Postprint: Emerging Topic Identification and Feature Association Based on Topic Modeling and Time Series Analysis

Authors: Li Yaqian, Sun Yuling, Zhao Wanyu

Date: 2023-10-08T00:00:00+00:00

Abstract

Purpose/Significance: Research on emerging topic identification that scientifically and effectively uncovers the correlation patterns of their characteristics can better serve practical needs while leveraging the innovative supporting function of scientific and technological intelligence research in disciplinary development.

Methods/Process: Beginning with the definition of emerging topic characteristics and integrating relevant theories and practices from emerging topic research and scientific impact evaluation, we establish a methodological framework for emerging topic identification that employs natural language processing, global principal component analysis, and time series analysis methods. This framework quantifies features such as topic consistency, novelty, impact, and growth, thereby completing the extraction, analysis, and identification of emerging topics in conjunction with trend forecasting. Building upon this foundation, we deeply explore the developmental patterns of emerging topics in the target field by employing Granger causality tests and cointegration analysis to conduct long-term equilibrium tests and causal relationship inferences regarding their characteristic correlation effects, analyzing the long-term correlation factors that influence emerging topic development and their functional relationships.

Results/Conclusion: We propose a comprehensive methodological suite for emerging topic identification and analysis of their correlated characteristics. To verify the feasibility and effectiveness of this approach, we select the wetland field for empirical research. By integrating topic identification with characteristic correlation effect analysis, we depict the dynamic development path of scientific impact for topics in this field and propose construction considerations for emerging topics from the perspective of correlated characteristics.

Full Text

Research on Emerging Topic Recognition and Feature Association Based on Topic Model and Time Series Analysis

Li Yaqian[1],[2], **Sun Yuling**[1],[2], **Zhao Wanyu**[1]

[1] National Science Library, Chinese Academy of Sciences, Beijing 100080
[2] Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100049

Abstract:
[Purpose/Significance] Conducting research on emerging research topic (ERT) identification and scientifically discovering their characteristic correlation patterns can better serve practical needs and leverage the innovative supporting role of scientific and technological information research in disciplinary development. Aiming to discover emerging research topics and their characteristic correlation effects scientifically and effectively, this paper carries out ERT identification and feature analysis while realizing the innovative supporting role of sci-tech information work. [Method/Process] Starting from the definition of emerging topic characteristics, and combining relevant theories and practices from emerging topic research and scientific impact assessment, this study establishes a methodological framework for emerging topic identification using natural language processing, global principal component analysis, and time series analysis. It quantifies features such as topic consistency, novelty, influence, and growth, and completes the extraction, analysis, and identification of emerging topics through trend forecasting. On the basis of emerging topic identification, the study deeply mines the development patterns of emerging topics in the target field, using Granger causality tests and cointegration analysis to conduct long-term equilibrium tests and causal relationship inference on their feature association effects, analyzing the long-term correlation factors influencing emerging topic development and their relationships. [Result/Conclusion] The study proposes a set of methods for emerging topic identification and analysis of their associated features. To verify the feasibility and effectiveness of the method, the wetland field is selected for empirical research. Combined with topic identification and feature association effect analysis, the dynamic development path of the scientific impact of topics in this field is depicted, and considerations for emerging topic construction are proposed from the perspective of associated features.

Keywords: trend forecasting; emerging research topic identification; characteristic correlation effect; cointegration analysis; panel data analysis

2. Background and Significance

With the rise of the fourth paradigm of scientific research, data-driven scientific inquiry is sinking from the knowledge layer to the data layer, and the formulation of scientific and technological development plans and related policies needs to keep pace with research dynamics. As an important carrier of knowledge flow, literature is a crucial data source for identifying disciplinary topics. Faced with massive textual data, how to scientifically and effectively discover emerging research topics from it serves as an important reference basis for research managers and researchers to layout and adjust research directions. At the same time, disciplinary topic development has "inertia" and "correlation/continuity," meaning that the temporal sequence changes and development of disciplinary topics are continuous and interconnected, with predictable development patterns within a certain period. The identification and trend prediction of emerging topics can help researchers understand research dynamics, assist funding organizations and decision-makers in optimizing the allocation of innovative resources, and further promote the development of research directions with growth potential.

