Research on Key Core Technology Identification based on Patent Semantic Analysis: A Case of the Biological breeding field

Shi Min[1],Quan Bin[1],Luo Jian[1*]

1Business School of Hunan Agricultural University,Changsha, 410125

[*] Luo Jian, 4658669@qq.com

Abstract: To achieve high-quality development, China must break through key core technologies. Accurately identifying key core technologies is crucial for the government or enterprises to optimize their research and development layout and actively seize the technological high ground. This study used LDA semantic analysis to classify patents in the field of biology, and selected topics in the field of biological breeding based on topic keywords; Conduct a competitive analysis of two key core technologies: gene editing technology and plant transgenic technology. It is recommended to closely monitor the latest research and development progress in these two technologies by the world's leading seed companies. Jointly with relevant entities, focus on tackling key core technologies, strengthen high-level open cooperation with domestic and international entities, accelerate the integration of informatics with biology, and promote the breakthrough of key core technologies in the field of biological breeding. These actions aim to provide decision support for the high-quality development of China's biological breeding sector.

Keywords: key core technologies;LDA semantic analysis;biological breeding;high-quality development; seed industry

"High-quality seeds lead to abundant harvests, and food security ensures the safety and well-being of the people." Since the 18th National Congress, China's seed industry has developed rapidly, with continuous improvements in independent innovation capabilities. However, there remain key issues that need to be addressed, such as inadequate protection and utilization of germplasm resources and a low level of breeding innovation. General Secretary Xi Jinping emphasized that "scientific and technological research must be problem-oriented, targeting the most urgent and pressing issues." In the context of comprehensively promoting the revitalization of the seed industry and achieving high-quality development in the sector, focusing efforts on breaking through the key core technologies in the seed industry is of utmost importance. The identification of key core technologies in the seed industry is the primary task in this regard. Only by accurately identifying these key core technologies can we conduct organized scientific research, solve the "Bottleneck" technological challenges, continuously enhance the independence, autonomy, and security of China's seed industry development, and firmly grasp the initiative in innovation and development.

In recent years, academic research on the identification of key core technologies has deepened, mainly focusing on critical fields such as photolithography machines and 5G communication. However, less attention has been given to the biological breeding sector. In the context of global climate change, tense geopolitical situations, and China's comprehensive efforts to promote the revitalization of the seed industry, the importance of identifying key core technologies in the biological breeding field has become increasingly prominent. Therefore, this paper uses semantic analysis methods to analyze patents in the field of biological breeding, identify key core technologies, and compare China's technological competitive capacity with leading global institutions to identify gaps. It also analyzes the R&D characteristics of global leaders, providing decision support for technological innovation in China's biological breeding sector.

1 Literature Review

1.1 Concept and Characteristics of Key Core Technologies

Key core technologies are national treasures, crucial for enhancing the leading role of technological innovation. They are the guarantee for achieving high-level technological self-reliance and strength, and they play an essential role in driving China's high-quality development and realizing the second centenary goal (Du Chuanzhong, 2023). According to the Modern Chinese Dictionary, "key" refers to the most important part or the decisive factor in a matter, while "core" refers to the central part. Scholars currently explain key core technologies from two main perspectives: ①Key core technologies are a composite concept formed by key technologies and core technologies. They are technologies that occupy a core position and play a crucial role in a specific industry or field during a particular historical period (Jiang Yao, Zhang Zhihe, Chen Jin, et al.).②The "key" in key core technologies represents their level of importance, while "core technologies" refer to the main body (Gu Shengzu, et al.). Key core technologies are considered as the most critical and decisive components within core technologies (Hu Xubo, et al.).

Scholars have also conducted research on the characteristics of key core technologies. Yang Wu, et al., believe that key core technologies possess foundational, critical, and competitive characteristics. On this basis, Jiang Yao, Chen Xu, and other scholars summarized the characteristics of key core technologies as frontier technology, complex innovativeness, and national strategic significance. Zhang Yuchen, et al., argued that key core technologies also exhibit systematic and highly knowledge-intensive characteristics. The characteristics of key core technologies are summarized in Table 1.

Table 1 Key core technology characteristics

Citation Source	Key Core Technology Characteristics	Characteristic Meaning
Jiang Yao [2], Chen Xu [3]	Frontier Technology	Technologies that have a continuous impact and consistently maintain a leading influence in the field.
Jiang Yao [2], Zhang Yuchen [10], Tan	Innovativeness/Originality	Technologies with high levels of originality and novelty, often developed from scientific discoveries or

Jinsong [4]		technological innovations, featuring pioneering exploration and significant breakthroughs.
Jiang Yao [2], Chen Xu [3], Xu Xia [5]	Core/Strategic National Significance	Technologies that occupy a core position in the entire technological system, are protected and recognized by multiple countries or regions, and contribute to the formation of a global leadership position for relevant domestic technologies.
Yang Wu [7], Huang Lucheng [12], Yu Jiang [8]	Foundational	Technologies that carry the core achievements of fundamental scientific research.
Yang Wu [7], Adomavicius [6], Song [9]	Critical	Technologies that are the foundation and key components of the entire technological system, with an overarching control over the direction of technological development.
Yang Wu [7], Wu Huabin [13], Frishammar [11], Xu Xia [5]	Competitiveness	Technologies that are important components of a company's core capabilities and serve as a key to transforming those capabilities into competitive advantages, containing significant economic and strategic value.
Zhang Yuchen [10]	High Knowledge Density	The technological system is essentially a collection of diverse technical knowledge, and key core technologies are the parts with the highest concentration of knowledge.
Zhang Yuchen [10]	Systematic	The breakthrough or advancement of key core technologies requires systematic collaborative capabilities.

Note: Source: Compiled by the research team

1.2 Research on Key Core Technology Identification Methods

The identification methods for key core technologies both domestically and internationally mainly include qualitative research methods and quantitative research methods (Xu Xia, 2022).

Qualitative research methods primarily rely on the Delphi method and questionnaire surveys. For example, Tang Zhiwei et al. identified the key core technologies in the electronic information industry by distributing questionnaires to academicians and experts within the industry. This method, based on the extensive collection of expert opinions and combined with industry and disciplinary development trends, forms technological predictions. It requires a high number and quality of experts and is subject to a certain degree of subjectivity and limitations (Li Weisi et al., 2022).

Quantitative analysis methods mainly include indicator analysis and text mining methods. Dong Kun et al., focusing on the blockchain industry in Shandong Province, identified 32 key core technologies through patent record indicators. However, this method results in a large granularity of the identified technologies, which reduces accuracy. As a result, some scholars have attempted to identify technologies at a finer granularity. Mao Jianqi et al. selected hot technologies based on co-occurrence frequency and then identified key core technologies through an indicator system (Mao Jianqi et al., 2022). Lü Kun et al. identified key technologies in the blockchain financial industry based on a combination of word segmentation methods and the LDA model (Lü Kun et al., 2022). Hu Kai et al. conducted research on the identification of industry-specific common technologies through patent text mining (Hu Kai et al., 2023).

