ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-202401.00317

Hotspot Interdisciplinary Topic Identification and Thematic Evolution Analysis in Genetic Engineering: An Interdisciplinary Perspective

Authors: Zhu Shiqin, Fan Dandan, Guo Tianyu, Zhu Shiqin

Date: 2024-01-26T00:00:00+00:00

Abstract

To more accurately capture research hotspots and development trends in interdisciplinary research, this study proposes a computational method for topic interdisciplinary degree and combines it with topic intensity to comprehensively identify hot interdisciplinary topics and predict their future development. This study conducts an empirical analysis on genetic engineering papers from the Web of Science database covering 2000-2019. First, the LDA model is adopted for topic mining; then hot interdisciplinary topics are identified through the calculation of topic intensity and topic interdisciplinary degree; finally, time windows are divided, trend charts of topic intensity and topic interdisciplinary degree are plotted, and the results are analyzed. The empirical results demonstrate that: the field of genetic engineering comprises 21 important topics, including 7 hot topics, 14 interdisciplinary topics, and 2 hot interdisciplinary topics; according to the changing trend of topic intensity, the 21 topics are classified into 3 rising topics, 7 declining topics, and 11 stable topics, with most topics showing an increasing trend in their degree of interdisciplinary integration.

Full Text

Preamble

Theme Evolution Analysis and Recognition of Hot Interdisciplinary Themes in Genetic Engineering from an Interdisciplinary Perspective

Zhu Shiqin, Fan Dandan, Guo Tianyu
East China University of Science and Technology, Shanghai, 200237

Abstract: To more accurately grasp research hotspots and development trends in interdisciplinary research, this study proposes an integrated approach based

on theme intensity and theme interdisciplinary degree to identify hot interdisciplinary themes and predict their future development. We selected papers in the field of genetic engineering from the Web of Science database (2000-2019) for empirical analysis. First, the LDA model was employed to mine themes. Then, hot interdisciplinary themes were identified by calculating theme intensity and interdisciplinary degree. Finally, time windows were divided to plot variation trends of theme intensity and interdisciplinary degree, followed by result analysis. Empirical results show that there are 21 important themes in genetic engineering, including 7 hot themes, 14 interdisciplinary themes, and 2 hot interdisciplinary themes. According to variation trends in theme intensity, the 21 themes are classified into 3 ascending themes, 7 descending themes, and 11 stable themes, with most themes showing an increasing trend in interdisciplinary degree.

Keywords: Interdisciplinary theme; Hot theme; Theme recognition; Theme evolution

Introduction

In his 1986 Nobel Foundation award ceremony speech, the chairman stated that old disciplinary boundaries between physics and chemistry, and between biology and medicine, had been broken through in all aspects. These fields not only intersect with each other but have also formed a continuum without distinct boundaries. In recent years, a series of scientific discoveries and technological innovation achievements have been widely distributed across interdisciplinary fields such as molecular biology, physical chemistry, and systems science. It can be foreseen that driven by major national strategic needs, multidisciplinary convergence and cross-boundary integration of technologies will become the norm, continuously spawning new disciplinary frontiers, new technological fields, and new innovation paradigms [1].

Interdisciplinary themes are common research themes formed through the fusion and penetration of two or more disciplines. They represent convergence points for knowledge integration [2], hubs for knowledge diffusion [3], and breakthrough points for achieving scientific and technological innovation [4]. Hot themes are important thematic areas that researchers maintain high attention to and conduct extensive research on during a certain period. Using scientific methods to identify research hotspots and their evolution trends can help researchers correctly grasp current dynamics, gain perspective on disciplinary development and academic progress, and clarify disciplinary priorities [5]. Currently, interdisciplinary research is mainly divided into macro and micro aspects [6]. This study adopts a micro perspective, taking the field of genetic engineering as an example, extracting themes through the LDA model, proposing a measurement method for theme interdisciplinary degree, and combining it with theme intensity indicators to identify hot interdisciplinary themes. Quantitative analysis

of variation trend characteristics of theme intensity and interdisciplinary degree in the field helps grasp interdisciplinary development trends and discover innovative research directions and themes.

1.1 Research on Interdisciplinary Theme Identification

Interdisciplinary theme identification typically employs citation analysis, lexical analysis, and topic modeling methods.

