AI Systems Engineering
https://aise.cultechpub.com/aise
E-ISSN: 3093-8449
Vol. 2, No. 2, November 2026
DOI:10.64229/kandhq75

Cultech Publishing

Article

Heterogeneous Graph Representation Learning and Research Topic Evolution Analysis Method for Open Academic Big Data

Zhihong Tang*

Jiangxi Institute of Technology, Jiangxi, China
*Correspondence author: Zhihong Tang, 12925114@qq.com

Abstract

The interplay of entity heterogeneity and temporal dynamic evolution in open academic big data makes the evolution paths of research topics difficult to characterize. This paper proposes a heterogeneous graph representation learning and research topic evolution analysis method for open academic big data, and defines the prediction target as the topic label assigned to a paper and its keyword set in the following year. A temporal heterogeneous academic information network containing papers, scholars, journals, and keywords is constructed first, and type-specific linear projection unifies multi-source heterogeneous features into a shared latent space. A Heterogeneous Graph Transformer captures the multi-hop neighborhood topology and higher-order semantics of academic entities through a relational adaptive scaling mechanism, while a dynamic graph neural network with continuous-time encoding and gated recurrent units updates the temporal hidden states of nodes. Topic drift paths are measured by combining OPTICS density clustering with the Wasserstein distance of optimal transport theory. Experiments on 1.2 million documents, 850,000 scholars, and 4.5 million citations show that the method reaches 93.25% accuracy and 91.87% macro F1 on the 2024 independent test set, a spatial clustering modularity of 0.842, and a topic drift RMSE of 0.041, with 11.4 hours of training on four graphics processing units. The method captures the multidimensional spatiotemporal evolution patterns of academic topics and suppresses long-range prediction errors.

Keywords

Heterogeneous graph neural network, Heterogeneous graph transformer, Topic evolution analysis, Academic big data

Article History

Received: 16 September 2026
Revised: 29 October 2026
Accepted: 19 November 2026
Available Online: 27 November 2026

Copyright

© 2026 by the authors. This article is published by the Cultech Publishing Sdn. Bhd. under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0): https://creativecommons.org/licenses/by/4.0

1. Introduction

With the explosive growth of open academic big data, massive amounts of academic documents, scholars, journals, and keywords have intertwined into a complex and dynamic multi-source network. How to accurately identify research topics from these heterogeneous and temporally evolving academic networks and quantitatively characterize their evolutionary trajectories has become a core issue in scientometrics, knowledge graphs, and research management. This question has been examined along three separate lines, bibliometric counting of entity co-occurrence, probabilistic topic models over document text, and embedding-based comparison of topic semantics, and each line reports the rise, split, and decline of research themes from a single view of the academic record [1].

This paper integrates four core entities–papers, scholars, journals, and keywords–and multiple relationships, designing a type-specific linear projection mechanism to uniformly map multi-source heterogeneous original features to a shared latent space. Through temporal window segmentation and continuous temporal encoding, it balances network structural integrity with fine-grained temporal details. A heterogeneous graph transformer (HGT) is introduced to capture multi-hop neighborhood topology and higher-order semantics through a relational adaptive scaling mechanism. Combined with a dynamic graph neural network (DGNN) fused with continuous temporal encoding and a gated recurrent unit (GRU), dynamic updates of node temporal states and long-range evolution tracking are achieved. This paper combines OPTICS density clustering to partition dynamic research topics and introduces Wasserstein distance to accurately measure topic drift and evolution paths between adjacent time steps. Different from dynamic topic models that displace topic vectors inside a text space and from dynamic graph networks that update node states without a topic-level readout, the proposed framework couples the two levels in one objective. The relational adaptive scaling of the HGT supplies type-aware neighborhood semantics, the continuous-time encoder keeps the exact arrival time of citations inside a yearly snapshot, the GRU carries the hidden state across snapshots, and the Wasserstein distance computed on clustered topic centers converts node-level drift into a topic-level evolution path. The contribution is the joint temporal-heterogeneous representation together with the transport-based drift measurement that operates on its output, rather than the isolated combination of four existing modules [2,3].

As open academic big data exhibits significant heterogeneity, dynamism, and multi-source association characteristics, the analysis of the evolution of research topics has shifted from single-dimensional frequency statistics to the comprehensive modeling needs that integrate semantic context, temporal dynamics, and interdisciplinary network structures. This trend has been fully reflected in the following empirical explorations in multiple fields. Wu et al. [4] linked topic modeling and bibliometrics to systematically draw a temporal map of the evolution of educational NLP from text processing to intelligent assessment, revealing the coupling mechanism between technological impetus and teaching needs. Shiu and Lehti-Shiu [5] used plant science as an example to identify and compare research topics in different periods using literature abstract text and topic models, and evaluated the patterns of topic rise, decline, and continuous evolution. ZareRavasan and Jeyaraj [6] identified the trajectory of the research on the commercial value of information systems from the productivity paradox to the socio-technical paradigm through long-cycle slicing. Wang et al. [7] proposed a new method for identifying interdisciplinary topics and their evolution paths based on the BERTopic model, combined with interdisciplinary citation relationships and text semantic representation. Huang et al. [8] introduced contextual word embedding into the evaluation of topic semantic consistency, revealing the phenomenon of "same name, different reality" in topic evolution and improving the semantic precision of topic evolution analysis.