In summary, current scholars mainly use single models for topic identification, predominantly focusing on industrial information fields, with limited research on the combination of multiple models for identifying key core technologies in the biological breeding field. This study employs the LDA-BERT model to perform topic analysis on patents in the field of biological breeding. Based on the established key core technology evaluation index system, it identifies key core technologies in this field and conducts topic evolution and technology competition analysis, providing a degree of innovation in the methodology.

2 Research Design

2.1 Research Approach

The identification method for key core technologies is divided into six specific stages:

①Data Collection and Preprocessing: Patent information is retrieved and downloaded from patent databases. The patent data is then cleaned and preprocessed, including word segmentation, to form a patent corpus.

②Topic Identification: The LDA-BERT model is established based on the BERT model and the LDA model, combining the advantages of both to perform in-depth topic mining.

③Topic Content Analysis: The key core technology evaluation index system is used to analyze the keywords, keyword contributions, and related patent content within each topic, in order to preliminarily identify key core technologies in the field of biological breeding.

④Topic Evolution Analysis: To study the evolution trend of key core technologies, the data for the identified key core technology topics is divided into four time windows. The cosine similarity between semantic vectors of topics in adjacent time windows is calculated to determine the correlation between topics, and topics with strong relationships are selected to map the evolution path of the research topics.

⑤Competitive Landscape Analysis of Key Core Technologies: Based on topic content and evolution analysis, key core technologies within the field are identified. The original patent data under each technology topic is used for comparative analysis to examine the current competitive landscape of key core technologies within the field.

The detailed research approach is illustrated in the figure below:

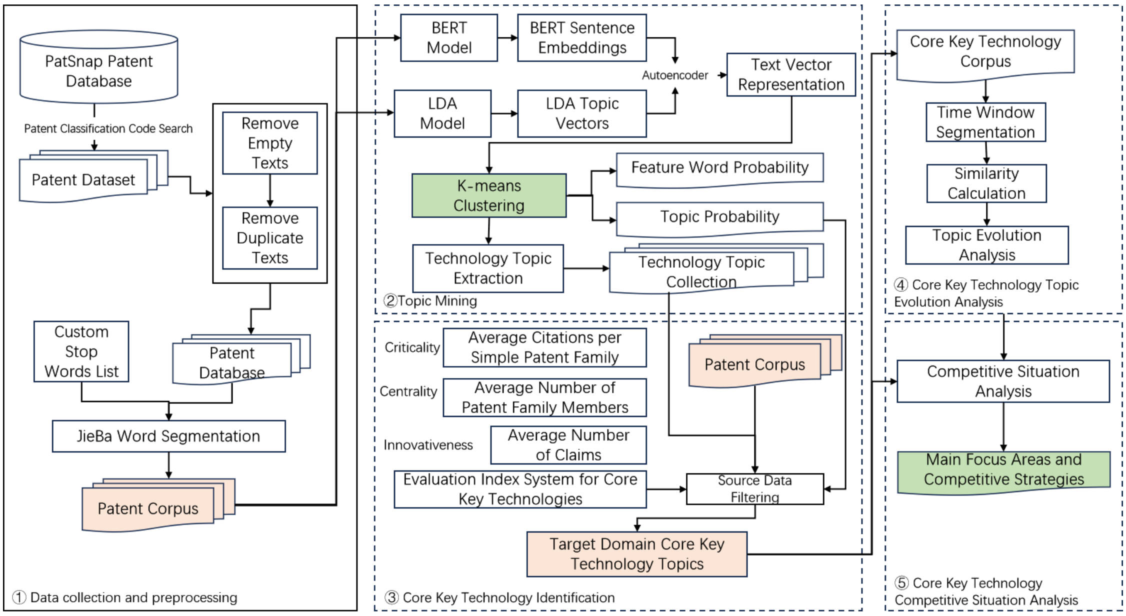
PatSnap Patent Database
Patent Classification Code Search
Patent Dataset
Remove Empty Texts
Remove Duplicate Texts
Custom Stop Words List
Patent Database
JieBa Word Segmentation
Patent Corpus
① Data collection and preprocessing
BERT Model
BERT Sentence Embeddings
Autoencoder
Text Vector Representation
LDA Model
LDA Topic Vectors
K-means Clustering
Feature Word Probability
Topic Probability
Technology Topic Extraction
Technology Topic Collection
②Topic Mining
Criticality
Average Citations per Simple Patent Family
Centrality
Average Number of Patent Family Members
Innovativeness
Average Number of Claims
Evaluation Index System for Core Key Technologies
Patent Corpus
Source Data Filtering
Target Domain Core Key Technology Topics
③ Core Key Technology Identification
Core Key Technology Corpus
Time Window Segmentation
Similarity Calculation
Topic Evolution Analysis
④ Core Key Technology Topic Evolution Analysis
Competitive Situation Analysis
Main Focus Areas and Competitive Strategies
⑤ Core Key Technology Competitive Situation Analysis

Figure 1 Design of key core technology identification process

2.2 Identification of Key Core Technology Topics

2.2.1 LDA Model

The Latent Dirichlet Allocation (LDA) model, proposed by Blei et al. in 2003, is a document topic generation model (Blei et al., 2003). The basic principle of the LDA topic model is to estimate the probability of word occurrences in a document through probability distributions, thereby discovering topics and keywords within the document. The joint probability distribution of the LDA topic model, which reduces the dimensionality of text representation, has been widely applied in the field of semantic mining (Zeng Ziming, 2019). The joint probability distribution of the LDA topic model is defined as shown in equation (1).

$$\mathrm{p}(\omega_{\mathrm{m}}, \mathrm{z}_{\mathrm{m}}, \theta_{\mathrm{m}}, \phi|\alpha, \beta) = \prod_{n=1}^{\mathrm{N_m}} \mathrm{p}\big(\omega_{m,n}\big|\phi_{z_{m,n}}\big) p\big(z_{m,n}\big|\theta_m\big) p(\theta_m|\alpha) p(\phi|\beta) \tag{1}$$

In the equation: ω_{m} represents the set of words in document m; $\mathrm{z_m}$ represents the set of topics corresponding to the words in document m; θ_{m} represents the topic distribution of document m; ϕ represents the topic-word distribution shared across all documents; α and β are the Dirichlet distribution parameters, used to generate the topic distribution θ_{m} and the topic-word distribution ϕ; $\mathrm{N_m}$ represents the number of words in document m; $\mathrm{z_{m,n}}$ represents the topic corresponding to the n-th word in document m; $\omega_{\mathrm{m,n}}$ represents the n-th word in document m. The specific process is illustrated in the figure below:

