dimensional space using Euclidean distances between topics, making the distances between LDA topics in this space as consistent as possible with those in high-dimensional space. The distance between topics indicates their correlation, which can be used for further analysis of core technology topics.

Based on the above analysis, this study explores using the LDAvis toolkit in R to draw interactive core technology topic visualization maps. The basic visualization layout and meaning are shown in Figure 6 [Figure 6: see original paper].

Figure 6 can be mainly divided into left and right parts. The left side visualizes LDA topics in two-dimensional space based on the MDS algorithm, where dots represent topics (the numbers in the dots are the LDA topic identification result numbers), and the size of the dots is determined by the number of documents corresponding to the topic. The right side shows the terms corresponding to the topic, sorted by generation probability. This figure is generated using Web front-end tools and has good interactive visualization effects. Taking topic1 as an example, clicking the topic1 dot interactively displays the subordinate terms of topic1 on the right. Clicking a term or chunk on the right can correspondingly display the corresponding topic. Based on the above processing steps, more comprehensive and intuitive core technology topic analysis can be conducted.

In addition, the parameter λ ($0 \leq \lambda \leq 1$) can be adjusted to control the topic-term relevance $r(term_w|topic_t)$, i.e., to control the display of different subordinate terms of a topic. The parameter λ calculation method is shown in formula (3):

$$r(w,k|\lambda) = \lambda log(\phi_{kw}) + (1-\lambda) log\left(\frac{\phi_{kw}}{p_w}\right) \quad \text{formula (3)}$$

Where w represents topic words, w {1, 2, 3, ···, V}; k represents topics, k {1, 2, 3, ···, K}; ϕ_{kw} represents Gibbs sampling parameters; p_w represents the distribution probability of topic word w. When $\lambda = 0$, it displays unique and relatively independent subordinate terms under the topic, i.e., these terms often appear only in that topic. When $\lambda = 1$, it displays subordinate terms with higher distribution probability, but these high-probability terms often do not belong exclusively to that topic and may also belong to other topics.

2.4 Characteristics and Advantages

Compared with the core technology topic identification framework based on citation features and text content features, the framework constructed in this study has the following characteristics and advantages:

From the perspective of result accuracy, it improves upon core technology topic identification methods based on the classic LDA model (which judges core technology topics by the number of topic-related documents, assuming that topics with more related documents are more likely to be core technology topics) by adding a topic association perspective for core technology topic determination.

From the perspective of result content, each core technology topic is composed of a set of noun chunks, which have stronger semantic expression capabilities than a set of words or patent numbers, making it easier for users to interpret.

From the perspective of result presentation, compared with static core technology knowledge graphs, the dynamic and interactive visualization graph form is more user-friendly and facilitates intelligence analysis.

3 Empirical Study

3.1 Data Source

This study uses patent data in the nano-agriculture field from January 1, 2010 to December 31, 2017, collected in the Derwent Innovations Index (DII) database as the data source. The DII database is a Web-based patent information database that includes more than 10 million basic invention patents from over 40 patent agencies (covering more than 100 countries) and more than 20 million patent information records, making it feasible and effective for studying global patent R&D status and technical breakthrough information in a technical topic. Therefore, using DII database nano-agriculture patent data as the data source for identifying core technology topics in the nano-agriculture field is feasible and effective.

In the DII database, the search query "Keyword = ' Nano agriculture*' " was used, with the search time span from January 1, 2010 to December 31, 2017,

yielding 4,937 results. The annual patent numbers are shown in Figure 7 [Figure 7: see original paper].

3.2 Semantically Enhanced LDA Topic Identification

The first step in LDA topic identification is to estimate the number of topics (K) in the input document collection. Current research mainly uses perplexity and log-likelihood value changes for estimation. The former decreases with increasing topic numbers, while the latter increases with increasing topic numbers. Generally, the topic number when both changes tend to flatten can be used as the estimated topic number. This study uses log-likelihood values to determine the optimal topic number.