At the application level, Xu et al. [9] used the Dynamic Topic Model (DTM) to model the time-series topics of STEAM education literature and track the content changes and popularity fluctuations of each topic. Taher Harikandeh et al. [10] mapped the literature to a high-dimensional semantic space and quantified the evolution trajectory by calculating the displacement of topic vectors at different time points. Zhang et al. [11] constructed a topic stage division method based on curve morphology, providing an operational quantitative measurement tool for the theory of the life cycle of scientific research topics. For engineering collaboration, Zhang et al. [12] took the cognitive-driven product concept design field as a case and comprehensively used co-word analysis, social network analysis and topic evolution model to identify the hot topics and migration paths of cross-domain collaborative research. Lee and Chang [13] used author keywords and MeSH (Medical Subject Headings) words to conduct topic identification and evolution analysis of literature in the field of medical informatics. Existing research is still insufficient in the collaborative depth of heterogeneous academic network representation and long-term time-series evolution tracking, and it is difficult to accurately depict the topic drift trajectory under multi-dimensional association. To this end, this paper proposes a research topic evolution analysis method that integrates HGT and DGNN to achieve efficient deconstruction of the complex evolution laws of open academic big data.

The subsequent structure of this paper is as follows: Section 2 reviews and summarizes related work; Section 3 elaborates on the specific methods for constructing temporal heterogeneous networks, HGT spatial semantic representation, DGNN dynamic evolution tracking, and topic evolution path measurement; Section 4 conducts multi-dimensional experimental evaluation and ablation analysis based on real academic big data; Section 5 summarizes the entire paper.

2. Methods

2.1 Problem Definition and Evaluation Protocol

The prediction task studied in this paper is defined as follows. Given the temporally heterogeneous academic information network observed up to year T, the model predicts, for every paper node and for the keyword nodes attached to it in year T+1, the topic label that these nodes receive in that year. The label space is the topic set retained by the clustering module on the target year, and the number of classes equals the number of retained topics in that year, so the prediction is a multi-class classification over K classes at each time step. The evaluation target is the topic label of the following year, and not a future keyword string, a future citation count, or a future co-authorship edge. Accuracy denotes the proportion of nodes whose predicted label equals the ground-truth label, macro F1 averages the class-wise F1 scores over all K classes with equal weight, micro F1 aggregates the counts of all classes into one confusion matrix, and RMSE drift denotes the root mean square error between the predicted Wasserstein distance of adjacent topic centers and the ground-truth drift value.

Ground-truth labels for the target year are produced independently of the learned representation. Two annotators with a background in computer science assign the label of each node by reading its title, abstract, and keyword list against a codebook built from the topic centers of the preceding year, and disagreements are resolved by a third annotator.

2.2 Construction of Temporally Heterogeneous Academic Information Network

The temporally heterogeneous academic information network is defined as $G=(V,E,T,\mathcal{A},\mathcal{R})$, where the node set V contains four entity types $\mathcal{A}$: Paper, Author, Venue, and Keyword; and the edge set E contains four relationship types $\mathcal{R}$: "Author-Published-Paper", "Paper-Citation-Paper", "Paper-Published in-Journal", and "Paper-Contains-Keyword". During construction, initial attribute features of each entity node are extracted: Papers are extracted using SBERT to generate 384-dimensional text vectors from abstracts and titles; Authors and Venues are represented by 16-dimensional numerical features based on historical publication and citation counts; and Keywords are initialized as 128-dimensional semantic vectors using Word2Vec. Each edge $e\in T$ is bound to a continuous timestamp $\tau\in T$ accurate to the year. When constructing the network, the continuous time axis is divided with a time window step of 1 year to generate a series of discrete heterogeneous subgraphs G_t arranged in time order. At the same time, the original continuous timestamps are retained for later fine-grained temporal coding to ensure that the network has both structural integrity and contains fine temporal evolution information [14,15]. The node type mapping function and the edge relationship mapping function are defined as $\phi(v):V\to\mathcal{A}$ and $\psi(e):E\to\mathcal{R}$, respectively. Based on this, the complete temporal network can be formally represented as a sequence of subgraphs at each time step:

$$\{G_t=(V_t,E_t,\mathcal{A},\mathcal{R})\mid t=1,2,...,T\} \tag{1}$$

T is the total number of time steps; V_t , E_t are the set of nodes and edges of the t -th subgraph, respectively; $\phi(v)$ returns the type of node v ; $\psi(e)$ returns the relationship type of edge e .