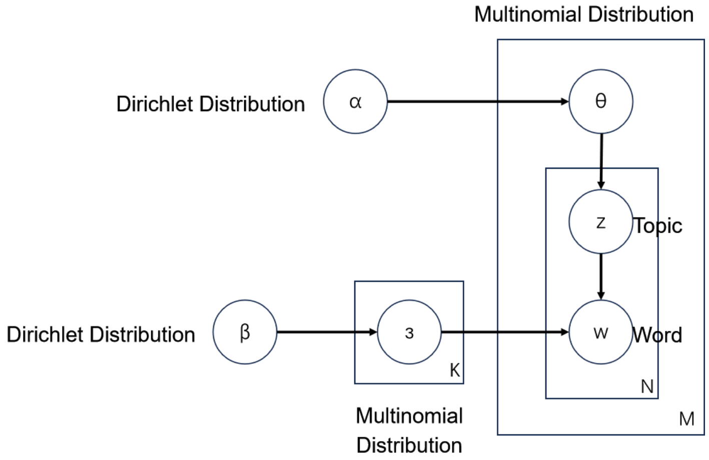

Figure 2 LDA model

2.2.2 BERT Model

BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained language representation method based on the Transformer model. Compared to traditional natural language processing models, BERT adopts a bidirectional Transformer encoder architecture, which enables it to capture the global dependencies of the input sequence more comprehensively and consider both contextual information. In the BERT model, the input text is first tokenized and mapped to corresponding embedding vectors. The multi-head self-attention mechanism is then applied to the vectors processed by the Transformer encoder, followed by residual connections and normalization. After being processed by a feed-forward neural network and a residual network, the BERT semantic feature vector is obtained, as shown in equation (2).

$$d = w_{ij}(\omega + \delta + \rho) \qquad （2）$$

In this equation: d represents the semantic feature vector; w_{ij} represents the weight coefficients; ω、δ、 respectively ρprepresent the word vectors, text vectors, and position vectors.

2.2.3 Vector Concatenation

In terms of text representation, the BERT model is used to generate sentence embedding vectors for documents to capture the semantic information of the text. The LDA model, on the other hand, generates probabilistic topic vectors for documents, reflecting the distribution of each document across different topics. To better integrate these two types of information, an autoencoder is used to concatenate these vectors, mapping the high-dimensional concatenated vectors into a lower-dimensional space. This process further extracts their latent semantic information and provides a more accurate and comprehensive feature representation for subsequent interdisciplinary topic identification, as shown in equation (3).

$$d'_m = (\mu; d_m) \qquad （3）$$

In this equation: d'_m represents the vectorized text representation after the fusion of BERT and LDA feature vectors; μ represents the probabilistic topic vector generated by the LDA topic model; d_m represents the sentence embedding vector generated by the BERT model.

2.2.4 Topic Clustering

To further determine key core technology topics, the UMAP (Uniform Manifold Approximation and

Projection) algorithm is used to reduce the dimensionality and visualize the latent space representations learned by the autoencoder. The K-means algorithm is then applied to cluster the latent space representations, aiming to discover underlying structures and patterns in the text data, and identify research topics with potential influence and significance. In topic clustering, it is typically necessary to determine the number of topics, which is usually assessed using evaluation metrics such as perplexity and topic coherence. In this study, perplexity is selected as the evaluation metric to determine the optimal number of topics, with the specific calculation methods shown in equations (4) and (5).

$$P（\mathrm{M}）=\exp\left\{-\frac{\sum_{m=1}^{M}\log_2 p(X_m)}{\sum_{m=1}^{M} N_m}\right\} \tag{4}$$

$$p(X_m)=p(z|m)p(x|z) \tag{5}$$

In these equations: P represents the perplexity function; M represents the number of documents in the test set; X_m represents the words in document m; $p(X_m)$ represents the probability of the word sequence in document m; $p(z|m)$ represents the probability of topic z given document m; $p(x|z)$ represents the probability of word x given topic z.

2.2.5 Construction of the Key Core Technology Evaluation Indicator System

Based on the summary of various scholars' research on the characteristics of key core technologies, an evaluation indicator system for key core technologies has been constructed, consisting of three indicators: criticality, centrality, and innovativeness, as shown in Table 2. For criticality, since patents in this field are frequently cited and referenced by subsequent patents, indicating that the patent plays a crucial role, the average citation number of the patent is selected as the indicator. For centrality, as patents are protected in multiple countries, indicating their core position, the average number of family members of the patent is chosen as the indicator for this characteristic. For innovativeness, the average number of patent claims is selected as the indicator, reflecting the breadth of effective innovation in the patent technology, and representing its innovativeness.

Table 2 Evaluation Indicator System for Key Core Technology

Key Core Technology Characteristic	Patent Indicator	Indicator Meaning	Indicator Calculation
Criticality	Average Citation Number of Simple Patent Families	The average number of citations by subsequent patents; a higher frequency indicates greater innovation influence on subsequent related technologies.	Average citation number of simple patent families: $N_1=\frac{F_2}{M}$, F $_1$ is the citation count of simple patent families for all patents in a given field T, M is the total number of patents in field T

Centrality	Average Number of Patent Family Members	The average number of countries where the same patent is protected; a higher average indicates stronger strategic protection of the technology.	Average number of patent family members:$N_2=\frac{F_2}{M}$ ， F_2 is the number of family members for all patents in field T, M is the total number of patents in field
Innovativeness	Average Number of Claims	The number of claims made by the patent; a higher number indicates a broader scope of effective innovation in the technology.	Average number of claims:$\mathrm{N}_3=\frac{\mathrm{F}_3}{\mathrm{M}}$, F_3 is the number of claims for all patents in field T, M is the total number of patents in field

2.3 Topic Evolution Analysis of Key Core Technologies

To more intuitively reveal the evolution trends of key core technology topics in the field of biological breeding, the research data is divided into four time windows. The probability distribution of topic words output by the LDA-BERT model is used as weights. By performing a weighted sum of topic words under the same topic, the topic probability vector is transformed into a topic semantic vector. The cosine similarity between topic semantic vectors in adjacent time windows is then calculated to assess the correlation between topics. Topics with strong correlations are selected to plot the evolution path of research topics. The cosine similarity calculation method is shown in equation (6).

$$\mathrm{sim}(\mathrm{T_i},\mathrm{T_j})=\frac{\sum_{\mathrm{K}=1}^{\mathrm{n}}\mathrm{x_k}(\mathrm{T_I})\cdot\mathrm{y_k}(\mathrm{T_j})}{\sqrt{(\sum_{\mathrm{k}=1}^{\mathrm{n}}\mathrm{x}_\mathrm{k}^2(\mathrm{T_i}))\cdot(\sum_{\mathrm{k}=1}^{\mathrm{n}}\mathrm{y}_\mathrm{k}^2(\mathrm{T_j}))}} \qquad （6）$$

In this equation: $\mathrm{sim}(\mathrm{T_i},\mathrm{T_j})$ represents the cosine similarity between vectors $\mathrm{T_i}$ and $\mathrm{T_j}$; $\mathrm{x_k}(\mathrm{T_I})$ and $\mathrm{y_k}(\mathrm{T_j})$ represent the k-th elements of vectors $\mathrm{T_i}$ and $\mathrm{T_j}$.

3 Experimental Process

3.1 Data Collection and Preprocessing

This study is based on the Zhihuiya patent database, where biological breeding is not specifically defined in the current patent classification. Therefore, the study collects patent data related to the biological breeding field as empirical research data. Patent classification numbers are used for searching, and technical topics are later filtered to obtain patents in the biological breeding field. The search was constructed using the primary strategic emerging industry classification “Biological Industry (A01H)” and the secondary classification “Biological Agriculture and Related Industries (C12N15)” as well as the classification code C12Q1/68 (nucleic acids). A total of 180,663 patents were retrieved. After excluding blank, duplicate, and empty abstract patents, 178,836 patents were obtained. The patent abstracts were then processed using Jieba word segmentation, during which a custom stopword list was created, and words with frequencies below 8 or above 90% were filtered out, resulting in a patent corpus.