(1) Citation Analysis Method. Citation analysis usually analyzes citation phenomena among various analytical objects such as journals, documents, themes, and authors based on citation relationships. In the field of interdisciplinary theme identification, citation analysis can also be used to identify cross-disciplinary themes. Chi R et al. [7] discovered the development of major research themes and their relationships based on co-citation network analysis. Adams J et al. [8] identified thematic content and determined theme clustering by constructing a bibliographic coupling network in AIDS research. P. Vugteveen et al. [9] mapped the disciplinary landscape and knowledge flow of the interdisciplinary field of river science based on journal citation relationships, and obtained research themes by clustering papers according to the similarity of cited references and associating them with major disciplines, but did not consider theme interdisciplinarity.

(2) Lexical Analysis Method. Lexical analysis uses words in documents as analysis objects, primarily employing word frequency statistics and co-word analysis to study hot themes. Xu et al. [10] took informatics as an example to determine interdisciplinary themes by calculating TI values, Bet values, and word frequency values, combining social network analysis with time series analysis to analyze the evolution of interdisciplinarity. Scholar Du Lijun [11] studied papers related to information retrieval in informatics and computer science, identifying cross-disciplinary research themes between the two disciplines through co-word matrix construction. In the field of genetic engineering vaccines, scholars Wei Ling [12] et al. used co-word analysis methods, employing patent co-word clustering and strategic diagrams to identify technical themes and their development status. Luo Rui [13] used subject term co-occurrence networks to represent knowledge networks and measured the state of knowledge networks using structural entropy to further identify scientific breakthroughs.

(3) Topic Modeling Method. Topic models applied to interdisciplinary theme identification mainly include the CTM model, AT model, and LDA model. The core computational problem of topic models is using visible documents to infer their latent thematic structure [14]. Latent Dirichlet Allocation (LDA), proposed by D. Blei [15] in 2003, is a common topic model widely applied in theme identification. For example, Zhang Bin [16] explored the formation of mixed disciplinary research themes from a clustering perspective using the LDA model. Chen Qiong et al. [17] used the LDA model to identify and divide themes in the field of medical informatics, then introduced DIV measurement indica-

tors to compare interdisciplinary situations. Han Zhengqi et al. [18] used the Rao-Stirling index to identify highly interdisciplinary literature and employed the LDA model to obtain research themes of highly interdisciplinary literature in nanotechnology, but did not consider theme interdisciplinary degree.

1.2 Research on Interdisciplinary Trend Evolution

Interdisciplinary research has attracted widespread attention and discussion in academic circles both domestically and internationally, with interdisciplinary trend evolution research in a vigorous development stage. Research objects are mainly journals and disciplinary fields.

(1) Journals as Research Objects. Silva [19], Leydesdorff [20], and other scholars measured differences between disciplines by constructing citation networks. Meng Xiangbao [21] studied core journals of foreign library and information science to understand the current status of interdisciplinary integration in foreign library and information science, discovering its knowledge sources and applications. R. Agarwal [22] and Yang Ruixian [23] both took journals as research objects, studying from the perspectives of references and citations. The former confirmed that the disciplinary boundaries of information systems are constantly expanding, while the latter studied the cross-disciplinary integration between library and information science and other disciplines.

(2) Disciplinary Fields as Research Objects. Carley and Porter [24] used Rao diversity as a measure of interdisciplinarity and analyzed citation patterns of paper collections in six thematic categories. The study found that mathematics had low interdisciplinarity while medicine had high interdisciplinarity, revealing trends of integration between disciplines. Levitt et al. [25] analyzed evolution among disciplines in Social Sciences Citation Index (SSCI) categories in three specific years (1980, 1990, and 2000). Cao Jiajun et al. [26] took the artificial intelligence field as the analysis object, revealing the distribution of core disciplinary categories within the field, and obtained associations and evolution between disciplines by calculating disciplinary similarity to understand development trends of various disciplines in AI. Deng and Xia [27] used social network analysis and disciplinary diversity measurement methods to find that disciplinary distribution in information behavior research was unbalanced.