Similar concepts to emerging topics abound, such as hot topics, frontier topics, and disruptive topics, which have evolved into concepts like general innovative topics, emerging frontier topics, and scientific frontiers, often leading to blurred conceptual boundaries in research and application. H. Xu et al. measured the research heat and development trends of the "emerging topic" concept family, pointing out that differences and intersections exist among the concepts, and that compared to frontier topics and disruptive topics, scholars' research interest in emerging topics is growing faster. The differences among emerging topic-related concepts are mainly reflected in the temporal dimension and innovation dimension. Hot topics, emerging topics, and frontier topics represent important research topics of the past, present, and future, respectively, with their degree of innovation gradually increasing over time and the difficulty of prediction also increasing.

In terms of emerging topic identification methods, scholars primarily utilize related techniques such as co-word analysis, citation analysis, and text mining analysis to extract and identify emerging topics from scientific literature. In recent years, discussions on emerging topic features have become increasingly common, with most scholars focusing on external historical features of literature, such as the historical evolution of textual topics and citation patterns, while paying less attention to future development trends. Wang Shan believes that emerging topics represent future trends in research fields, and their trend analysis and interpretation are particularly important. As related research interest continues to grow, identification methods are becoming increasingly diversified and scientific. However, there remains a lack of good connection between clear conceptual definitions of emerging research topics and proposed operational indicators. Therefore, how to mine the association relationships between emerging topics and their features, adopt effective feature schemes, and extract long-term correlation variables through constructing scientifically rigorous prediction mod-

els and using appropriate analytical methods can provide some references for emerging topic identification.

3. Emerging Topic Identification Method Framework

The emerging topic identification and analysis framework proposed by the author mainly consists of four parts (see Figure 1 [Figure 1: see original paper]). For textual data, LDA topic identification is used to generate topic time series, which are combined with ARIMA models and global principal components to quantify topic features and construct an emerging topic identification scheme. On the basis of emerging topic identification, panel cointegration analysis and Granger causality inference are comprehensively employed to mine the long-term relationships and association effects among observed variables, analyzing the long-term association relationships between emerging topics and their features.

3.2 Quantitative Indicators for Emerging Topic Features

3.2.1 Future High Growth Potential Future high growth potential refers to topics having good development potential in the future. This study primarily uses the ARIMA model to predict future trends based on topic intensity data. The ARIMA(p,d,q) model includes an AR process, MA process, and differencing integration process, with three main parameters: p is the number of autoregressive terms, d is the order of differencing for stationarity, and q is the number of moving average terms. The ARIMA model can be expressed as: MATH_1

3.2.2 Novelty The measurement of novelty is a key part of identifying novel topics. Y. N. Tu et al. used publication time to calculate a novelty index. Bai Jingyi et al. added topic lifecycle theory to define novelty, as shown in formula (2): MATH_2

3.2.3 Consistency and Coherence Consistency and coherence refer to topics that have existed for a period of time and have a continuous and stable development trend. Q. Wang et al. defined topic coherence as the degree of looseness of topic links, measured by the ratio of citation count to publication count within the field (consistency index), with a threshold of 1. S. Xu believed that coherence depends on whether the topic extraction method can ensure sufficiently coherent topics. Bai Rujiang et al. suggested that by time slicing, topics meeting set standards in continuous time intervals satisfy coherence requirements. This study comprehensively employs adjacent time slicing and consistency index calculation methods to measure consistency and coherence features.

3.2.4 Scientific Impact and Growth The paradigm for scientific impact assessment includes quantity, quality, and effectiveness theories, involving three

dimensions: conditions for research output, presentation carriers, and dissemination, as well as indicators such as research intensity, research performance, research support capacity, degree of institutional diversification, and research output dissemination capacity. For the analysis of scientific impact of emerging topics, there are cases of single and multiple indicators. This study, based on the scientific impact assessment paradigm, selects citation count, author count, institutional scale, and disciplinary richness as comprehensive observation indicators for scientific impact.