Currently, before determining the optimal topic number, a priori estimation of the number of topics contained in the dataset is needed. This study estimates that the number of topics in the downloaded patent dataset is within 100. Therefore, iterative experiments were conducted to determine the optimal topic number, with K ranging from 1-100 in steps of 25, running 1,000 iterations for each topic number to obtain each K and its corresponding log-likelihood value, as shown in Figure 8 [Figure 8: see original paper].

As can be seen from Figure 8, when the topic number is 90, the log-likelihood value of the LDA model tends to stabilize, and reaches its maximum at 97. Therefore, this experiment selects K = 97 topics. After determining the LDA topic number, the topicmodels package in R is used for LDA topic identification, and the topic identification results are saved locally for chunk annotation. After obtaining the initial LDA topic identification results, Python is used for chunk extraction as described above. Partial chunk extraction results are shown in Figure 9 [Figure 9: see original paper]. Then, based on the chunk extraction results, the initial LDA topic identification results are annotated with chunks to obtain semantically enhanced LDA topics (Chunk-LDA topics), which are saved locally in the form of a Chunk-LDA topic-document matrix. Partial results are shown in Figure 10 [Figure 10: see original paper].

3.3 Core Technology Topic Identification Based on SNA

Based on the previous step's data processing results (semantically enhanced LDA topic identification), Social Network Analysis (SNA) methods are used to further process LDA topic identification results: on the basis of LDA topic identification results, an LDA topic social network graph is constructed, and core technology topics are determined through centrality indicators. Social network analysis was conducted on the identified 97 topics, and an LDA topic visualization network was constructed, with results shown in Figure 11 [Figure 11: see original paper], where node size is determined by centrality.

Through calculation and sorting, the ranking of core technology topics and their $C_i(T)$ values are obtained, as shown in Table 4 . Combined with visualization results, core topics in the nano-agriculture field can be intuitively discovered

through centrality analysis, such as Topic1, Topic40, Topic60, etc., which are located at the core positions of the nano-agriculture field topic network, thus can be judged as core technology topics.

Although the SNA-based method can identify core technology topics in the nano-agriculture field, its interpretation and analysis still have certain difficulties and cannot meet actual intelligence analysis needs. Therefore, this study further processes the results using Web front-end technology for interactive visualization processing to enhance result readability and analysis dimensions. Finally, based on the core technology topic visualization map, core technology topics in the nano-agriculture field are analyzed.

3.4 Core Technology Topic Visualization Analysis Based on Chunk-LDAvis

Based on the core technology topic identification results from the previous step, the top 15 core technology topics are selected, and the LDAvis toolkit is used to draw interactive core technology topic visualization maps for the nano-agriculture field. Figure 12 [Figure 12: see original paper] shows the static results of core technology topic visualization in the nano-agriculture field. The dynamic, interactive visualization results have been uploaded to a self-built website and can be accessed online (https://www.informationscience.top/coretechnologytopic/). In the webpage, the dots representing topic numbers on the left can be clicked, and hovering the mouse over a dot displays the Top-30 noun chunks constituting that topic. The noun chunks on the right can also be clicked, and hovering over a noun chunk displays the corresponding topic.

In Figure 12, it can be found that the positions of 7 core topics (Topic1, Topic40, Topic59, Topic60, Topic73, Topic12, and Topic33) are basically consistent with their positions in the LDA topic network graph. However, the node size in Figure 12 is proportional to topic probability, thus differing from the topic node size in the LDA topic network graph (which is proportional to topic centrality).

Based on the above results, comprehensive analysis of core technology topics in the nano-agriculture field is conducted. The top 3 core technology topics are selected for specific analysis:

(1) Topic1 - Nano-pesticides. Analysis of the subordinate phrases under Topic1 reveals three main forms of nanotechnology applications in pesticides: Using nano-processing technology to nano-size pesticide active ingredients into nano-dispersions, nano-emulsions, nano-particles, or nano-microspheres, increasing the specific surface area of pesticide formulations, improving oil solubility or water miscibility, enhancing dispersibility and stability in water, and promoting absorption. Such nano-pesticides include thiacloprid nano-particles, pyraclostrobin pesticides, cyfluthrin nano-emulsion compositions, and nano-particles of benzoxazole and phenyl compounds. Using nano-carriers to load pesticides, improving the stability of environmentally sensitive pesticides, enhancing

drug adhesion and permeability on crop surfaces, reducing loss. Incorporating metals or inorganic materials into pesticides to enhance bactericidal and photocatalytic effects, promoting pesticide decomposition and reducing pesticide residues, such as new photocatalyst insecticides and nano-titanium dioxide composite pesticides. Additionally, some new nano-pesticides and pest control slow-release agents can increase plant pest resistance or fungal resistance, inhibit microbial growth and reproduction, ensure plant robustness, and have good weeding efficiency.