3.2 Key Core Technology Topic Mining

Before topic mining, perplexity was calculated, and a perplexity curve was established to determine the optimal number of topics for each dataset using the elbow method.

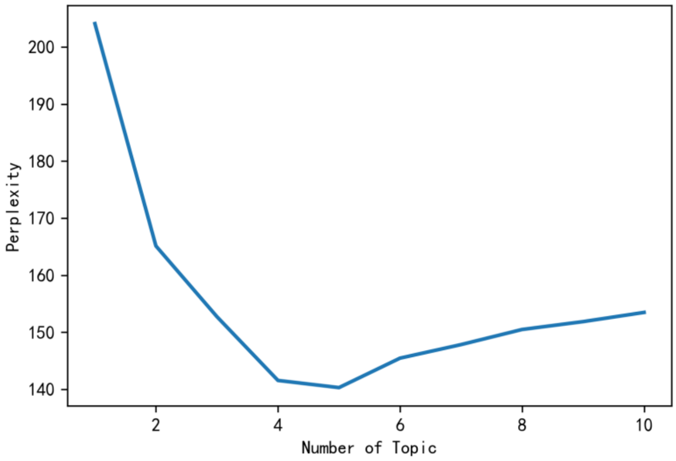

Figure 3 Confusion Degree of Different Themes

Based on the perplexity curve for different numbers of topics shown in Figure 3, the elbow method indicates that the optimal number of topics is K=5. Based on this, the LDA visualization code was used with K=5 to obtain the details of the five topics, as shown in Figure 4. It was observed that the overlap between Topics 1 and 3 was too high when five topics were set. However, when K=6, the topic perplexity remained at a lower level, and the visualization of the topics showed clearer boundaries between the topics, indicating that the topic classification was more distinct compared to when K=5. Therefore, K=6 was ultimately chosen as the optimal number of topics for topic identification.

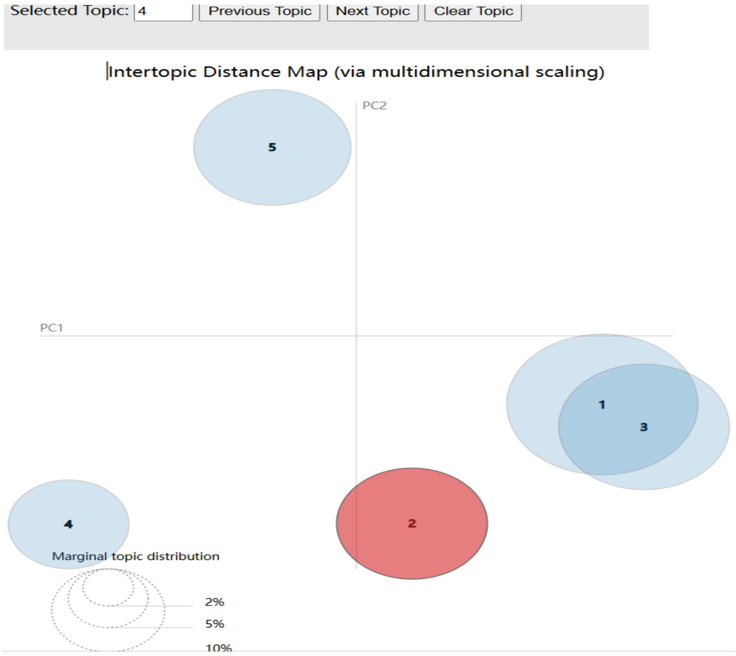
Selected Topic: 4
Previous Topic
Next Topic
Clear Topic
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
5
1
3
4
2
Marginal topic distribution
2%
5%
10%

Figure 4 Topic Visualization Image when K=5

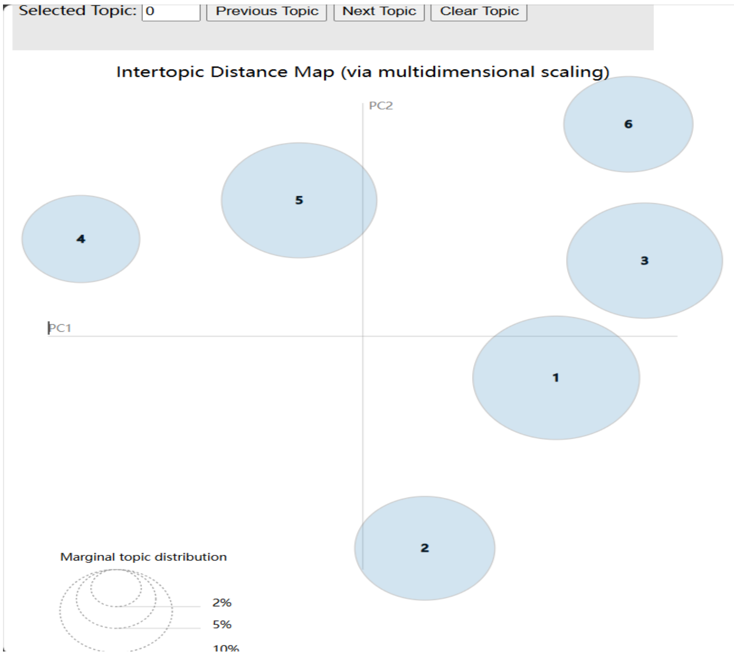
Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
6
5
4
3
1
2
Marginal topic distribution
2%
5%
10%

Figure 5 Topic Visualization Image when K=6

The topic classification was ultimately conducted with K=6, resulting in the top 10 characteristic words and the number of patents for each topic, as shown in Table 3.

Table 3 Some topics and their characteristic words

Topic Number	Top 10 Characteristic Words	Patent Count
Topic 0	pcr, genome, identification, kit, locus, sequencing, fragment, snp, mutation, base	36,839
Topic 1	amino acid, strain, transformation, bacillus, transgenic, synthesis, E. coli, microorganism, yeast, bacteria	31,197
Topic 2	virus, rna, promoter, transcription, antibody, editing, plasmid, targeting, immunity, tumor	43,346
Topic 3	new variety, tomato, backcross, flower, fruit, blooming, inbred line, self-pollination, green, cabbage type	21,939
Topic 4	culture medium, induction, pollination, rooting, explants, pollen, utility model, tissue culture, proliferation, propagation	20,316
Topic 5	rice, breeding, plant, traits, selection, callus, material, transgenic, resistance, wheat	25,205

This study focuses on the technological topics in the field of biological breeding. However, there is no specific category in the patent classification to categorize these technologies. To ensure the accuracy and effectiveness of topic clustering, the abstracts of high-contribution patents for all topics were read, and the applicant institutions for patents within each topic were statistically analyzed. It was found that Topic 2 includes keywords such as “virus,” “tumor,” and “immunity.” Moreover, the top five institutions by patent application number in this topic are all internationally renowned pharmaceutical companies: “IONIS PHARMACEUTICALS,” “REGENERON PHARMACEUTICALS,” “OPKO CURNA,” “ALNYLAM PHARMACEUTICALS,” and “ROCHE AG.” Therefore, it is clear that this topic primarily belongs to the biopharmaceutical field. Finally, by interpreting the characteristic words of each topic and combining them with patent abstracts, five key topics in the field of biological breeding were identified.