Specific indicators include: - **Topic Citation Index (TCI):** MATH_3 - **Topic Author Index (TAT):** MATH_4 - **Topic Category Index (TCG):** MATH_5 - **Topic Institution Index (TIS):** MATH_6 - **Topic Intensity (TI):** MATH_7

Growth measurement is reflected in citation growth, author growth, institutional scale expansion, and convergence of different disciplines, measured as changes in adjacent time data. Taking topic intensity growth as an example, the measurement formula is: MATH_8

3.3 Topic Feature Association Analysis

To deeply mine the internal development patterns of emerging topics in the target field, this study employs topic feature association analysis methods. Nobel laureate C. W. J. Granger proposed cointegration analysis and Granger causality test methods based on "prediction" in 2003, providing statistical tests for long-term relationships between variables to determine causal associations. For panel data containing cross-sectional individual features and temporal variation features, C. W. Kao et al. proposed panel cointegration test methods, and E. I. Dumitrescu and C. Hurlin expanded the testing methods for Granger causality in panel data, enabling better analysis of the association mechanism between independent and dependent variables.

4. Empirical Analysis of Emerging Topic Identification

4.1 Data Collection

The author conducts empirical analysis using research paper data from the "wetland" field, retrieving from the SCIE (SCI-Expanded) and SSCI (Social Sciences Citation Index) databases in the Web of Science Core Collection. After sorting wetland types and expressions and designing retrieval strategies using related keywords, with title, abstract, author keywords, and keywords as identification fields, the retrieval formula TI=((wetlands or wetland or "wet land" or "wet lands" or marsh or swamp* or peatland* or "peat land*" or bog or bogs or mire or mires or fen or fens or everglade* or mangrove*)) not TS=("swamp crayfish" or "marsh sandpiper" or "marsh mallow" or "marsh harbour") was used for topic retrieval. The retrieval period was limited to January 1, 2000, to December 31, 2020, with retrieval conducted in September 2020. Literature types were limited

to “article” and “review,” yielding a total of 24,449 wetland-related documents. The annual distribution of papers is shown in Figure 4 [Figure 4: see original paper], showing good development momentum and stable growth.

4.2 Topic Identification and Data Extraction

The author uses Python for topic identification, selecting models with 1-175 topics and comprehensively comparing perplexity and coherence performance. Perplexity uses probability to calculate a topic model's performance on test sets, with lower values indicating better models. Perplexity analysis results show insignificant differentiation. C_v, U_{mass}, C_{npmi}, and C_{uci} coherence are all consistency indicators measuring whether words within topics support each other. In the consistency indicator results, 26 topics are optimal, as shown in Figure 5 [Figure 5: see original paper].

Through natural language processing such as word segmentation and topic modeling, the topic-keyword distribution is exported, yielding 26 research topics in the wetland field (see Table 1). Combined with manual interpretation and translation, the wetland field includes constructed wetland regeneration, wetland ecological monitoring, environmental climate change response, wetland pollution component analysis, wetland biodiversity conservation, wetland gas emission flux models and monitoring, degraded wetland system restoration, wetland cycling system analysis, regional wetland management, wetland restoration standard techniques, and wetland ecological protection.

4.3 Emerging Topic Identification Analysis

4.3.1 Consistency and Coherence Analysis To detect the consistency and coherence of wetland field topics, time slicing was used to calculate topic consistency indices for 2016-2020 and 2011-2015, with results shown in Figure 6 [Figure 6: see original paper]. The horizontal axis represents topic numbers, and the vertical axis represents topic consistency index calculation results. Topic consistency indices in adjacent time intervals are far above the set threshold, indicating that the 26 research topics identified by the topic model are tightly connected and meet consistency and coherence requirements.

4.3.2 Potential High Growth Potential Analysis For potential high growth potential, the author constructs ARIMA models to predict future topic trends. To avoid potential autocorrelation and heteroskedasticity issues, data were log-transformed before stationarity testing. Three test types were used: trend and intercept (c,t), intercept only (c,0), and no trend no intercept (0,0), with the test type determined by significance. Stationarity test results are shown in Table 2 . In topic intensity sequences, after differencing, sequences for topics 1, 5, 6, 12, 21, and 23 became stable, while the rest were stationary sequences, enabling modeling.