In summary, scholars have conducted substantial and fruitful research on interdisciplinary theme identification and interdisciplinary trend evolution from perspectives of disciplines and journals. Using topic modeling for theme identification can overcome the lag of citation analysis and the defect of traditional co-word analysis that cannot reflect semantic associations between word pairs. Topic modeling has been applied in disciplinary fields such as medical informatics and nanotechnology, fully proving its feasibility in interdisciplinary fields, but no research has applied it to genetic engineering. Therefore, this study selects the LDA model that can analyze latent semantic information to extract research themes, combines theme intensity with the proposed calculation method

for theme interdisciplinary degree to identify hot interdisciplinary themes, and expands the research objects of interdisciplinary trend evolution to the genetic engineering field to explore change trends of cross-disciplinary themes.

2.1 Research Framework

To identify hot interdisciplinary themes in genetic engineering and conduct theme evolution analysis, the research framework design is proposed as shown in Figure 1 [Figure 1: see original paper].

First, obtain the genetic engineering paper collection from Web of Science, perform operations such as deduplication, deletion of missing values, word frequency statistics, and stop word removal, and use Python natural language processing to extract author keywords (DE) and extended keywords (ID) from documents as the corpus source for theme identification research. Second, use the LDA topic model for theme mining, calculate theme intensity and theme interdisciplinary degree and determine thresholds, and identify hot interdisciplinary themes based on both thresholds. Finally, conduct theme classification and evolution trend presentation from two aspects of theme intensity and theme interdisciplinary degree, and make reasonable analysis of theme development trends by combining disciplinary categories that contribute to theme development.

2.2.1 LDA Model-Based Hot Interdisciplinary Theme Identification Method

This study adopts the LDA topic model for genetic engineering theme research with obvious advantages: facing the massive data of genetic engineering literature, the LDA topic model demonstrates powerful text processing capabilities, enabling computer language-based mining of author keywords and extended keywords to extract more expressive feature words and more accurately mine genetic engineering themes; the greatest advantage of the LDA topic model is combining theme mining with theme evolution, which can analyze theme evolution trends while obtaining themes to grasp research directions in the field.

(1) LDA Model-Based Theme Mining. First, create a word dictionary for the preprocessed corpus, assign an index to each individual word, use the created dictionary to transform the document list into a matrix; second, use the Gensim model to establish an LDA model object, obtain the optimal number of topics K according to calculated perplexity, then run and train the LDA model on the matrix, output the topic-word probability distribution matrix with words output in descending order of frequency, select the top 10 vocabulary items with highest probability under each topic to represent the topic, and identify the topic by combining other output vocabulary. During this process, generate the document-topic probability distribution matrix and the topic-term probability distribution matrix.

(2) Theme Intensity Measurement. Hot theme mining and theme evolution can be measured by calculating theme intensity, which reflects the importance and attention level of a theme. It mines hot themes by comparing theme intensities of different themes under the same time window, and reveals theme evolution characteristics and trends by analyzing theme intensity changes of the same theme across continuous, different time windows. Theme intensity is calculated through vocabulary probability distribution extracted from context under the LDA topic model.

Theme intensity is mainly obtained through the constructed document-topic probability distribution matrix to get the probability of each topic generated by each document.

$$\theta_{tz} = \frac{1}{D_t} \sum_{d=1}^{D_t} \theta_{dz}$$

where θ_{tz} represents theme intensity in time period t, obtained by taking the average of topic posterior probabilities; θ_{dz} represents the proportion of theme z in document d, and D_t represents the number of documents in time period t.

After calculating theme intensity for each theme, a threshold is determined to filter out themes with higher attention. Regarding theme intensity threshold determination, this study adopts the calculation method proposed by Wu Chake et al. [28], with the formula:

$$T = \frac{1}{K} \sum_{z=1}^{K} \theta_{tz}$$

where T is the theme intensity threshold, K represents the number of themes, and D_t represents the text collection. When theme intensity is greater than threshold T, the theme can be judged as a hot theme in the current time window.

(3) Theme Interdisciplinary Degree Measurement. The theme interdisciplinary degree is measured according to the constructed formula. The main idea is to obtain the document set contained under each theme after getting the document-topic probability distribution matrix, and calculate the interdisciplinary degree of each document. This study adopts the Rao-Stirling index as a comprehensive measurement indicator for document interdisciplinary degree, referred to as R in this paper. The R indicator comprehensively measures the interdisciplinary degree of a single paper from diversity, balance, and disparity. If the references of a paper belong to very similar disciplinary categories, the paper has a low interdisciplinary degree; conversely, it is higher.

$$R = \sum_{i,j} p_x p_j S_{i,j}$$

where p_x represents the proportion of cited frequency of discipline i to the total cited frequency of all disciplines, and $S_{i,j}$ represents the similarity between discipline x and discipline j in the discipline similarity matrix.