Table 4 Topics in the field of biological breeding

Topic Number	Top 10 Characteristic Words
Topic 0	pcr, genome, identification, kit, locus, sequencing, fragment, snp, mutation, base
Topic 1	amino acid, strain, transformation, bacillus, transgenic, synthesis, E. coli, microorganism, yeast, bacteria
Topic 3	new variety, tomato, backcross, flower, fruit, blooming, inbred line, self-pollination, green, cabbage type
Topic 4	culture medium, induction, pollination, rooting, explants, pollen, utility model, tissue culture, proliferation, propagation
Topic 5	rice, breeding, plant, traits, selection, callus, material, transgenic, resistance, wheat

3.3 Topic Content Analysis

Based on the established key core technology evaluation indicator system, the average citation number of simple patent families, the average number of patent family members, and the average number of claims are calculated from the perspectives of criticality, centrality, and innovativeness for the 15 topics in the field of biological breeding. These indicators are used to identify the key core technologies in the biological breeding field. The indicator values for various entities in the biological breeding field are shown in Table 5.

Table 5 Relevant indicator values of various topics in the field of biological breeding

Topic Number	Average Number of Claims	Average Number of Patent Family Members	Average Citation Number of Simple Patent Families
Topic0	14.71	9.93	40.10
Topic1	18.88	16.05	48.55
Topic3	5.09	2.22	3.15
Topic4	7.12	3.33	5.99
Topic5	11.40	6.26	16.16

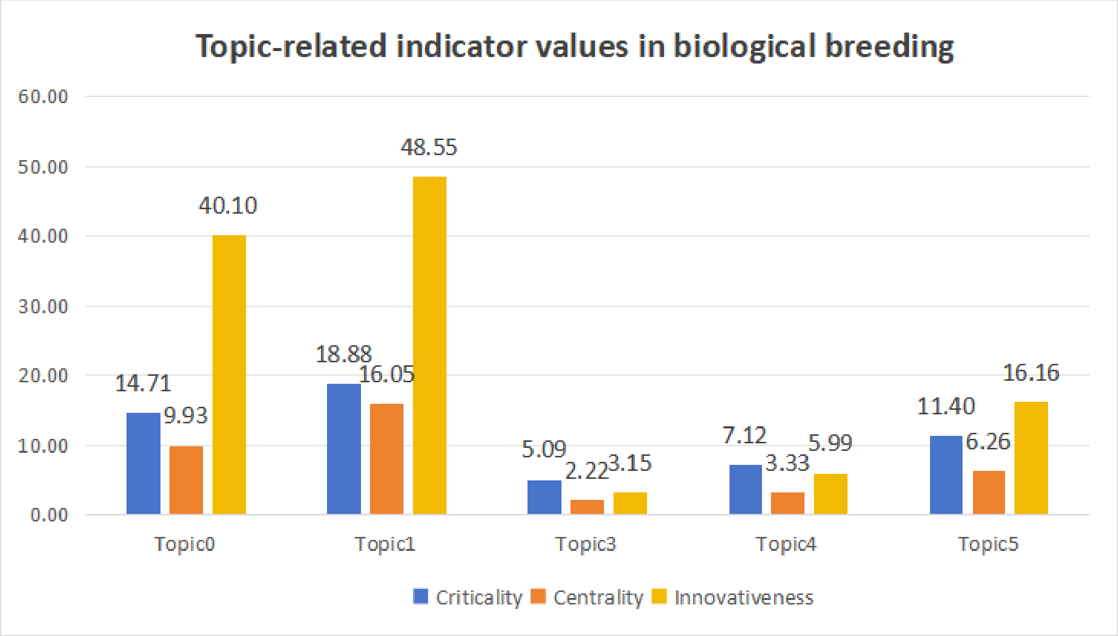

Figure 6 Indicator values related to various topics in biological breeding

Through the analysis of the data in Table 5 and the comparison of the data in Figure 6, it can be observed that Topic 0 and Topic 1 have a clear advantage in terms of criticality over the other topics, and they also rank among the top two in terms of centrality and innovativeness. Based on this, these two topics are identified as key core technologies in the field of biological breeding. The characteristic words and high-contribution patents for both topics were reviewed, and the following interpretations and naming were made:

①Topic0-Gene Analysis and Detection Technologies: This refers to a set of technologies used to study and interpret individual DNA sequences, such as polymerase chain reaction (PCR), gene sequencing, microarrays, and gene editing. In the field of biological breeding, these technologies can be used to identify and select genes with desirable traits, accelerating the breeding process and improving the quality and adaptability of crops or livestock, thus more efficiently cultivating new varieties that meet the demands of modern agriculture. The characteristic words for this topic mainly include PCR, genome, identification, kit, etc., indicating that the focus is primarily on gene analysis and detection technologies. To further define the topic, a review of high-contribution patents revealed that the patent CN111051537A, with a contribution of 0.9825, provides a method for determining a set of SNP loci, while CN113308550A, with a contribution of 0.9724, discloses a method for detecting sheep CRY2 gene insertion/deletion polymorphisms. Therefore, Topic 0 is named Gene Analysis and Detection Technologies.

②Topic1-Synthetic Biology: This refers to the design and construction of new biological components, devices, and systems to provide tools for improving crops and livestock. It enables precise editing or insertion of specific genes to create new varieties with ideal traits, accelerating traditional breeding processes and making customized organisms possible. The characteristic words for this topic mainly include amino acids, strains, transformation, bacillus, transgenics, synthesis, etc. A review of high-contribution patent literature shows that this topic focuses on the field of synthetic biology, emphasizing technologies and products related to synthetic biology. For example, patent JP6357702B2, with a contribution of 0.982, involves a new heat-stable fiber disaccharide hydrolase; IN289992B (providing a method for synthesizing modified N-glycosylation profiles of target proteins in plants or plant cells); CN114207129A (providing reagents and methods for replicating, transcribing, and translating in semisynthetic organisms), among others. These patents emphasize the design and construction of new biological components and systems, or the modification of existing biological systems, to provide strong support for biological breeding. Therefore, this topic is named Synthetic Biology.