(2) Topic40 - Agricultural Equipment and Devices. Analysis of this topic' s specific content reveals that nanotechnology applications in agricultural equipment and devices mainly concentrate in three aspects: Irrigation systems, water purification systems, and aquaculture systems. Nanotechnology applications in water use, treatment, and multi-systems for irrigation, purification, and aquaculture are mainly reflected in nano-purification aerators and nano-bubble generators using nano-tubes for drainage and water purification, nano-carbon cloth for insulation, and nano-support plates and nano-trays for load-bearing and containment. Greenhouse devices. Nanotechnology applications in greenhouse devices concentrate on using carbon nanotubes to collect solar thermal energy, using nano-carbon cloth and nano-glass for insulation, using nano-coatings for power generation and sterilization, and using nano-power generation glass and nano-grids for photosynthesis, etc. Self-propelled combine harvesters, seeders, and fertilizer applicators. Nanotechnology applications in such agricultural machinery mainly concentrate on nano-frames and nano-baffles for support, stress, and protection, nano-transport systems for transportation, nano-tubes and nano-bags for collection, and nano-fibers for decontamination.

(3) Topic59 - Agricultural Environment Improvement. Analysis of this topic' s specific content shows that due to their huge specific surface area and modifiable functional groups, nanomaterials easily combine with organic compounds and heavy metal particles and other pollutants in the environment, playing an increasingly important role in agricultural environment improvement. Current research focuses on using zinc oxide/diatomite nano-composite materials for sewage treatment; nano-titanium dioxide as a biological adsorbent can have excellent adsorption capacity, high heavy metal selectivity, and high degradation ability to remove organic pollutants, pathogenic bacteria, and microorganisms; how to use graphene oxide and iron oxide magnetic nano-particles to prepare magnetic nano-bactericides is also an important content of this topic. Using nanotechnology to nano-size silver, nano-silver sterilization has spectrum antibacterial, strong bactericidal, strong permeability, and persistent antibacterial characteristics, and is widely used in agricultural environment improvement, especially in sewage treatment and antibacterial sterilization. In sewage treatment, this topic mainly studies titanium dioxide nano-composite hydrogel softening reactors for sewage softening, using titanium dioxide photocatalysts to remove algae, and using nano-silver composite materials to decompose and degrade organic toxins and their bactericidal applications. In antibacterial sterilization, it mainly studies using silver aqueous polyurethane for antibacterial,

sterilization, and deodorization, using silver nano-particle stable suspensions, adsorbing silver ions onto silica nano-particles as biocides, or preparing biocide compositions by mixing hydroxypyridone compounds and silver compounds to effectively inhibit and eliminate microorganisms in water.

3.5 Verification of Core Technology Topic Identification Results

The empirical results are compared with the results of specific nano-agriculture field patent analysis practice work to verify the feasibility and effectiveness of the core technology topic identification method proposed in this study. In specific practice work (nano-agriculture field patent trend research analysis work, with the same original data), Thomson Innovation (TI) under Clarivate Analytics was used to draw a patent map of the nano-agriculture field (see Figure 13 [Figure 13: see original paper]). The altitude of peaks in the patent map represents the density of literature on specific topics and shows the relative relationships between different topics, which can be used for core technology topic analysis.

Analysis of Figure 13 reveals that core technology topics in the nano-agriculture field mainly include seven topics: pesticides, fertilizers, agricultural equipment and devices, agricultural product processing, agricultural planting and cultivation, agricultural environment improvement, and animal/plant genetic breeding and nano-detection. Through comparison and verification with the core technology topic identification results obtained in this study, it can be found that the identified core technology topics Topic1 - Nano-pesticides, Topic40 - Agricultural Equipment and Devices, and Topic59 - Agricultural Environment Improvement correspond to results 1, 3, and 6 in the patent map, which can verify the feasibility and effectiveness of the method proposed in this paper to a certain extent.