After unit root testing, ACF and PACF plots for order determination, combined

with information criteria (AIC, SC, and HQ minimum principle) and parameter comparison, ARIMA model forms were determined. Due to extensive process data, Table 3 shows only final model parameter order results, with Topic 5 as an example to demonstrate the modeling process.

As shown in Figure 7 [Figure 7: see original paper], Topic 5's autocorrelation plot cuts off at lag 3, and partial autocorrelation plot cuts off at lag 1, with model parameter p taking 0-3 orders and parameter q taking 0-1, yielding 8 possible combinations. Through information criteria comparison, the optimal model form was determined (see Figure 8 [Figure 8: see original paper]). Based on this, topic trend fitting and prediction analysis were conducted. The left side of Figure 9 [Figure 9: see original paper] shows the 2000-2018 topic intensity trend fitted by the ARIMA model, showing growth; the right side shows Topic 5's 5-year future trend prediction results, showing stable performance.

4.3.3 Impact and Growth Analysis Time-series global principal component analysis uses comprehensive variables to replace original global variables, capturing main impact features. By calculating annual measurement indicators for 2001-2018, a 260$\times$18 time-series data table was obtained with 4,680 data points, showing correlations among indicators (see Figure 10 [Figure 10: see original paper]). To eliminate dimensional effects, standardization was applied. The Bartlett test statistic was 9,135.283 with p-value approaching 0, and KMO test value greater than 0.7, suitable for principal component analysis.

Calculating initial and factor solutions for global principal component analysis, based on the eigenvalue greater than 1 principle, principal components F1 and F2 were selected, carrying 43.375% and 32.519% of original data information, respectively. The first principal component has positive and large loads on all 5 impact indicators, forming a comprehensive impact factor. The second principal component more reflects topic growth, forming a growth factor.

Using component score coefficients, analytical expressions for the two principal components were obtained: - **Impact Factor:** MATH_9 - **Growth Factor:** MATH_{10}

To better explain the practical significance of principal components, topic two-dimensional distribution can be observed through data standardization and principal component score calculation, as shown in Figure 11 [Figure 11: see original paper]. Topics 7, 13, 16, 23, 24, 25, and 26 show synergistic development effects of high growth and high impact, indicating that high-growth emerging topics can achieve more scientific impact. Topics 1, 2, 4, 5, 6, 10, 11, 12, 14, 18, 19, and 21 show certain substitution effects between growth and impact. Topics 3, 8, 9, 12, and 15 are distributed near the origin, with relatively stable development of impact and growth features.

4.3.4 Emerging Topic Identification Results Comprehensively analyzing various dimensional features of wetland field topics reveals: All 26 topics cal-

culated by the topic model meet consistency and coherence requirements. Potential high growth potential analysis shows that during 2000-2018, most topic intensities showed stable or rising trends; in the next 5 years, topics 5, 6, 7, 9, 13, 14, 15, 16, 17, 18, 22, 23, 25, and 26 have significant potential high growth potential, with expected positive development trends. Novelty performance is good for topics 2, 7, 9, 11, 12, 13, 15, 16, 17, 23, and 25. Joint analysis of growth and impact shows that topics 3, 7, 13, 16, 17, 23, 24, 25, and 26 have good feature performance.

Emerging topic multidimensional identification results are shown in Figure 12 [Figure 12: see original paper]. Results indicate that topics meeting the definition of emerging topics in the wetland field are topics 7, 13, 15, 16, 17, and 25, namely degraded wetland system restoration, wetland microbial gene research, wetland ecological compensation, wetland quantitative survey research, wetland bacterial community system governance analysis, and wetland ecological response to climate change.

4.4 Emerging Topic Feature Association Analysis

Emerging topics have the potential to develop into future hot topics and serve as the foundation for frontier topic incubation. Based on emerging topic identification, deeply mining the long-term relationships of associated features of emerging topics can better understand them and has practical significance.

This study, based on the scientific evaluation system, selects main measurement indicators reflecting research intensity, research performance, institutional diversification, and output dissemination capacity. For panel data composed of emerging topics, topic feature association analysis is conducted (including citation features, author features, institutional scale, and disciplinary richness). To avoid potential heteroskedasticity, data were log-transformed before LLC stationarity testing, where topic multidisciplinary features have unit root processes (first-order integration) while other variables are zero-order integrated.