The annual subgraphs in Equation (1) serve solely to organize network snapshots and define the macro-level update steps for the DGNN; they do not imply that all temporal information is coarsened into an annual-only model. Each edge retains its original continuous timestamp. To explicitly incorporate these continuous timestamps into the model, we aggregate the continuous timestamps of all edges incident to a given node within an annual snapshot to derive a continuous temporal representation for that node. If a node has no incident edges within an annual snapshot, this representation defaults to the annual timestamp. This node-level continuous temporal representation is subsequently fed into a temporal encoding function (described in Section 2.4) and concatenated with spatial embeddings to serve as input for the DGNN. In essence, the annual subgraphs provide the structural integrity of the network and the sequence of updates, while the continuous timestamps offer fine-grained temporal encoding within each snapshot–together forming a modeling approach characterized by "discrete snapshots combined with continuous temporal encoding."

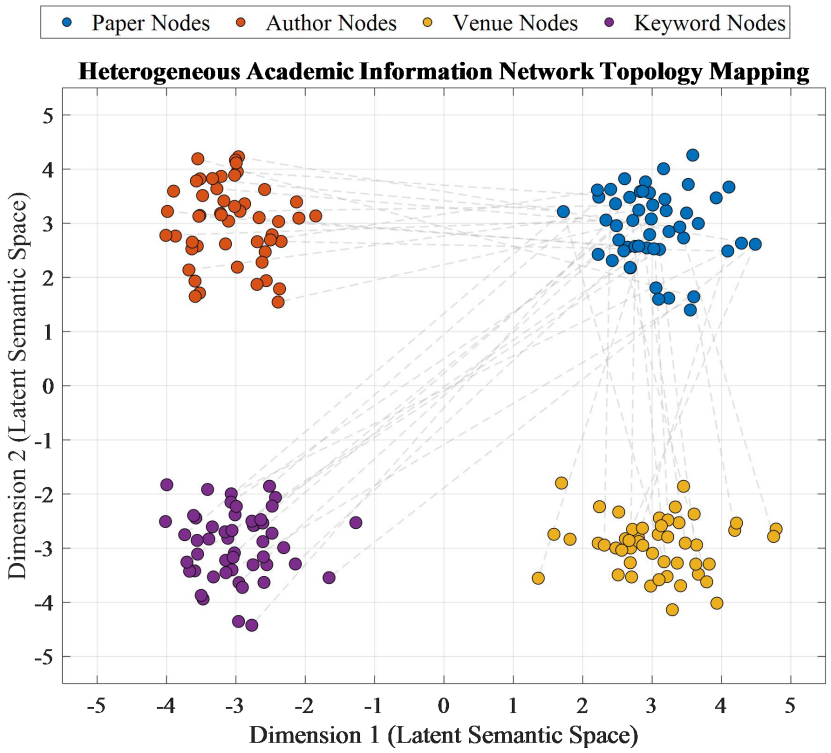

Figure 1. Multidimensional Topological Mapping of Temporally Heterogeneous Networks.

Figure 1 shows the multidimensional mapping distribution of four types of heterogeneous entities—papers, scholars, journals, and keywords—and their multiple interactions in the latent space. Each type of entity presents a clear semantic boundary through different spatial clustering regions, and the dashed lines connecting the entities reflect the multi-source topological association characteristics at the bottom layer of the academic network [16,17].

For any node v, the dimension of its original feature vector x_v^0 varies with $\phi(v)$. The original dimensions of papers, scholars, journals, and keywords are 384, 16, 16, and 128, respectively. To uniformly map the heterogeneous original features to a shared representation space, a type-specific linear projection matrix is constructed for each type of node, mapping the original features to a unified d-dimensional latent space:

$$h_v^0 = W_{\phi(v)} x_v^0, \forall v \in V \tag{2}$$

$W_{\phi(v)}$ is the learnable linear projection matrix of type $\phi(v)$; h_v^0 is the initial embedding vector of node v after projection.

The uniform dimension after projection is set to 128. The continuous timestamps of the edges are preserved by an independent timestamp function $\tau(e): E \to \mathbb{R}$, providing a time reference accurate to the year for subsequent dynamic evolution modeling. Finally, each discrete subgraph G_t not only contains the topological structure within the time window, but also preserves the initial embedding of the nodes after projection and the accurate continuous timestamps of each edge [18].