This study proposes a calculation method for theme interdisciplinary degree, determining the theme's interdisciplinary degree based on the mean of interdisciplinary degrees of all documents under a theme.

$$R_t = \frac{1}{m} \sum_{i=1}^{m} R_i$$

where R_t represents the interdisciplinary degree of theme t, m represents the number of documents contained in the theme, and R_i is the interdisciplinary degree of document i.

After calculating each theme's interdisciplinary degree, a threshold is determined to filter out themes with higher interdisciplinary degrees, with the formula:

$$I = \frac{1}{K} \sum_{t=1}^{K} R_t$$

where I is the theme interdisciplinary degree threshold, K represents the number of themes, and R_t represents the interdisciplinary degree of theme t.

2.2.2 Theme Evolution Trend Analysis Method

Theme evolution trend analysis includes theme intensity variation trend analysis and theme interdisciplinary degree variation trend analysis. Existing research [29] summarizes three different evolution methods according to different ways of introducing time: Joint method, pre-discretization analysis method, and post-discretization analysis method. This study adopts post-discretized analysis. This method first ignores time, takes the entire text collection as the analysis text, obtains the topic-term probability distribution matrix and document-topic probability distribution matrix through the LDA topic model, discretizes documents into each time window according to their publication year; finally, calculates theme intensity of each theme in continuous time windows sequentially through formula (1), and classifies themes into categories based on rising and falling intensity trends.

3. Empirical Study

To mine hot interdisciplinary themes in genetic engineering and conduct theme evolution analysis, this study calculated and obtained variation trends of theme intensity and interdisciplinary degree for each theme, and carried out empirical research.

3.1 Data Collection and Processing

Research data were sourced from the genetic engineering field in Web of Science, with the retrieval strategy shown in Table 1 . After deduplication and deletion of invalid content, a total of 51,954 documents were obtained.

Table 1: Literature Retrieval Strategy

TI=("gene* engineering" or "DNA engineering" or "gene* manipulat*" or "DNA manipulat*" or "gene* recombinat*" or "transgen*" or "gene* clon*" or "molecular clon*") or AK=("gene* engineering" or "DNA engineering" or "gene* manipulat*" or "DNA manipulat*" or "gene* recombinat*" or "transgen*" or "gene* clon*" or "molecular clon*")

Source Database: SCI-Expanded Database Article

Second, Python natural language processing was used to extract author keywords (DE), extended keywords (ID), disciplines, publication year, title, abstract, etc. from documents as texts for analysis. Third, word frequency statistics were performed on the analysis texts, high-frequency words without discriminative power and meaningless interference words were deleted according to statistical results, synonyms were merged, stop words were removed, and words with different parts of speech were lemmatized.

3.2 Hot Interdisciplinary Theme Identification

(1) Theme Extraction. This study used the Gensim model to establish an LDA model object and obtained the optimal number of topics K according to calculated perplexity. The perplexity distribution curve under different topic numbers from 2000 to 2019 is shown in Figure 2. When the perplexity value fluctuates to a gentle state at a relatively small value, or when a relatively obvious inflection point appears, the inflection point represents the best fitting degree of the topic model and the optimal theme extraction effect. Therefore, K was determined to be 21.

Figure 2 [Figure 2: see original paper]: Perplexity Distribution Curve of Genetic Engineering Under Different Topic Numbers

The LDA model was run and trained on the corpus, generating the document-topic probability distribution matrix and topic-term probability distribution matrix during this process. The topic-term probability distribution matrix was output, with characteristic words under each theme sorted from largest to smallest according to their definition degree for the theme. The higher the distribution probability under the theme, the more forward its position. The top 10 keywords with highest probability under each theme were selected to represent the theme, with partial themes shown in Figure 3 [Figure 3: see original paper]. This study summarized a total of 21 themes.