3.4 Topic Evolution Analysis

To conduct a deeper analysis of the key core technology topics identified earlier, the evolution trends of the two topics were studied. The patent data for both topics were split according to four time windows: 2002–2006 (Phase 1), 2007–2011 (Phase 2), 2012–2016 (Phase 3), and 2017–2022 (Phase 4). Based on equations (4) to (6), a Python program was written to use the abstracts of the patents as the analysis content. After a series of data preprocessing steps, the LDA-BERT model was used for topic recognition and extraction. The optimal number of topics was determined by calculating the perplexity, followed by topic clustering. For each identified research topic, the top 10 words based on probability rankings were selected as the topic keywords. Ultimately, Topic 0 identified 40, 50, 50, and 40 keywords in the four time windows, respectively; Topic 1 identified 40, 40, 80, and 60 keywords in the four time windows, respectively. To visually present and better understand these topics, further analysis of the meaning of the topic words and the representative patents under each topic was conducted. Each topic was then named based on the weight of the keywords in the corresponding topic, as shown in Table 6.

Table 6 Evolutionary Path of Research Topics

Topic	Time Window	Number of Topics	Topic Name
Topic-0	Phase 1	4	Fundamental Molecular Biology Tools, Crop Variety Improvement, Genetic Improvement and Microbial Applications, Hybrid Breeding and Gene Editing
	Phase 2	5	Plant Metabolic Engineering and Stress Resistance Research, Amino Acid Metabolism and Herbicide Resistance, Crop Breeding and Transgenic Crops, Microbial Resource Exploration and Industrial Applications, Proteomics and Functional Genomics
	Phase 3	5	Microbial Genetic Engineering, Herbicide Resistance and Tolerance Mechanisms, Crop Breeding and Variety Improvement, Bacterial Gene Editing and Applications, Protein Engineering and Structural Biology
	Phase 4	4	Gene Detection Technologies and Reagent Kit Development, CRISPR-Cas9 Gene Editing Technology, Virus Detection and Nucleic Acid Purification Technologies, Molecular Marker-Assisted Selection and Precision Breeding
Topic-1	Phase 1	4	Protein Engineering, Crop Improvement and Transgenic Technologies, Breeding and Molecular Tools, Hybrid Breeding and Gene Modification

	Phase 2	4	Crop Breeding and Transgenic Crops, Protein Structure and Function Analysis, Microbial Metabolic Engineering, Crop Stress Resistance and Yield
	Phase 3	8	Abiotic Stress Response, Protein Engineering in Crops, Yeast Systems and Gene Editing, Crop Variety Improvement, Herbicide Resistance Breeding Strategies, Biosynthesis of Carbohydrates, Protein Sequence Feature Analysis, Microbial Communities and Industrial Fermentation
	Phase 4	6	Cotton Fiber Quality and Stress Resistance Improvement, Transgenic Crops, Proteomics and Gene Expression Regulation, Microbial Applications and Biotechnology, Microbial Diversity and Synthetic Biology, Microbial Metabolites and Biocatalysis

To further explore the evolutionary trajectories of key core technology topics in the field of biological breeding, we used Python programming to quantify the similarity between topics in adjacent time periods. By calculating the average similarity scores between all topics, a filtering benchmark was established. This was used to focus on those topics that exhibited high similarity for more detailed visualization analysis. The specific evolution paths of the two key core technology topics, Topic-0 and Topic-1, are shown in Figures 7 and 8. The lines connecting the topics represent the direction and relationship of the topic evolution flow, with the thickness of the lines indicating the cosine similarity between the topics. The thicker the line, the closer the evolutionary relationship between the topics.

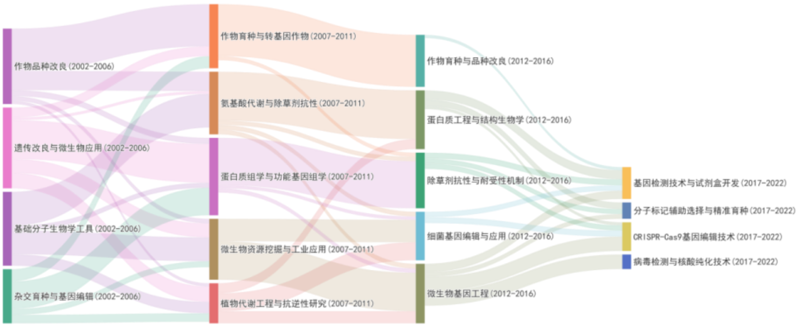

Figure 7: Evolution Path of Topic-0

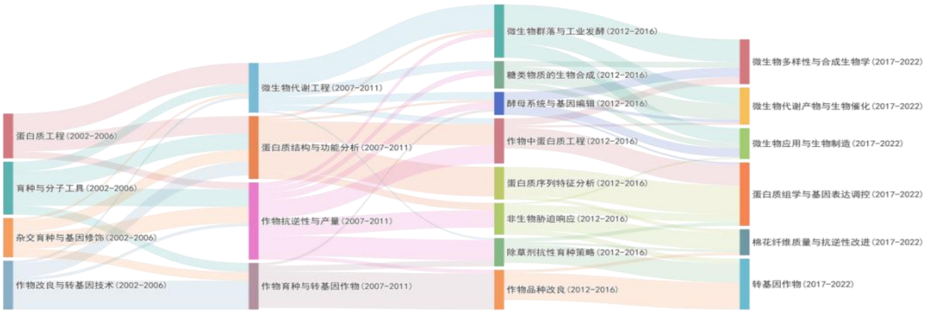

Figure 8: Evolution Path of Topic-1

The evolutionary paths of key core technology topics in the field of biological breeding, as visually displayed in Figures 7 and 8, intuitively demonstrate the development trends and evolutionary directions of research topics in different stages. From the topic evolution map, the following observations can be made:

In Figure 7, the evolution of the gene analysis and detection technologies topic shows that themes such as variety improvement and microbial applications exhibit inheritance, differentiation, and merging. The evolution begins with variety improvement, genetic enhancement, and basic tool research, followed by the incorporation of transgenic breeding, specific resistance studies, resource exploration, and functional gene utilization. Currently, the focus has shifted to technologies for studying and interpreting individual DNA sequences, such as polymerase chain reaction (PCR), gene sequencing, microarrays, and gene editing.

In Figure 8, the evolution of the synthetic biology technology topic shows a high degree of inheritance and evolution. The topic starts with protein research, breeding and molecular tools, and crop improvement, followed by an evolution into specific protein structures and functions, protein sequence research, and the development of improved crop varieties. Today, it focuses on proteomics research, microbial application studies, and transgenic crop research.

From Figures 7 and 8, it is evident that the evolutionary paths of gene analysis and detection technologies, as well as synthetic biology technologies, have become more diverse and are converging, indicating a process of transition from initial exploration to deeper understanding and application. This trend helps to enhance research efficiency, promote technological innovation, and ensure optimal utilization of limited resources.

3.5 Competitive Landscape Analysis of Key Core Technologies

Patent holder counts and semantic volume are used to represent the technological competitive capabilities of patent holders in these fields, which leads to the compilation of a list of the top ten institutions globally and in China for these two fields, as shown in Tables 7 and 8.