2.3 Spatial Semantic Representation and Topological Mapping Based on HGT

To address the semantic heterogeneity of different nodes and relationships in academic heterogeneous networks, a HGT is employed for spatial semantic representation. First, for different types of nodes, independent learnable linear projection matrices are used to map the initial features of heterogeneous nodes to the same 128-dimensional latent space. In multi-head attention computation, for each edge relation ρ connecting source node s and target node o, relation-specific matrices K_i^ρ and Q_i^ρ are introduced to calculate attention weights. The projected target node features are used as the Query, and the source node features as the Key and Value. The number of attention heads is set to $H = 8$, and the semantic relevance between nodes is calculated. For the i-th attention head, given the Query projection $Q_i(o)$ of target node o and the Key projection $K_i(s)$ of source node s, the attention weight $ATT_i(s,o,\rho)$ under edge relation ρ is calculated through Softmax normalization:

$$ATT_i(s,o,\rho) = \text{Softmax}_{\forall s \in N(o)}\left(\frac{Q_i(o)K_i(s)^\top}{\sqrt{d_k}} \cdot \mu_{\phi(s),\psi(e),\phi(o)}\right) \tag{3}$$

$Q_i(o)$ is the Query projection vector of the target node o in the i -th head; $K_i(s)$ is the Key projection vector of the source node s in the i -th head; d_k is the key vector dimension of each attention head; $\mu_{\phi(s),\psi(e),\phi(o)}$ are prior adaptive scaling factors based on the source node type, edge relation type, and target node type.

During message passing, the message vector $MSG_I(s,o,\rho)$ passed from the source node s to the target node o by the i -th attention head is:

$$MSG_I(s,o,\rho) = Linear^i_{\phi(s)}(h_s) \cdot v_{\psi(e)} \tag{4}$$

$Linear^i_{\phi(s)}$ is the i -th head linear transformation specific to the source node type $\phi(s)$; h_s is the current layer embedding of the source node; $v_{\psi(e)}$ is the edge weight scaling factor specific to the relation type $\psi_{(e)}$.

The multi-head attention output is obtained by concatenating the weighted messages of all attention heads, and skip connections and layer normalization (LayerNorm) are used to fuse the neighborhood aggregation features with the original node features, with the Dropout rate set to 0.2. After three layers of HGT convolutional iterations, a 128-dimensional node structure embedding matrix Z_t is generated, containing high-order neighborhood topology and multimodal fusion semantics. The forward propagation of the entire HGT layer is uniformly represented as (layer l to layer $l+1$):

$$h_o^{l+I} = \text{LayerNorm}\left(h_o^l + \text{Dropout}\left(\sum_{i=I}^{H} \sum_{S \in N(o)} ATT_i(s,o,\rho) \cdot MSG_I(s,o,\rho)\right)\right) \tag{5}$$

h_o^l represents the embedding vector of the target node o at layer l ; h_o^{l+1} represents the updated embedding. Finally, the embeddings of all nodes after layer 3 are organized by time step t to obtain a spatial structure embedding matrix, where each row corresponds to the spatial representation of a node at that time step.

1000 nodes were randomly sampled, and the mean, variance, and average L2 norm of their embedding vectors after convolution through layers 0 (initial projection), 1, 2, and 3 of the HGT were recorded. The statistical results are shown in Table 1.

Table 1. Statistical distribution of node embeddings after convolution through each HGT layer.

Network Layer	Mean	Variance	Average L2 Norm	Cosine Similarity to Input
Layer 0 (Input Projection)	0.0214	0.8923	1.000	1.0000
Layer 1 HGT Conv	0.0187	0.8456	0.978	0.8624
Layer 2 HGT Conv	0.0152	0.8124	0.953	0.7318
Layer 3 HGT Conv	0.0128	0.7981	0.927	0.6345

2.4 Dynamic Temporal Evolution Tracking Based on DGNN

To capture the continuous state transitions of heterogeneous graph embeddings along the time axis, a DGNN with continuous timestamp features is introduced. The discrete temporal node embeddings output by the HGT are used as static inputs, while continuous temporal encoding is performed on the interaction timestamps of the nodes [19,20]. The temporal encoding function uses an adaptive sine/cosine harmonic function to map the scalar timestamp τ to a 32-dimensional continuous temporal feature vector, which is then concatenated with the node spatial embeddings to form a 160-dimensional input feature. For any timestamp τ , the k -th dimension of the continuous temporal encoding $TEnc(\tau) \in \mathbb{R}^{32}$ is:

$$TEnc(\tau)_k = \begin{cases} \sin\left(\tau / \omega_{base}^{2k/32}\right), & k \text{ is an even number} \\ \cos\left(\tau / \omega_{base}^{2k/32}\right), & k \text{ is an odd number} \end{cases} \tag{6}$$

where k is the absolute dimension index of the 32-dimensional temporal encoding vector; even indices $k = 0,2,30$ use the sine branch, and odd indices $k = 1,3,31$ use the cosine branch. The exponent 2k/32 is computed from the same absolute index k in each branch. ω_{base} is a fixed positive scalar hyperparameter that controls the frequency scale (base frequency) of the harmonic encoding and is not learned or updated by gradients; it is set to $\omega_{\text{base}} = 10000.0$ in all experiments.