Figure 3: Topic-Term Distribution in Genetic Engineering Field (Partial)

(2) Theme Intensity and Interdisciplinary Degree Measurement. By calculating theme intensity and theme interdisciplinary degree, this study conducted comparative analysis of each theme's intensity and interdisciplinary degree from 2000 to 2019, as shown in Table 2 .

Table 2: Theme Intensity and Interdisciplinary Degree of Themes in Genetic Engineering Field

No.	Theme	Interdisciplinary Degree
T1	Transcriptome Sequencing Technology	
T2	Transgenic Drugs	
T3	Transgenic Crops	
T4	Gene Cloning Technology	
T5	Alzheimer's Disease	
T6	Tumor	
T7	Vaccine	
T8	Synaptic Plasticity	
T9	Atherosclerotic Disease	
T10	Frontotemporal Dementia	
T11	Plant Insect Resistance	
T12	Transgenic Animals	
T13	Gene Therapy	
T14	Biological Inheritance	
T15	Plant Disease Resistance	
T16	Phytoremediation Technology	
T17	Apoptosis	
T18	Amyotrophic Lateral Sclerosis	
T19	Biodiversity Conservation	
T20	Organism Immune Response	
T21	Abiotic Stress	

The theme intensity threshold for the genetic engineering field was calculated as 0.0476, yielding 7 hot themes: "Transcriptome Sequencing Technology," "Gene Cloning Technology," "Alzheimer's Disease," "Transgenic Animal Research," "Plant Disease Resistance," "Apoptosis," and "Abiotic Stress." The theme interdisciplinary degree threshold was calculated as 0.3995, yielding 14 interdisciplinary themes: "Transgenic Drugs," "Transgenic Crops," "Alzheimer's Disease," "Tumor," "Vaccine," "Synaptic Plasticity," "Atherosclerotic Disease," "Frontotemporal Dementia," "Plant Insect Resistance," "Gene Therapy," "Biological Inheritance," "Apoptosis," "Amyotrophic Lateral Sclerosis," and "Organism Immune Response."

Among these, themes with both theme intensity and theme interdisciplinary degree exceeding thresholds were "Apoptosis" and "Alzheimer's Disease." They are both interdisciplinary themes and hot themes—hot interdisciplinary themes—as shown in Figure 4 [Figure 4: see original paper].

3.3 Theme Evolution and Analysis

Figure 4: Distribution of Theme Intensity and Interdisciplinary Degree

Taking year as the unit, the publication year of documents was obtained, and theme intensity values and interdisciplinary degree values of each theme in continuous time windows were calculated. Variation trend charts of genetic engineering theme intensity and theme interdisciplinary degree were drawn respectively. By observing changes in theme intensity and interdisciplinary degree of each theme across continuous time windows, theme evolution characteristics were summarized, and each theme was categorized into ascending themes, descending themes, and stable themes.

Figure 5 [Figure 5: see original paper]: Variation Trend of Theme Intensity and Interdisciplinary Degree in Genetic Engineering

Themes "T5 Alzheimer's Disease" and "T17 Apoptosis" are both interdisciplinary themes and hot themes. As shown in Figure 5, except for "T21 Abiotic Stress," the interdisciplinary degree of other themes basically shows an upward trend. In genetic engineering research, themes are becoming increasingly interdisciplinary, with cross-disciplinary collaboration becoming more evident. The intensity of 3 themes is continuously increasing, 7 themes show a declining trend, and the remaining 11 themes show small intensity changes with a stable trend. By classifying themes into ascending, descending, and stable types based on theme intensity variation trends, and combining theme interdisciplinary degree, key themes are analyzed.

(1) Ascending Themes. Figure 5 shows there are 3 ascending themes: T5 Alzheimer's Disease, T11 Plant Insect Resistance, and T21 Abiotic Stress. Among them, "T21 Abiotic Stress" is a hot research theme, and "T5 Alzheimer's Disease" is a hot interdisciplinary theme. The hot interdisciplinary theme "T5 Alzheimer's Disease" is selected for focused analysis:

Over the 20-year period, the theme intensity of "T5 Alzheimer's Disease" has remained high, showing an overall upward trend, indicating that this theme is still a research hotspot receiving high attention from scholars. Although theme intensity declined in 2004, 2009, and 2010, the decline was not significant and rebounded promptly the following year, not affecting the overall upward trend. The interdisciplinary degree of this theme also shows a stable upward trend, indicating that more and more disciplines are participating in Alzheimer's disease research. Combined with specific discipline analysis, Neurosciences and Biochemistry & Molecular Biology have made significant contributions to this theme. Current Alzheimer's disease research mainly includes: pathogenesis, risk factors, diagnosis, and treatment. The complexity and comprehensiveness of these research questions require breaking disciplinary barriers for multidisciplinary collaborative research. It can be reasonably predicted that as the aging population gradually increases, the theme intensity of Alzheimer's disease will continue to rise, and its research heat will inevitably drive participation from

more disciplines.