3.6 Discussion

Compared with core technology topic identification methods based on the classic LDA model, the method proposed in this paper improves the insufficient semantic information of single topic words in classic LDA results through semantic chunk annotation. On the other hand, compared with simply relying on topic distribution probability to judge core topics, it proposes a method based on social network and multidimensional scaling analysis to identify associations between topics and their visualization. Compared with current core patent topic analysis methods based on keyword and classification code co-occurrence, the core technology topic identification method based on Chunk-LDAvis proposed in this paper is more targeted and readable (not single keywords or classification codes, but basic knowledge units are noun chunks representing subjects or objects), and can interactively visualize and analyze core technology topics in a technical field, improving the readability of identification results.

However, this method also has certain limitations. For example, regarding the problem of insufficient semantic information of core technology topics, this paper

solves it by constructing Chunk-LDA topics, which are obtained through semi-automatic methods, resulting in certain deficiencies in analysis efficiency. Therefore, more effective machine learning methods need to be explored to achieve automated construction of Chunk-LDA topics. In addition, patent topic representation methods based on semantic TRIZ can also solve the problem of insufficient semantic information in current core technology topic identification research, i.e., representing basic patent knowledge units as SAO (Subject-Action-Object) structures, and then dividing different dimensions to achieve multi-level core technology topic analysis at macro, meso, and micro levels.

This paper proposes a core technology topic identification method based on Chunk-LDAvis that can be used to analyze core technology topics in a patent field. The main innovations are: (1) proposing a new LDA topic analysis method based on semantic chunk annotation, and (2) using Web front-end technology to achieve visualization analysis of the implicit relationships of core technology topics. Finally, taking the nano-agriculture field as an example, 4,937 patent documents from 2010 to 2017 were selected as the data source, and an empirical study was conducted using the proposed core technology topic identification method, proving that the method is feasible and effective. However, the proposed core technology topic identification method has two main deficiencies: (1) Chunk-LDA topics are obtained through semi-automatic methods, resulting in low analysis efficiency when the volume of patent data to be analyzed is too large; (2) it cannot fully analyze the development trends of core technology topics. Therefore, the next step is to conduct automated construction of Chunk-LDA topics and identification research on core technology topic evolution paths to achieve dynamic tracking of core technology topics.

References

[1] Wang Xiaoyue, Bai Rujiang. Research on Automatic Classification Technology for Massive Network Academic Literature [M]. Beijing: People' s Publishing House, 2015: 40-42.
[2] Schankerman M, Pakes A. Estimates of the value of patent rights in European countries during the post-1950 period [J]. Economic journal, 1986, 96(384): 1052-1076.
[3] Xu Haiyun, Yue Hui, Lei Bingxu, et al. Core patent mining based on co-occurrence of patent technology efficacy keywords and patent citations [J]. Library and Information Service, 2014, 58(4): 59-67.
[4] Yuan Run, Qian Guo. Rough set theory model for identifying core patents [J]. Library and Information Service, 2015, 59(2): 123-130.
[5] Ma Yongtao, Zhang Xu, Fu Junying, et al. Review of core patents and their identification methods [J]. Journal of Intelligence, 2014, 33(5): 38-43, 70.
[6] Kwon O, Seo J, Noh K, et al. Categorizing influential patents using bibliometric analysis of patent citation network [J]. Information-an international interdisciplinary journal, 2007, 10(3): 313-326.
[7] Choi C, Park Y. Monitoring the organic structure of technology based on