The temporal encoding and the spatial node embedding z_v^t of the HGT output are concatenated to obtain the complete input feature $x_v^t \in \mathbb{R}^{160}$ of node *v* at time t :

$$x_v^t = \text{Concat}\left(z_v^t, TEnc(t)\right) \tag{7}$$

The core of dynamic evolution tracking is GRU, whose internal structure includes update gate and reset gate mechanisms. The memory unit takes the hidden state d_v^{t-1} of the previous time step and the node feature x_v^t of the fused temporal encoding at the current time step as input, and dynamically and selectively retains long-term historical theme evolution information and filters short-term noise through the gating mechanism [21,22]. The calculation of the temporal update gate u_t and the reset gate r_t, as well as the generation of the candidate hidden state $\tilde{d}_t$, follow the following gating mechanism:

$$u_t = \sigma(W_z x_v^t + U_z d_v^{t-1} + b_z) \tag{8}$$

$$r_t = \sigma(W_r x_v^t + U_r d_v^{t-1} + b_r) \tag{9}$$

$$\tilde{d}_t = \tanh\left(W_h x_v^t + U_h(r_t \odot d_v^{t-1}) + b_h\right) \tag{10}$$

σ is the Sigmoid activation function; W_z, U_z, b_z are the input weight matrix, cyclic weight matrix, and bias term of the update gate, respectively; W_r, U_r, b_r are the corresponding parameters of the reset gate; W_h, U_h, b_h are the corresponding parameters of the candidate hidden state; $\tilde{d}_t$ is the candidate hidden state.

Finally, the current hidden state d_v^t is updated by interpolation between the previous state and the candidate state through the update gate:

$$d_v^t = u_t \odot d_v^{t-I} + (1 - u_t) \odot \tilde{d}_t \tag{11}$$

This vector records the evolution trajectory of the node and its associated topics.

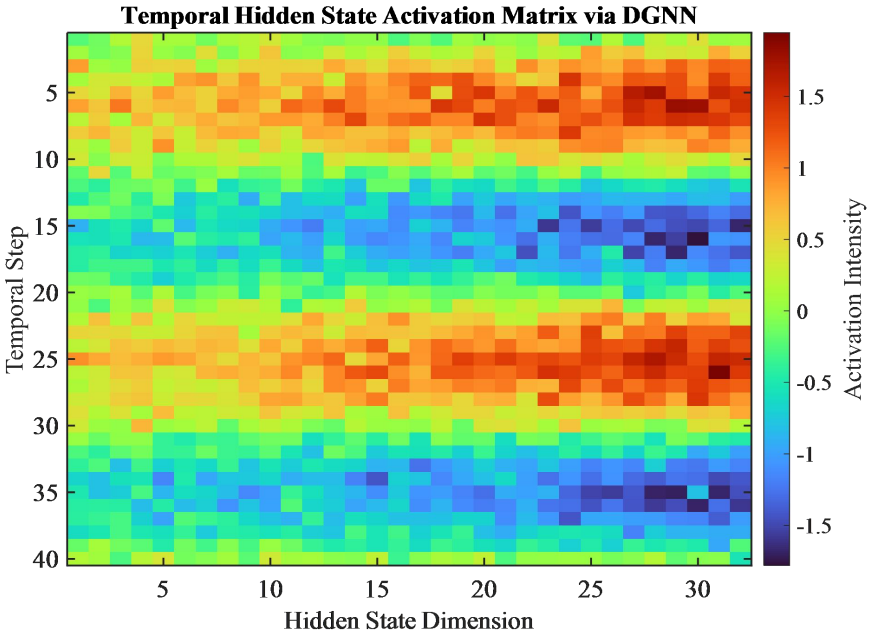

Figure 2. Temporal hidden state heatmap.

The heatmap of Figure 2 intuitively shows the activation intensity of multidimensional features of academic entity nodes after gating update at different time steps, reflecting the dynamic distribution of feature dimensions over time [23,24].

The network training process uses the AdamW optimizer with a learning rate of 0.001 and a weight decay of 0.0001. The model is input to the subgraph frame by frame in chronological order, and updates the 128-dimensional dynamic hidden state d_v^t of each node at each time step in turn, accurately recording the evolution trajectory of the entity and its associated topics.