Table 7 Top 10 Institutions Holding Patents in Gene Analysis and Testing Technologies

Global Top 10				China Top 10			
Project Institution	Patent Count	Cumulative Semantic Volume	Average Semantic Volume	Project Institution	Patent Count	Cumulative Semantic Volume	Average Semantic Volume
Monsanto Technology Company	1571	1064	0.677	Huazhong Agricultural University	441	262	0.595
Huazhong Agricultural University	441	262	0.595	China Agricultural University	409	261	0.640

China Agricultural University	409	261	0.640	South China Agricultural University	290	174	0.601
Dow AgroSciences LLC	352	185	0.525	Zhejiang University	240	137	0.571
South China Agricultural University	290	174	0.601	Beijing Academy of Agriculture and Forestry	238	129	0.545
Illumina, Inc.	259	158	0.611	Syngenta (China)	217	109	0.502
Zhejiang University	240	137	0.571	Nanjing Agricultural University	199	107	0.542
Beijing Academy of Agriculture and Forestry	238	129	0.545	Chinese Academy of Agricultural Sciences, Beijing Institute of Animal Husbandry and Veterinary Research	180	115	0.639
Syngenta (China)	217	109	0.502	Northwest A&F University	189	119	0.634
Rijk Zwaan Seed Group	193	115	0.598	Yangzhou University	149	86	0.577

Table 8 Top 10 Institutions Holding Synthetic Biology Patents

Global Top 10				China Top 10			
Project Institution	Patent Count	Cumulative Semantic Volume	Average Semantic Volume	Project Institution	Patent Count	Cumulative Semantic Volume	Average Semantic Volume
Monsanto Technology Company	1736	1212	0.698	Jiangnan University	336	193	0.575
Pioneer	1399	890	0.636	Syngenta (China)	236	137	0.583

International Seed Co.							
MS Technologies LLC	551	454	0.825	Tianjin University	139	76	0.551
BASF SE	444	313	0.705	Nanjing Agricultural University	126	68	0.546
STINE SEED FARM	364	317	0.871	Shanghai Jiao Tong University	104	57	0.551
Dow AgroSciences LLC	496	301	0.607	Chinese Academy of Agricultural Sciences, Biotechnology Research Institute	99	55	0.556
Evogene Ltd.	362	319	0.883	Zhejiang University	99	51	0.521
Jiangnan University	336	193	0.575	Institute of Microbiology, Chinese Academy of Sciences	95	59	0.623
Bayer AG	257	164	0.639	Tianjin University of Science and Technology	77	42	0.546
Syngenta (China)	236	137	0.583	Kunming University of Science and Technology	76	39	0.514

The institutions that appear in the global top 10 for both fields are Monsanto Technology Company, Dow AgroSciences LLC, Pioneer International Seed Co., and Syngenta Group, all of which are among the top 10 global seed industry companies and renowned international leaders. In the field of gene analysis and testing technologies, additional institutions include Huazhong Agricultural University, China Agricultural University, and six other research institutions, as well as Rijk Zwaan Seed Group. In the field of synthetic biology, well-known seed companies like BASF SE and Bayer AG, along with Evogene Ltd., Jiangnan University, and Nanjing Agricultural University, also appear. Notably, Evogene Ltd. (EVOGENE LTD.) uses advanced computational technologies for predictive research in biology, establishing itself in the field through computational biology; Rijk Zwaan Seed Group (RIJK ZWAAN ZAADTEELT EN ZAADHANDEL B.V.) is a leading global vegetable and flower seed breeding company; MS Technologies LLC (MS TECHNOLOGIES, LLC.) focuses on agricultural technologies, offering various agronomic products and solutions; and Illumina, Inc. is a global leader in gene sequencing and microarray technologies, specializing in the development, manufacturing, and sales of systems for large-scale analysis of genetic variation and biological functions.

By comparing data from both domestic and international top 10 institutions, it was found that Chinese top 10 institutions hold a combined total of 2,552 patents in gene analysis and testing technologies and 1,387 patents in synthetic biology, representing 60% and 22% of the patent count of the global top 10, respectively. In terms of cumulative semantic volume, Chinese top 10 institutions hold 1,499 and 777, accounting for 57% and 18% of the global top 10, respectively. For average semantic volume, Chinese top 10 institutions have an average semantic volume of 0.584 and 0.556 in the two key core technology fields, compared to 0.586 and 0.702 for global top 10 institutions. Based on this, it can be concluded that China has a relative advantage in gene analysis and testing technologies, where research is relatively more abundant, whereas synthetic biology technologies still need further development in both quality and quantity. An in-depth exploration of patent sharing and transfer among international leading institutions reveals that, in addition to independent R&D, international seed industry companies also emphasize collaborative R&D and technology transfer, with frequent cooperation between similar companies. Additionally, in the development of these two key core technologies, international entities are predominantly led by companies, whereas Chinese research is mostly driven by universities and research institutes, which may explain the lower quality of patents in China compared to international seed industry companies. Therefore, it is evident that for China to develop the biological breeding field, it must strengthen collaboration and technology transfer between enterprises and complementary technology institutions, particularly with international leading institutions; focus on the integration of informatics and biology to promote scientific research breakthroughs; and enhance the role of enterprises in innovation, leveraging the strengths of both enterprises and research institutes to achieve breakthroughs in key core technologies in the biological breeding field.

4 Conclusion

The integration of the LDA topic model and BERT model enables precise technical topic mining based on patent literature in the field of biological breeding, as well as the analysis of the evolutionary relationships of key core technologies in biological breeding. By embedding the BERT model, the output of topic vector features can more accurately represent the content of the text. The combination of LDA and BERT models takes into account both the semantic features of word vectors and the context of the text. Compared to the single LDA topic model, this approach can better represent the content of the topics. Additionally, by constructing "criticality" based on the "average citation number of patent families," "centrality" based on the "average number of patent family members," and "innovativeness" based on the "average number of claims," this study achieves measurement and analysis of key core technology topics in the biological breeding field. Empirical analysis results show: 1)From the perspective of topic identification, a topic identification method based on the LDA-BERT model was constructed, which identified six technical topics. Through keyword analysis and the indicator system, two key core technology topics were identified: Gene Analysis and Detection Technologies, and Synthetic Biology. 2)From the perspective of topic evolution visualization, Gene Analysis and Detection Technologies rapidly evolved from the demand for improved breeding to precision breeding based on gene analysis, while Synthetic Biology evolved from protein research to the use of synthetic biology technologies to create new varieties with ideal traits based on various protein characteristics. 3)From the perspective of the competitive landscape of key core technologies, China has an advantage in Gene Analysis and Detection Technologies, but compared to seed industry giants like Monsanto, there is a gap in research quality. Furthermore, research and development in China's biological breeding field is mainly led by research institutes, with lower involvement from seed industry companies in innovation, which lags behind international high-level innovation.

References:

[1] Cheng Yu, Ye Xingqing, Ning Xia, et al. The Major "Bottlenecks" and Policy Approaches for China's Technological Self-Reliance in the Seed Industry [J]. *China Rural Economy*, 2022(8): 35-51.

[2] Jiang Yao, Chen Xu, Zhang Lingkai. A Three-Stage Recognition Study of "Bottleneck" Technologies under the Patent Perspective—A Case Study of Chip Materials [J]. *Information Journal*: 1-9.

[3] Chen Xu, Jiang Yao, Xiong Yan, et al. Identification and Analysis of Key Core "Bottleneck" Technologies Based on Patents—A Case Study of the Integrated Circuit Industry [J]. *Information Journal*: 1-8.

