Identification of Hot Cross-disciplinary Topics and Thematic Evolution Analysis in Genetic Engineering: An Interdisciplinary Perspective (Postprint)

Authors: Zhu Shiqin, Fan Dandan, Guo Tianyu

Date: 2024-06-13T00:00:00+00:00

Abstract

To more accurately grasp the research hotspots and development trends in interdisciplinary fields, this study proposes a computational method for measuring the interdisciplinary degree of topics, and comprehensively identifies hot interdisciplinary topics by integrating topic strength to predict future development trends of each topic. The study selects papers in the field of genetic engineering from the Web of Science database during 2000-2019 for empirical analysis. First, the LDA model is employed to extract topics. Then, hot interdisciplinary topics are identified by calculating topic strength and topic interdisciplinary degree. Finally, time windows are constructed to plot trend charts of topic strength and topic interdisciplinary degree, and the results are analyzed. Empirical results indicate that there are 21 important topics in the field of genetic engineering, encompassing 7 hot topics, 14 interdisciplinary topics, and 2 hot interdisciplinary topics. According to the changing trends of topic strength, the 21 topics are classified into 3 rising topics, 7 declining topics, and 11 stable topics, with most topics showing an upward trend in interdisciplinary degree.

Full Text

Preamble

To accurately grasp research hotspots and development trends in interdisciplinary fields, this study proposes a method for calculating the degree of interdisciplinary integration within topics and identifies hot interdisciplinary topics by combining topic intensity metrics. The approach employs the Latent Dirichlet Allocation (LDA) model to extract themes from genetic engineering literature, calculates topic intensity and interdisciplinary degree to identify hot interdisciplinary topics, and analyzes future development trends by dividing

time windows and mapping evolutionary patterns. Empirical analysis of genetic engineering publications from 2000–2019 reveals 24 topics in the field, including 8 hot interdisciplinary themes. Based on intensity trends, these are categorized into 4 rising topics, 3 declining topics, 4 stable topics, and 8 hot topics covering 5 important themes. Most topics demonstrate increasing interdisciplinary integration. Keywords: interdisciplinary research; genetic engineering; topic modeling; topic evolution; Rao-Stirling index

1. Literature Review

1.1 Interdisciplinary Topic Identification Methods

Interdisciplinary topic identification typically employs three approaches: citation analysis, lexical analysis, and topic modeling. Interdisciplinary topics emerge from the fusion and permeation of two or more disciplines, serving as hubs for knowledge diffusion and breakthrough points for technological innovation. Citation analysis examines citation relationships among papers, authors, and other objects to identify cross-disciplinary themes. For instance, Adams et al. used co-citation networks to map HIV/AIDS research, while Carley and Porter applied journal coupling networks to identify interdisciplinary clusters. However, citation analysis suffers from temporal lag and cannot capture semantic relationships.

Lexical analysis uses term frequency and co-word analysis to identify interdisciplinary topics. Studies by Xu et al. and Luo et al. employed co-word networks to reveal knowledge structures in information retrieval and genetic engineering vaccine research. While effective, this method fails to capture semantic relationships between term pairs.

Topic modeling, particularly LDA, overcomes these limitations by analyzing latent semantic information. Zhang applied LDA to explore interdisciplinary topics from a clustering perspective, while Chen et al. used it to identify themes in medical informatics. This approach has proven effective in nanotechnology and medical fields, making it suitable for genetic engineering research.

1.2 Interdisciplinary Evolution Research

Interdisciplinary evolution studies primarily focus on macro-level trends and micro-level topic development. Journal-level analyses by Leydesdorff and Silva measured interdisciplinarity through citation networks, while Agarwal and Vugteveen examined knowledge flows between disciplines. The Rao-Stirling diversity index has become a standard metric, measuring interdisciplinary degree through variety, balance, and disparity in reference distributions. Recent studies by Deng et al. and Cao et al. combined social network analysis with diversity measures to map interdisciplinary dynamics in information behavior and artificial intelligence research.

Despite these advances, most studies focus on journals or disciplines rather than

micro-level topics. This study addresses this gap by applying LDA to genetic engineering, integrating topic intensity and interdisciplinary degree to identify hot interdisciplinary themes and analyze their evolution.

2. Research Design and Methods

2.1 Data Collection and Processing

Data were collected from the Science Citation Index Expanded (SCI-Expanded) database for genetic engineering literature from 2000–2019. The search strategy included terms such as "gene* engineering," "DNA manipulat*," "gene recombinat*," and "transgen*." After deduplication and removal of missing values, author keywords and expanded keywords were extracted using Python. Data cleaning involved removing stop words, high-frequency non-discriminative terms, and meaningless interference words, while merging synonyms and performing lemmatization.

2.2 Topic Modeling with LDA

The LDA model was implemented using Gensim to process the preprocessed corpus. The optimal number of topics K was determined by calculating perplexity, with the 拐点 (inflection point) indicating best model fit. The model generates two key matrices: the document-topic probability distribution and the topic-word probability distribution. For each topic, the top 30 terms by probability were selected as characteristic words for topic labeling.

2.3 Topic Intensity Measurement

Topic intensity reflects a theme's importance and attention level within a field, calculated as the average posterior probability of topic occurrence across documents:

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

where tz represents topic intensity in time period t, θ_{dz} is the proportion of topic z in document d, and D_t is the number of documents in period t. The threshold T for identifying hot topics follows Wu et al.'s method:

$$T = \frac{\sum_{d=1}^{D} \sum_{z=1}^{K} \theta_{dz}}{D \times K}$$

Topics with intensity exceeding T are designated as hot topics.