2.5 Topic Evolution Path and Trajectory Measurement

The topic evolution path analysis module presents the evolution trajectory by performing distance fitting and structural clustering on the continuous temporal node vector matrix generated by DGNN. First, at each time step t, the dynamic hidden state d_v^t of keyword nodes and paper nodes is extracted, and the cosine distance and Euclidean distance between

nodes are calculated to form a comprehensive similarity matrix [25,26]. For any two nodes m,n, their comprehensive distance $D(m,n)$ is a weighted combination of cosine distance and Euclidean distance:

$$D(m,n) = \alpha \cdot \left(1 - \frac{d_m^t \cdot d_n^t}{\left\|d_m^t\right\| \cdot \left\|d_n^t\right\|}\right) + (1-\alpha) \cdot \left\|d_m^t - d_n^t\right\|_2 \tag{12}$$

Let m,n be any two nodes; $D(m,n)$ be the combined distance between them; a be the weighting coefficient balancing the contributions of the two distances (set to 0.5); d_m^t, d_n^t be the dynamic hidden states of nodes m,n at time step t, respectively.

Since the cosine distance and the Euclidean distance in Equation (12) have different dimensions and ranges of values, they must be normalized before being combined. The cosine distance is normalized to the interval [0, 1]. The Euclidean distance is normalized based on the 95th percentile of all pairwise Euclidean distances between keyword nodes and paper nodes at the current time step, and then truncated to the interval [0, 1]. Once normalized, the two distances are combined using a weighted sum as per Equation (12), with α set to 0.5. To verify the appropriateness of the chosen value for α, a sensitivity analysis was conducted; the results are presented in Table 2 below.

Table 2. Sensitivity analysis of the weighting coefficient α.

α	**Macro F1 (%)**	**RMSE Drift**
0.1	90.92	0.046
0.2	91.31	0.044
0.3	91.58	0.043
0.4	91.79	0.042
0.5	91.87	0.041
0.6	91.81	0.042
0.7	91.62	0.043
0.8	91.34	0.044
0.9	90.98	0.046

As shown in the table 2, the Macro F1 score is highest and the RMSE Drift is lowest when α = 0.5; thus, α = 0.5 is a stable and reasonable setting.

The OPTICS density clustering algorithm is used to cluster the keyword vectors at the current time step. The density neighborhood radius ϵ is dynamically set to 0.3, and the minimum number of generated points is set to 5. The dense keyword clusters generated by the aggregation are defined as the research topics at a specific time step. Let K_t topic clusters be obtained at time step t. The topic center vector μ_c^t of each cluster is obtained by averaging the hidden states of all keyword nodes within that cluster .

$$\mu_c^t = \frac{1}{\left|K_c^t\right|} \sum\nolimits_{v \in K_c^t} d_v^t \tag{13}$$

K_c^t is the set of keyword nodes belonging to this cluster; $\left|K_c^t\right|$ is the cardinality (number of keywords) of this set.

The term μ_c^t obtained from Equation (13) serves solely as a central representation for a topic cluster and as an anchor for matching topics across adjacent time steps; it is not used directly as a probability density in the calculations for Equations (14) and (15). In Equations (14) and (15), the terms p and q (as well as p_i and q_i) are derived from empirical distributions constructed using the dynamic hidden states of the keyword nodes within each topic cluster. Specifically, for each topic cluster, the dynamic hidden states of the constituent keyword nodes are treated as samples; these are then used–via soft assignment or kernel density estimation over a shared discrete interval–to obtain p_i and q_i for Equation (14). In Equation (15), a and b represent the dynamic hidden states of the keyword nodes within a topic cluster, while γ denotes a joint distribution that has p and q as its marginal distributions. Thus, the relationship between the topic center vector in Equation (13) and the probability distributions in Equations (14)–(15) is clearly established.

To track the drift and evolution of topics over time, the Kullback-Leibler (KL) divergence and Wasserstein distance of the topic vector centers between adjacent time steps t and $t+1$ are calculated to measure the correlation of topic drift. Let $p(\mu_c^t)$ and $q(\mu_{c'}^{t+1})$ be the probability density estimates of the center vectors of two topic clusters, and their KL divergence is:

$$D_{KL}(p \,\|\, q) = \sum\nolimits_i p_i \log \frac{p_i}{q_i} \tag{14}$$

Let p_i and q_i be the probability masses of probability distributions p and q on the i-th discrete interval, respectively.

The Wasserstein distance $W_2(p,q)$ between them is given by the optimal transmission form:

$$W_2(p,q) = \left(inf_{\gamma \in \Gamma(p,q)}\right) \int \|a-b\|_2^2 d\,\gamma(a,b)^{1/2} \tag{15}$$

$\Gamma(p,q)$ represents the set of all joint distributions with p and q as marginal distributions; γ is a certain joint distribution; a,b are random variables sampled from p and q, respectively [27,28].