[4] Tan Jinsong, Song Juan, Wang Kexin, et al. Breaking Through Key Core "Bottleneck" Technologies from the Perspective of Innovation Ecosystems—A Case Study of China's High-Speed Train Traction Systems [J/OL]. *Nankai Management Review*: 1-28 [2023-12-20].

[5] Xu Xia, Wu Fuxiang, Wang Bing. Research on Key Core Technology Identification Based on International Patent Classification [J/OL]. *Information Journal*, 2022, 41(10): 74-81.

[6] ADOMAVICIUS G, BOCKSTEDT J C, GUPTA A, et al. Technology Roles and Paths of Influence in an Ecosystem Model of Technology Evolution [J/OL]. *Information Technology and Management*, 2007, 8(2): 185-202.

[7] Yang Wu, Yang Dafei. Research on Industry Core Technology Identification Based on Patent Data—A Case Study of the 5G Mobile Communication Industry [J]. *Information Journal*, 2019, 38(3): 39-45+52.

[8] Yu Jiang, Chen Feng, Zhang Yue, et al. Forging Strong National Key Technologies: Exploration of the Laws and Systematic Construction of Key Core Technology Breakthroughs [J]. *Bulletin of the Chinese Academy of Sciences*, 2019, 34(03): 339-343.

[9] SONG K, KIM K, LEE S. Identifying Promising Technologies Using Patents: A Retrospective Feature Analysis and a Prospective Needs Analysis on Outlier Patents [J/OL]. *Technological Forecasting and Social Change*, 2018, 128: 118-132.

[10] Zhang Yuchen, Tan Li. Concept Definition, Feature Analysis, and Breakthrough Paths of Key Core Technologies [J/OL]. *China Science and Technology Forum*, 2023(2): 20-29.

[11] Frishamma J, Ericsson K, Patel P C. The Dark Side of Knowledge Transfer: Exploring Knowledge Leakage in Joint R&D Projects [J]. *Technovation*, 2015, 41/42: 75-88.

[12] Huang Lucheng, Liu Chunwen, Wu Feifei, et al. Core Technology Identification Model Based on NPCIA and Application Research [J]. *Science Research Management*, 2020, 38(11): 1998-2007.

[13] Jia Qian, Zheng Huaiguo, Zhao Jingjuan. Patent Layout of Multinational Seed Companies in Crop Breeding and Implications for China [J/OL]. *China Biotechnology Journal*, 2022, 42(10): 112-124.

[14] Wu Huabin, Xu Qingrui, Li Yang. Cultivating and Enhancing Enterprise Core Capabilities under Innovation Leadership—A Vertical Case Study of Haier Group [J]. *Nankai Management Review*, 2019, 22(5): 28-37.

[15] Chen Xu, Jiang Yao, Xiong Yan, et al. Identification and Application of Key Core “Bottleneck” Problems: A Case Study of AI Chips [J]. *China Science and Technology Forum*, 2023(09): 17-27.

[16] HU R, MA W, LIN W, et al. Technology Topic Identification and Trend Prediction of New Energy Vehicles Using LDA Modeling [J/OL]. *Complexity*, 2022, 2022: 1-20.

[17] Wang Xiuhong, Wang Xin, Wang Shaofan, et al. A Disruptive Technology Identification Method Based on SimCSE-LDA and Anomaly Detection—A Case Study of Agricultural Robots [J/OL]. *Information Theory and Practice*, 2023, 46(5): 135-143.

[18] Lü Kun, Chen Xiaoyu, Jing Jipeng. Key Technology Identification Research of Blockchain Financial Industry Based on Combined Word Segmentation and LDA Model [J/OL]. *Library and Information Work*, 2022, 66(19): 110-121.

[19] Chen Yuxin, Lu Jun, Han Yi. Research on Disruptive Technology Identification Based on Patent Literature—A Case Study of Artificial Intelligence [J]. *Journal of Information Science*, 2022, 41(11): 1124-1133.

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

[21] Hu Kai, Xie Fen, Yang Binyu, et al. Identification and Application Research on Key Common Technologies in Industries Based on Patent Text Mining [J]. *Technology Management Research*, 2023, 43(8): 21-31.

[22] Cui Zunkang, Li Danyang, Xu Xiaoting, et al. Global Innovation Layout and Competitive Landscape of Food Crop Biological Breeding Technologies—Based on Core Patent Data Mining [J]. *China Agricultural Science and Technology Bulletin*, 2022, 24(05): 1-14.

[23] Tang Zhiwei, Li Yuxuan, Zhang Longpeng. Identification Methods and Breakthrough Paths for "Bottleneck" Technologies in the Context of Sino-US Trade Frictions—A Case Study of the Electronic Information Industry [J]. *Science and Technology Progress and Countermeasures*, 2021, 38(1): 1-9.

[24] Jiang Yao, Chen Xu, Hu Bin. Two-Stage Funnel Selection Model for "Bottleneck" Key Core Technologies and Application Research [J]. *Information Journal*, 2023, 42(3): 94-101.

[25] LI X, FAN M, ZHOU Y, et al. Monitoring and Forecasting the Development Trends of Nano Generator Technology Using Citation Analysis and Text Mining [J]. *Nano Energy*, 2020, 71: 104636.

[26] Liu Ziqiang, Xu Haiyun, Yue Lixin, et al. Core Technology Topic Identification Method Based on Chunk-LDAvis and Application Research [J/OL]. *Library and Information Work*, 2019, 63(9): 73-84.

[27] Yang Dafei, Yang Wu, Tian Xuejiao, et al. Core Technology Identification Model Construction and Empirical Study Based on Patent Data [J]. *Information Journal*, 2021, 40(2): 47-54.

[28] ALTUNTAS S, DERELI T, KUSIAK A. Forecasting Technology Success Based on Patent Data [J]. *Technological Forecasting and Social Change*, 2015, 96(7): 202-214.

[29] Mao Jianqi, Du Yanting, Miao Chenglin, et al. Key Core Technology Identification Model Construction and Application Based on Patent Co-classification—A Case Study of Photolithography Technology [J/OL]. *Information Journal*, 2022, 41(11): 48-54.

[30] Yang Yanping, Dong Yu, Han Tao. Research on Industry Key Technology Identification Methods Based on Patent Co-citation Clustering and Combined Analysis—A Case Study of Crop Breeding Technology [J/OL]. *Library and Information Work*, 2016, 60(19): 143-148+124.

[31] Hu Xubo, Yuan Changhong. Key Core Technologies: Concepts, Features, and Breakthrough Factors [J/OL]. *Science Research Management*, 2022, 40(1): 4-11.

[32] Luo Jian, Cai Lijun, Shi Min. Two-Stage Emerging Technology Identification Research Based on Patents—A Case Study of Image Recognition Technology [J]. *Information Science*, 2019, 37(12): 57-62.

[33] KWON O J, SEO J, NOH K R, et al. Categorizing Influential Patents Using Bibliometric Analysis of Patent Citations Network [J]. *Information—An International Interdisciplinary Journal*, 2007, 10(3): 313-326.