4.4.1 Long-term Equilibrium Analysis As the data are not integrated at the same order, cointegration testing is needed to determine long-term stable relationships. In Kao-test cointegration testing, the null hypothesis is that no cointegration relationship exists between topic intensity and topic feature data. Based on significance comparison of 5 test statistics including DF and adjusted ADF, conclusions reject the null hypothesis (see Table 4), indicating cointegration relationships exist. Long-term stable relationships exist between topic intensity sequences and topic external feature dimensions, enabling further causal relationship analysis.

According to the cointegration equation: increases in topic institutions, topic author numbers, and topic citation frequency have positive long-term equilibrium relationships with topic intensity; increases in topic disciplinary richness have negative long-term equilibrium relationships with topic intensity, as shown

in Table 5 .

4.4.2 Granger Causality Test Granger causality test is a test of predictive ability. Its basic principle is: assuming mutual influence between variables A and B, if A's lagged variables significantly affect B, then A is Granger-cause of B, and vice versa.

After confirming cointegration relationships between topic intensity and various dimensional features, with unclear direction of effects, the author first uses the Granger causality test method proposed by A. Juodis et al. to test variable exogeneity, determining whether topic feature joint dimensional changes are endogenous factors of topic intensity changes. Test results show that the joint effect of institutions, authors, citations, and disciplinary richness on topic intensity is significant at the 0.05 level, indicating these four variables' joint changes are endogenous factors of topic intensity changes. To study specific causal relationships between variables, further Granger causality tests were conducted, with results shown in Table 7 .

Analyzing Table 7 Granger causality test results yields the following conclusions: (1) For emerging topics in the wetland field, bidirectional Granger causality exists between topic intensity and topic institutional numbers and topic author numbers. This indicates that growth of research scholars in the field promotes development of emerging topics, and topic intensity growth also attracts new scholars to conduct related research, verifying cluster effects and showing that talent development and topic development are mutually reinforcing active modes. This reflects the effectiveness of research support institutions' incentive policies in the wetland field, suggesting a project-first, talent-based execution approach for future disciplinary topic development. (2) In the wetland field, unidirectional causality exists between topic intensity and topic disciplinary richness and topic citations. Specifically, good topic intensity development is the cause of topic disciplinary richness, but disciplinary richness is not the cause of good topic intensity development; topic intensity growth is the cause of citation frequency increase, while citation frequency increase is the cause of topic intensity change. The practical implication is that topic intensity has a unidirectional effect on topic richness—topic intensity expands over time, promoting disciplinary diversification in the wetland field. However, disciplinary richness development does not significantly optimize topic intensity growth, indicating that promoting disciplinary richness cannot directly promote healthy topic intensity growth in this field. Blindly pursuing disciplinary richness in the wetland field may lead to excessive topic fragmentation, making it difficult to achieve "large yet refined." Additionally, citation patterns represent topic attention shifts, and topic intensity growth's pulling effect on citations is not significant in the short term. Conversely, citation frequency increase has a significant promoting effect on topic intensity development, making it a "weathervane" for topic intensity development in this field.

Conclusion

From paper data, this study proposes a set of identification and association analysis methods based on emerging topic features. In feature extraction, combining emerging topic theories and practices, improvements were made in novelty and other aspects, adding potential high growth indicators, and selecting comprehensive feature consideration schemes for impact and growth. This study extracts research topics and distributions through topic models, uses trend prediction models and analysis methods to analyze topic future trends, combines global principal component analysis to depict dynamic development paths of topic growth and impact, and completes emerging topic identification based on comprehensive topic performance. To better identify emerging topics, the author uses cointegration analysis and Granger causality tests to mine feature association relationships of emerging topics. The study finds bidirectional association effects between topic intensity and institutional numbers and author scale, positive impact of topic citation frequency on topic development, and unidirectional promoting effect of topic intensity on topic diversity. Therefore, the author proposes adhering to a project-first, talent-based innovation policy execution approach and offers some thoughts on developing emerging topics. The author repeatedly considered feature scientificity and identification comprehensiveness, comprehensively employing natural language processing, multivariate statistical analysis, and time series analysis methods to determine emerging topic identification and feature analysis methods, which have certain reference value for objectively understanding research topic dynamics in fields and making research layout decisions.