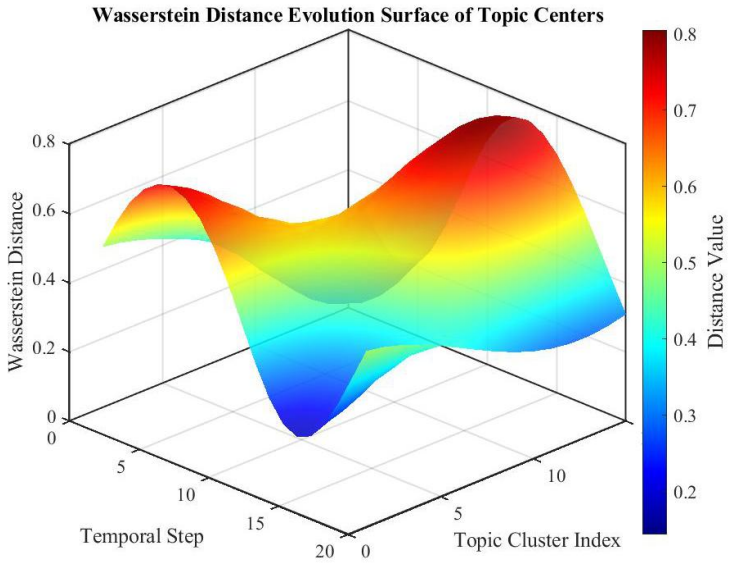

Figure 3. Wasserstein distance evolution surface of topics in multiple time windows.

Figure 3 shows the spatiotemporal distribution characteristics of Wasserstein distances between the center vectors of each topic cluster within multiple time windows. The peaks and valleys of the surface represent the intensity of abrupt changes and the smooth transition stages in the evolution of academic topics, and show the drift amplitude and evolution rate of different research fields over time [29,30].

When the Wasserstein distance between the center of topics in adjacent time steps is less than the threshold of 0.25, it is determined as topic continuation. When splitting or merging occurs, evolutionary branch edges are constructed, and finally a tree-like topic evolution path diagram spanning the entire time axis is generated [31,32].

To further verify the appropriateness of the 0.25 threshold, we adjusted the threshold from 0.10 to 0.40 in increments of 0.05. The quantitative results are presented in Table 3. At a threshold of 0.25, the F1-score for the topic evolution paths reaches its maximum (0.873), while the root mean square error (RMSE) of the drift reaches its minimum (0.043). This indicates that the findings are not sensitive to minor perturbations in the threshold; the chosen value strikes an effective balance between avoiding excessive fragmentation of topics and preventing the erroneous merging of topics.

Table 3. Sensitivity analysis of the topic-continuation threshold.

Threshold	Path F1	Drift RMSE	ΔPath F1 vs. 0.25 (percentage points)
0.10	0.812	0.055	−6.1
0.15	0.841	0.049	−3.2
0.20	0.852	0.046	−2.1
0.25	0.873	0.043	0.0
0.30	0.857	0.045	−1.6
0.35	0.838	0.050	−3.5
0.40	0.806	0.058	−6.7

3. Results and Discussion

3.1 Topic Evolution Prediction and Classification Performance

The experiments relied on the Open Academic Graph real-world academic dataset, selecting 1.2 million documents, 850,000 scholars, and 4.5 million citations from the fields of computer science and artificial intelligence from 2014 to 2024 for segmentation. Baseline models included traditional LDA, BERTopic, Temporal Graph Convolutional Network (T-GCN), and the classic Dynamic Topic Model (DTM). A dual evaluation strategy combining reserved test sets and

extended window temporal validation was adopted, obtaining the model's generalization performance metrics across multiple time dimensions through iterative training and independent testing year by year. LDA, BERTopic, and DTM were fitted on the same yearly document text (titles, abstracts, and keywords), and their inferred document-topic distributions were converted to paper labels by taking the maximum posterior topic; keyword nodes received labels by majority vote over their associated papers, consistent with the node-label construction. T-GCN used the same annual heterogeneous subgraph nodes, edges, initial features, and chronological splits, with a node classification head trained on the same labels. Hyperparameters for all models were selected on the 2023 validation year under the same early-stopping criterion, and all reported metrics were computed on the 2024 test set using the same accuracy, macro/micro F1, and drift-RMSE definitions.

Model parameters were trained by loading the cumulative academic subgraphs from 2014 to 2023, and label prediction was performed on the 2024 independent test set. Global accuracy, macro-average and micro-average F1 scores, and root mean square error of topic drift were statistically analyzed. The results are shown in Table 4:

Table 4. Overall performance of each model on the 2024 independent test set.