2.4 Interdisciplinary Degree Measurement

The Rao-Stirling (R) index measures single-paper interdisciplinary degree through three dimensions: variety, balance, and disparity:

$$R = \sum_{i,j} d_{ij} p_i p_j$$

where p_i and p_j represent the proportions of cited disciplines i and j, and d_{ij} measures disciplinary distance. The topic-level interdisciplinary degree R_t is calculated as the mean R-index across all documents in the topic:

$$R_t = \frac{1}{m} \sum_{n=1}^{m} R_n$$

where m is the number of documents in the topic. The threshold I for interdisciplinary topics is:

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

Topics exceeding both intensity threshold T and interdisciplinary threshold I are identified as hot interdisciplinary topics.

2.5 Evolutionary Analysis

Time-series analysis was conducted by dividing the study period into annual windows. Topic intensity and interdisciplinary degree were calculated for each window, and trends were visualized. Topics were categorized as rising, declining, or stable based on their intensity trajectories.

3. Empirical Study

3.1 Topic Extraction and Identification

The LDA model was trained on the genetic engineering corpus, yielding 24 topics. Perplexity analysis identified the optimal topic number K = 24. Table 1 presents partial results of the topic-word probability distribution, showing characteristic terms for each topic.

** Topic-Word Probability Distribution (Partial)**

Topic	Characteristic Terms
T1	transcriptome sequencing, RNA-seq, gene expression, bioinformatics

Topic	Characteristic Terms
T2	transgenic drugs, recombinant protein, therapeutic protein, pharmaceutical
T3	transgenic crops, insect resistance, Bt gene, agricultural biotechnology
…	…

3.2 Hot Interdisciplinary Topic Identification

Topic intensity and interdisciplinary degree were calculated for all 24 topics (Table 2). The intensity threshold $T = 0.042$ and interdisciplinary threshold $I = 0.387$ were computed. Topics exceeding both thresholds were identified as hot interdisciplinary themes: transgenic drugs, transgenic crops, Alzheimer's disease, gene cloning technology, apoptosis, and abiotic stress.

** Topic Intensity and Interdisciplinary Degree**

Topic	Intensity	Interdisciplinary Degree
Transcriptome sequencing	0.051	0.421
Transgenic drugs	0.048	0.412
Transgenic crops	0.045	0.398
Alzheimer's disease	0.056	0.435
…	…	…

3.3 Evolutionary Trend Analysis

Annual analysis from 2000–2019 revealed distinct patterns:

Rising Topics (4): Alzheimer's disease, apoptosis, plant insect resistance, and abiotic stress showed increasing intensity and interdisciplinary degree. Alzheimer's disease research attracted growing attention from neuroscience, biochemistry, and molecular biology, with intensity rising steadily despite minor fluctuations in 2009 and 2010. Apoptosis maintained high, stable intensity with gradually increasing interdisciplinary engagement, reflecting its broad applications in medicine and agriculture.

Declining Topics (3): Gene therapy, transcriptome sequencing, and transgenic crops exhibited decreasing intensity. Transgenic crops peaked around 2010 before declining, though interdisciplinary degree continued rising. Vaccine research, while declining in intensity after 2010, saw increased interdisciplinary collaboration, particularly during the COVID-19 pandemic when funding and cross-disciplinary efforts surged.

Stable Topics (4): Plant disease resistance, biodiversity conservation, and plant remediation showed stable intensity with rising interdisciplinary degree, indicating maturing research areas with expanding collaborative networks.

[Figure 1: see original paper] Evolutionary Trends of Topic Intensity and Interdisciplinary Degree

The figure illustrates annual changes for each topic, showing how hot interdisciplinary themes like Alzheimer's disease and apoptosis maintain high intensity while increasing in interdisciplinary integration.

4. Conclusions and Limitations

This study demonstrates that integrating LDA topic modeling with Rao-Stirling interdisciplinary metrics effectively identifies hot interdisciplinary topics in genetic engineering. Key findings include:

1. **24 topics** were extracted, with **6 hot interdisciplinary themes** identified through dual-threshold filtering.
2. **Topic intensity is dynamic**: Rising topics like Alzheimer's disease and apoptosis reflect evolving research priorities, while declining topics such as transgenic crops indicate shifting focus.
3. **Interdisciplinary integration is increasing**: Most topics show rising interdisciplinary degree, particularly hot themes like vaccines and Alzheimer's disease, which attract diverse disciplinary contributions.
4. **Cross-disciplinary collaboration is deepening**: The complexity of research problems drives breaking disciplinary boundaries, as seen in neuroscience and molecular biology contributions to Alzheimer's research.

The LDA-Rao-Stirling framework provides a replicable method for other fields. However, limitations include potential bias in optimal topic number selection via perplexity and incomplete topic naming. Future research should combine expert interviews with computational analysis to deepen understanding of interdisciplinary dynamics.

[Figure 2: see original paper] Distribution of Topics by Intensity and Interdisciplinary Degree

The scatter plot visualizes the positioning of all 24 topics relative to the thresholds, clearly demarcating hot interdisciplinary themes in the upper-right quadrant.

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