the patent development paths [J]. Technological forecasting and social change, 2009, 76(6): 754-768.
[8] Hsu C W, Chang P L, Hsiung C M, et al. Charting the evolution of biohydrogen production technology through patent analysis [J]. Biomass & bioenergy, 2015, 76(5): 1-10.
[9] Zhang Xin, Ma Ruimin. Research on core patent discovery based on improved PageRank algorithm [J]. Library and Information Service, 2018, 62(10): 106-115.
[10] Kang Chuanbo, Wang Wei, Mu Xiaomin, et al. Comprehensive value model for core patent identification [J]. Information Science, 2018, 36(2): 67-70.
[11] Wang Y, Bai H J, Stanton M, et al. PLDA: parallel latent Dirichlet allocation for large-scale applications [C]//International conference on algorithmic aspects in information and management. San Francisco: Springer-verlag, 2009: 301-314.
[12] Newman M E J, Girvan M. Finding and evaluating community structure in networks [J]. Physical review, 2004, 69(2): 108-118.
[13] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks [J]. Journal of statistical mechanics: theory and experiment, 2008, 30(2): 155-168.
[14] Lee Y G, Song Y I. Selecting the key research areas in nanotechnology field using technology cluster analysis: a case study based on National R&D Programs in South Korea [J]. Technovation, 2007, 27(12): 57-64.
[15] Luan Chunjuan, Zeng Guoping. Research on measurement of core technology fields based on SNA [J]. Library and Information Service, 2011, 55(6): 33-35.
[16] Fan Yu, Fu Hongguang, Wen Yi. Patent information clustering technology based on LDA model [J]. Computer Applications, 2013, 33(S1): 87-89, 93.
[17] Li Jiajia, Ma Tiejun. Core technology identification and trend analysis of wind energy based on patent data [J]. Science and Technology Management Research, 2017(12): 129-136.
[18] Yi Huifang, Wu Hong, Ma Yongxin, et al. Patent technology topic analysis based on LDA and strategic coordinates: taking the graphene field as an example [J]. Journal of Intelligence, 2018, 37(5): 97-102.
[19] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation [J]. Journal of machine learning research, 2003(3): 993-1022.
[20] Blei D M, Lafferty J. Dynamic topic models [C]//Proceedings of the 23rd international conference on machine learning. New York: ACM, 2006: 113-120.
[21] Wang X, McCallum A, Wei X. Topical N-Grams: phrase and topic discovery, with an application to information retrieval [C]//IEEE international conference on data mining. Omaha: IEEE Computer Society, 2007: 697-702.
[22] ElKishky A, Song Y, Voss C R, et al. Scalable topical phrase mining from text corpora [J]. Proceedings of the VLDB endowment, 2014, 8(3): 305-316.
[23] Li B, Wang B, Zhou R, et al. CITPM: A cluster-based iterative topical phrase mining framework [C]//International conference on database systems for advanced applications. Dallas: Springer International Publishing, 2016: 197-213.

[24] Sievert C, Shirley K. LDAvis: a method for visualizing and interpreting topics [C]//Proceedings of the workshop on interactive language learning, visualization, and interfaces. Baltimore: Association for Computational Linguistics, 2014: 63-70.
[25] Fan Yunman, Ma Jianxia. Research on emerging topic detection based on LDA and emerging topic feature analysis [J]. Journal of the China Society for Scientific and Technical Information, 2014, 33(7): 698-711.
[26] Zhang Qin, Zhang Zhixiong. Research on topic phrase mining method based on PhraseLDA model [J]. Library and Information Service, 2017, 61(8): 120-125.
[27] Landauer T K, Dumais S T. A solution to Plato' s problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge [J]. Psychological review, 1997, 104(2): 211-240.
[28] Shen C, Li T, Ding C H Q. Integrating clustering and multi-document summarization by bi-mixture probabilistic latent semantic analysis (PLSA) with sentence bases [C]//AAAI conference on artificial intelligence. San Francisco: AAAI Press, 2011: 914-919.
[29] Bonacich P B. Factoring and weighting approaches to status scores and clique identification [J]. Journal of mathematical sociology, 1972, 2(1): 113-120.
[30] Sturrock K, Rocha J. A multidimensional scaling stress evaluation table [J]. Field methods, 2016, 12(1): 49-60.

Author Contributions

Liu Ziqiang: Designed the overall research framework and wrote the paper;
Xu Haiyun: Proposed the research idea and guided paper revision;
Yue Lixin: Wrote the results analysis section;
Fang Shu: Guided paper revision.

Note: Figure translations are in progress. See original paper for figures.