The emerging topic identification and analysis method proposed by the author mainly proceeds from the perspective of scientific literature. Since emerging topics are comprehensive features of research content in a field, their research value is reflected in science, policy, economy, and other aspects, while literature is only an important object reflecting innovative changes in research topics. In addition to scientific literature, research objects also include policy texts and patent data. Therefore, future research could attempt to fuse multi-source texts for comprehensive emerging topic identification research.

References

[1] Liu ZQ, Wang XY, Bai RJ. Research on visualization analysis methods for disciplinary topic evolution from a multi-dimensional perspective: A case study of big data research in China's library and information science field[J]. Journal of Library Science in China, 2016, 42(6): 67-84.

[2] Wang S. Progress in research frontier detection methods[J]. Information Science, 2019, 37(10): 164-169.

[3] XU H, WINNINK J, YUE Z, et al. Multidimensional scientometric indicators for the detection of emerging research topics[J]. Technological forecasting and social change, 2021, 163: 1-25.

[4] LU C, HOU H, DING Y, et al. Review of internatinonal studies on discovering emerging topics[J/OL]. Journal of the China Society for Scientific and Technical Information, 2019[2021-09-13]. http://en.cnki.com.cn/Article_{en}/CJFDTotal-QBXB201901011.htm.

[5] LIU G Y, HU J M, WANG H L. A co-word analysis of digital library field in China[J]. Scientometrics, 2012, 91(1): 203-217.

[6] CHI R, YOUNG J. The interdisciplinary structure of research on intercultural relations: a co-citation network analysis study[J]. Scientometrics, 2013, 96(1): 147-171.

[7] SONG M, KIM S Y. Detecting the knowledge structure of bioinformatics by mining full-text collections[J]. Scientometrics, 2013, 96(1): 183-201.

[8] Zhong HX. A review of emerging trend detection research[J]. Modern Information, 2017, 37(12): 162-167.

[9] XU S, HAO L, AN X, et al. Emerging research topics detection with multiple machine learning models[J]. Journal of informetrics, 2019, 13(4): 100983.

[10] Liu XL, Tan ZY. Research progress on emerging technology topic identification methods[J]. Library and Information Service, 2020, 64(11): 145-152.

[11] DE SOLLA PRICE D J. Networks of scientific papers[J]. Science, 1965, 149(3683): 510-515.

[12] OHNIWA R L, HIBINO A, TAKEYASU K. Trends in research foci in life science fields over the last 30 years monitored by emerging topics[J]. Scientometrics, 2010, 85(1): 111-127.

[13] TU Y N, SENG J L. Indices of novelty for emerging topic detection[J]. Information processing & management, 2012, 48(2): 303-325.

[14] ROTOLO D, HICKS D, MARTIN B R. What is an emerging technology?[J]. Research policy, 2015, 44(10): 1827-1843.

[15] WANG Q. A bibliometric model for identifying emerging research topics[J]. Journal of the Association for Information Science and Technology, 2018, 69(2): 290-304.

[16] SMALL H. Co-citation in the scientific literature: a new measure of the relationship between two documents[J]. Journal of the American Society for Information Science, 1973, 24(4): 265-269.

[17] CHEN C. CiteSpace II: detecting and visualizing emerging trends and transient patterns in scientific literature[J]. Journal of the American Society for Information Science and Technology, 2006, 57(3): 359-377.

[18] Wang YP. Research progress on topic discovery and evolution of scientific literature based on topic models in China[J]. Library and Information Service, 2016, 60(3): 130-137.

[19] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of machine learning research, 2003, 3(4/5): 993-1022.

[20] GERRISH S, BLEI D M. A language-based approach to measuring scholarly impact[C/OL]. [2021-08-10]. https://openreview.net/forum?id=HJ-9EsbdWr.

[21] XU M, LI G, WANG X. Detecting emerging topics by exploiting probability burst and association rule mining: a case study of library and information science[J]. Malaysian journal of library & information science, 2020, 25(1): 47-66.

[22] Li J, Xu LL. Comparison and analysis of research hotspot trend prediction models based on machine learning algorithms: BP neural network, support vector machine and LSTM models[J]. Modern Information, 2019, 39(4): 23-33.