(2) Descending Themes. Between 2000 and 2019, 7 descending themes include: T1 Transcriptome Sequencing Technology, T3 Transgenic Crops, T7 Vaccine, T9 Atherosclerotic Disease, T10 Frontotemporal Dementia, T13 Gene Therapy, and T19 Biodiversity Conservation. Among them, "T1 Transcriptome Sequencing Technology" is a hot theme, while "T3 Transgenic Crops," "T7 Vaccine," "T9 Atherosclerotic Disease," "T10 Frontotemporal Dementia," and "T13 Gene Therapy" are interdisciplinary themes. The interdisciplinary theme "T7 Vaccine" is selected for focused analysis:

In the history of human struggle against diseases, vaccination is an effective means to eliminate and control infectious diseases. With the development of science and technology, vaccines for various diseases have been developed, such as hepatitis B vaccine and rabies vaccine, and their development technologies are constantly improving. The theme intensity of vaccines peaked in 2001, then showed a fluctuating downward trend, but showed an upward trend with an obvious inflection point in 2017. This shows that although vaccine research heat shows a downward trend, it rebounded after 2017. The interdisciplinary degree of this theme shows an upward trend, becoming an interdisciplinary theme in 2005 with an interdisciplinary degree of 0.4095. Scholars from Immunology and Biochemistry & Molecular Biology have been focusing on vaccine research. Combined with the novel coronavirus that emerged at the end of 2019 and caused a pandemic, countries around the world increased investment in funds, personnel, and other aspects. Through interdisciplinary and cross-boundary integration, transformative technologies effectively promoted vaccine research and development, greatly shortening the vaccine development cycle and successfully developing multiple types of COVID-19 vaccines. Both theme intensity and theme interdisciplinary degree of this theme increased in 2019. It can be predicted that with people's awakening health consciousness of "prevention first" and the occurrence of public health events, vaccine research heat will continue to rise in the future and is expected to become a research hotspot.

(3) Stable Themes. There are 11 stable themes: T2 Transgenic Drugs, T4 Gene Cloning Technology, T6 Tumor, T8 Synaptic Plasticity, T12 Transgenic Animals, T14 Biological Inheritance, T15 Plant Disease Resistance, T16 Phytoremediation Technology, T17 Apoptosis, T18 Amyotrophic Lateral Sclerosis, and T20 Organism Immune Response. Among them, "T4 Gene Cloning Technology," "T15 Plant Disease Resistance," and "T12 Transgenic Animals" are hot themes; "T2 Transgenic Drugs," "T6 Tumor," "T8 Synaptic Plasticity," "T11 Plant Insect Resistance," "T14 Biological Inheritance," "T18 Amyotrophic Lateral Sclerosis," and "T20 Organism Immune Response" are interdisciplinary themes; and "T17 Apoptosis" is a hot interdisciplinary theme. The hot interdisciplinary theme "T17 Apoptosis" is selected for focused analysis:

Apoptosis refers to the process of active cell death controlled by genes after normal cells undergo physiological or pathological stimulation. The theme intensity of apoptosis has remained at a high level with a stable trend; the theme

interdisciplinary degree is stable with a slight increase, meaning this theme has always been a focus of scholars and is applied in multiple fields. Since the entire cell generates protrusions containing organelles, nuclei, and cytoplasmic apoptotic fragments during apoptosis, which are then phagocytosed by other cells, scholars have used this characteristic to further understand cells in organisms, bringing brand-new research directions to medicine, development, animal husbandry, and other fields. This shows that apoptosis research is widely applied across disciplinary fields, helping multiple disciplines overcome challenges. However, to date, apoptosis detection methods and pathways have not been thoroughly studied. To more thoroughly understand apoptosis mechanisms and more effectively treat diseases, scholars from related disciplines will continue to pay attention to this theme and attempt more in-depth research on apoptosis.