[23] Bai JY, Yan DW, Chen Q. Research on emerging topic trend prediction based on topic model and curve fitting[J]. Information Theory and Practice, 2020, 43(7): 130-136, 193.

[24] KONTOSTATHIS A, GALITSKY L M, POTTENGER W M, et al. A survey of emerging trend detection in textual data mining[C]//BERRY M W. Survey of text mining: clustering, classification, and retrieval. New York: Springer, 2004: 185-224.

[25] LEE C, KWON O, KIM M, et al. Early identification of emerging technologies: a machine learning approach using multiple patent indicators[J]. Technological forecasting and social change, 2018, 127: 291-303.

[26] Yue LX, Zhou XY, Chen YN. Research on trend prediction of information architecture research topics based on ARIMA model[J]. Knowledge of Library and Information Science, 2019(5): 54-63, 72.

[27] Yue LX, Liu ZQ, Hu ZY. Research on hot topic evolution analysis methods for trend prediction[J]. Data Analysis and Knowledge Discovery, 2020, 4(6): 54-63.

[28] Liu ZQ, Xu HY, Yue LX, et al. Research on the lag effect of topic diffusion evolution for research frontier prediction[J]. Journal of the China Society for Scientific and Technical Information, 2018, 37(10): 967-977.

[29] Bai RJ, Liu BW, Leng FH. Research on identification of future emerging scientific research frontiers based on multi-dimensional indicators[J]. Journal of the China Society for Scientific and Technical Information, 2020, 39(7): 747-757.

[30] Wang Q, Tan ZY, Qian L. A review of social impact assessment of scientific research[J]. Library and Information Service, 2015, 59(14): 143-148.

[31] GONZALEZ-ALCAIDE G, GORRAIZ J, HERVAS-OLIVER J L. On the use of bibliometric indicators for the analysis of emerging topics and their evolution: spin-offs as a case study[J]. Profesional de la informacion, 2018, 27(3): 493-510.

[32] GUO H, WEINGART S, BÖRNER K. Mixed-indicators model for identifying emerging research areas[J]. Scientometrics, 2011, 89(1): 421-435.

[33] Wan LL, Gan JF, Yu XY. Time series global principal component analysis method and application based on Matlab[J]. East China Economic Management, 2010, 24(1): 148-151.

[34] CHEN B, TSUTSUI S, DING Y, et al. Understanding the topic evolution in a scientific domain: an exploratory study for the field of information retrieval[J]. Journal of informetrics, 2017, 11(4): 1175-1189.

[35] ENGLE R F, GRANGER C W J. Co-integration and error correction: representation, estimation, and testing[J]. Econometrica, 1987, 55(2): 251-276.

[36] KAO C W, CHIANG M H. On the estimation and inference of a cointegrated regression in panel data[J]. Advances ecoometrics, 2000, 15: 179-222.

[37] DUMITRESCU E I, HURLIN C. Testing for granger non-causality in heterogeneous panels[J]. Economic modelling, 2012, 29(4): 1450-1460.

[38] Qiao F, Yao J. Application of time series global principal component analysis in dynamic depiction of economic development[J]. Application of Statistics and Management, 2003(2): 1-5.

[39] Luo R, Xu HY, Dong K. A review of research frontier identification methods[J]. Library and Information Service, 2018, 62(23): 119-131.

[40] Huang XB, Wu G. A review of research frontier detection methods in disciplinary fields[J]. Journal of the China Society for Scientific and Technical Information, 2019, 38(8): 872-880.

[41] Gu ZS. Research on the relationship between trade openness and carbon dioxide emissions in China[J]. Academic Forum, 2012, 35(8): 109-112.

[42] JUODIS A, KARAVIAS Y, SARAFIDIS V. A homogeneous approach to testing for Granger non-causality in heterogeneous panels[J]. Empirical Economics, 2021, 60(1). DOI: 10.1007/s00181-020-01885-3.

Author Contributions

Li Yaqian: Research framework construction, data analysis, article writing
Sun Yuling: Paper guidance, manuscript revision
Zhao Wanyu: Data collection and preprocessing

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