Conclusion

This study mined and extracted theme content in the genetic engineering field, identified 21 important themes through the topic-term probability distribution generated by the LDA topic model, then identified hot interdisciplinary themes in genetic engineering based on theme intensity and theme interdisciplinary degree thresholds, and conducted theme analysis. Document themes were divided by time, and by calculating intensity and interdisciplinary degree of each theme in different and continuous time windows, variation trend charts of genetic engineering research theme intensity and interdisciplinary degree were drawn to obtain theme evolution trends. This study can draw the following conclusions:

(1) **Identify hot interdisciplinary themes in disciplinary fields.** Using the LDA model can quickly obtain and identify 21 key themes in genetic engineering. Meanwhile, combining the Rao-Stirling index to measure theme interdisciplinary degree in this field can quickly discover 2 hot interdisciplinary themes in genetic engineering from massive literature. This method also has applicability for hot interdisciplinary theme research in other fields.

(2) **Theme intensity has dynamic variability.** From the perspective of theme intensity, hot research themes in genetic engineering include: "Transcriptome Sequencing Technology," "Gene Cloning Technology," "Alzheimer's Disease," "Transgenic Animal Research," "Plant Disease Resistance," "Apoptosis," and "Abiotic Stress." Influenced by continuous development of biology and information technology or factors such as sudden diseases and politics, theme popularity in genetic engineering changes across different time windows, with related scholars increasing or decreasing attention to themes accordingly.

(3) **Interdisciplinary integration degree in genetic engineering further deepens.** From the perspective of interdisciplinary degree, most themes show an upward trend in interdisciplinary degree, indicating that genetic engineering currently attaches increasing importance to interdisci-

plinary research. Among them, the theme intensity and interdisciplinary degree of "Alzheimer's Disease" continue to rise, with more and more disciplines investing in research, allowing for further in-depth study; the theme intensity and interdisciplinary degree of "Apoptosis" remain stably at high levels, continuously attracting attention from numerous scholars as it reveals the mysteries of life at the microscopic level.

This study still has certain limitations. When obtaining the optimal number of topics based on perplexity calculation results, there is a risk of low theme discriminability due to excessive topic numbers; using the LDA topic model to obtain the top ten terms under each topic based on probability, the naming results after summarization cannot completely cover all content under the theme, resulting in certain deviations; analysis of genetic engineering themes is limited by time and professional expertise, and future research can deepen analysis of hot interdisciplinary themes through literature reading and expert interviews.

References

[1] Wu Chaohui, Zhao Ena. Serving National Strategic Needs through Interdisciplinary Integration [N]. People's Daily, 2020-11-04(012).

[2] Jia Xiali. Research on Interdisciplinary Knowledge Integration Based on Theme Co-occurrence [D]. Beijing: University of Chinese Academy of Sciences (National Science Library, Chinese Academy of Sciences), 2022.

[3] Zhu Shiqin, Fan Dandan, Gou Wenjing. Correlation Study between Interdisciplinarity and Knowledge Diffusion: A Case Study of Genetic Engineering Field [J]. Modern Information, 2022, 42(09):121-131.

[4] Li Chunjing, Liu Zhonglin. Analysis of Interdisciplinary Models in Modern Scientific Development: An Analytical Framework for Interdisciplinary Models [J]. Science Research, 2004(03):244-248.

[5] Ma Feicheng. Focusing on Disciplinary Hotspots and Perspective Academic Progress [J]. Information and Documentation Services, 2022, 43(01):13-14+22.

[6] Xu Haiyun, Yin Chunxiao, Guo Ting, et al. Review of Interdisciplinary Research [J]. Library and Information Service, 2015, 59(05):119-127.

[7] Chi R, Young J. The interdisciplinary Structure of Research on intercultural Relations: A Co-citation Network Analysis Study [J]. Scientometrics, 2013, 96(1).

[8] Adams J, Light R. Mapping Interdisciplinary Fields: Efficiencies, Gaps and Redundancies in HIV/AIDS Research [J]. Plos One, 2014, 9(12):e115092.

[9] Vugteveen P, Lenders R, Van den Besselaar P. The dynamics of interdisciplinary research fields: the case of river research [J]. Scientometrics, 2014, 100(1):73-96.

[10] Xu H, Guo T, Yue Z, et al. Interdisciplinary topics of information science: a study based on the terms interdisciplinarity index series [J]. Scientometrics, 2016, 106(5):583-601.

[11] Du Lijun. Evolution Analysis of Information Retrieval Research Themes from an Interdisciplinary Perspective: A Case Study of Informatics and Computer Science [J]. Information Technology and Informatization, 2020(01):178-183.

[12] Wei Ling, Xu Haiyun, Liu Chunjiang, et al. Research on Theme Discovery in Technical Fields: A Case Study of Genetic Engineering Vaccine Field [J]. Digital Library Forum, 2017(01):37-45.

[13] Luo Rui, Xu Haiyun, Liu Yahui. Scientific Breakthrough Theme Identification Based on Structural Entropy: A Case Study of Genetic Engineering Vaccine Field [J]. Information Studies: Theory & Application, 2021, 44(05):106-114+99.

[14] Blei D M. Probabilistic topic models [J]. Communications of the ACM, 2012, 55(4):77-84.

[15] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. Journal of machine Learning research, 2003(03):993-1022.

[16] Zhang Bin. Interdisciplinary Theme Exploration: From a Theme Clustering Perspective [J]. Information Science, 2020, 38(10):49-55.

[17] Chen Qiong, Zhu Qinghua, Min Hua, et al. Research on Interdisciplinary Feature Identification Method Based on Domain Themes: A Case Study of Medical Informatics [J]. Modern Information, 2022, 42(04):11-24.

[18] Han Zhengqi, Liu Xiaoping, Kou Jingjing. Domain Interdisciplinary Theme Identification Based on Rao-Stirling Index and LDA Model: A Case Study of Nanotechnology [J]. Information Science, 2020, 38(02):116-124.

[19] Silva F N, Rodrigues F A, Oliveira O N, et al. Quantifying the interdisciplinarity of scientific journals and fields [J]. Journal of Informetrics, 2013, 7(2):469-477.

[20] Leydesdorff L, de Moya-Anegón, F, Guerrero-Bote, V P. Journal Maps, Interactive Overlays, and the Measurement of Interdisciplinarity on the Basis of Scopus Data (1996-2012) [J]. Journal of the Association for Information Science & Technology, 2013, 66(5):1001-1016.

[21] Meng Xiangbao. Interdisciplinary Integration and Development of Library and Information Science: Citation Analysis Based on 35 Foreign Core Journals [J]. Library and Information Knowledge, 2012, (5):50-58.

[22] Agarwal R. On the intellectual structure and evolution of ISR [J]. Information systems research, 2016, 27(3):471-477.

[23] Yang Ruixian, Jiang Xiaohan. Knowledge Structure and Evolution of Interdisciplinary from Citation Perspectives of Disciplines and Journals: An Em-

pirical Study of Library and Information Science [J]. Library and Information Service, 2018, 62(05):30-39.

[24] Carley, S, Porter, A L. A forward diversity index [J]. Scientometrics. 2012, 90(2), 407-427.

[25] Levitt J M, Thelwall M, Oppenheim C. Variations between subjects in the extent to which the social sciences have become more interdisciplinary [J]. Journal of the Association for Information Science & Technology, 2011, 62(6):1118-1129.

[26] Cao Jiajun, Wang Yuefen, Chen Shengzhi, et al. Research on Distribution Status and Evolution among Disciplines in Multidisciplinary Comprehensive Research Fields [J]. Journal of the China Society for Scientific and Technical Information, 2020, 39(05):459-468.

[27] Deng S L, Xia S D. Mapping the interdisciplinarity in information behavior research: a quantitative study using diversity measure and co-occurrence analysis [J]. Scientometrics, 2020, 124(4):489-513.

[28] Wu Chake, Wang Shuyi. Research on Theme Discovery and Evolution of Domestic Library Science Based on LDA [J]. New Century Library, 2019(07):90-96.

[29] Shan Bin, Li Fang. Review of LDA-based Topic Evolution Research Methods [J]. Journal of Chinese Information Processing, 2010, 24(06):43-49, 68.

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

Source: ChinaXiv — Machine translation. Verify with original.