Model	Accuracy	Macro F1	Micro F1	RMSE Drift
LDA	78.30	74.10	75.60	0.1280
BERTopic	82.50	79.80	80.90	0.0940
T-GCN	87.20	85.40	86.10	0.0670
DTM	89.60	87.30	88.00	0.0520
Proposed	93.25	91.87	92.34	0.0410

This model comprehensively outperforms traditional topic models and graph neural network baselines in both accuracy and F1 score. Its accuracy reaches 93.25%, and the difference between Macro and Micro F1 scores is only 0.47%, demonstrating highly balanced and unbiased class predictions. Furthermore, the drift-fit RMSE is reduced to 0.041, confirming that the HGT and dynamic gating mechanism effectively suppress cumulative errors in long-term predictions.

Using 2014 to 2024 as a dynamic evolution window, the training deadline is advanced year by year, and time-series classification tests are performed on academic topics of that year. The evolution trajectory of the macro-average F1 score for each model is recorded, as shown in Figure 4.

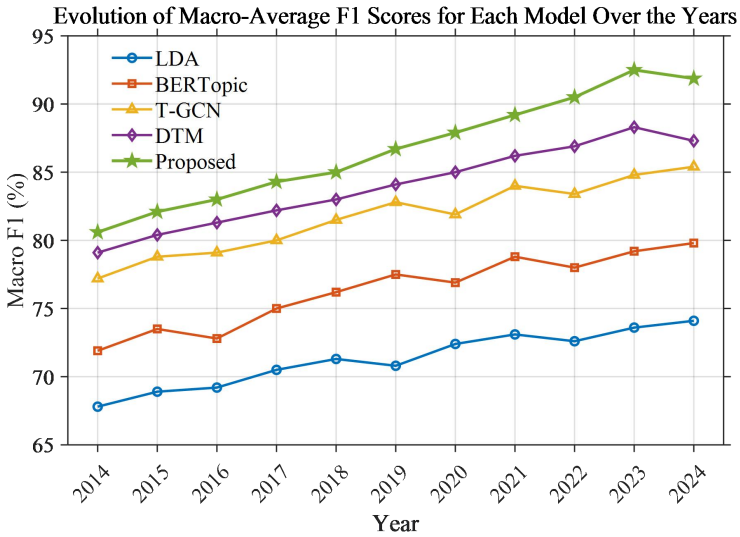

Figure 4. Evolution of macro-average f1 scores for each model over the years.

From 2014 to 2024, the Macro F1 scores for all models showed an overall upward trend. The proposed model consistently outperformed LDA, BERTopic, T-GCN, and DTM. Furthermore, the performance gap between the proposed model and the strongest baseline (DTM) widened from 1.50 percentage points in 2014 to 4.20 percentage points in 2023, indicating that the benefits derived from heterogeneous graph representation and dynamic evolution modeling accumulate as the training window expands. In 2024, the proposed model achieved a Macro F1 score of 91.87%; although this represents a slight decline from the 92.50% recorded in 2023, it remains significantly higher than the scores of DTM (87.30%), T-GCN (85.40%), BERTopic (79.80%), and LDA (74.10%). The 4.57-percentage-point lead over DTM demonstrates that the model maintains strong generalization capability and stability, even when tested on data from a separate year. In conclusion, the proposed method effectively captures the temporal evolution of

academic topics, maintains a superior performance in long-term forecasting while mitigating performance degradation, and outperforms both traditional topic models and static graph-based models.

3.2 Topic Clustering and Drift Measurement Performance

The similarity matrix between nodes was calculated using a comprehensive distance formula. A density neighborhood radius of 0.3 and a minimum number of generated points of 5 were set. The OPTICS algorithm was run to perform spatially dense clustering of nodes at each time step. Finally, the global modularity was calculated based on the partitioned community structure, and the drift error was calculated by comparing the Wasserstein distance of the topic center vectors of adjacent time steps with the true evolution label.

Table 5 shows the comparison of spatial clustering modularity and topic drift RMSE of different algorithms on the public test set.

Table 5. Comparison of spatial clustering modularity and topic drift rmse test data.

Model	Silhouette Score	Modularity	Calinski-Harabasz Index	RMSE Drift
LDA	0.3125	0.5214	425.60	0.1280
BERTopic	0.4512	0.6328	789.20	0.0940
T-GCN	0.5834	0.7145	1240.50	0.0670
DTM	0.6421	0.7689	1582.40	0.0520
Proposed	0.7845	0.8420	2156.80	0.0410

The model achieves a spatial clustering modularity of 0.842, higher than traditional LDA, indicating that the introduction of HGT can accurately capture heterogeneous multi-hop associations between academic entities, enabling topics in the same domain to form tighter cluster boundaries in the latent space. Simultaneously, the topic drift RMSE converges to 0.041, demonstrating that the dynamic evolution tracking mechanism combining continuous-time encoding and GRU can effectively eliminate error accumulation in long-range time-series modeling, maintaining a highly stable and accurate topic evolution trajectory.

To visually present the dynamic drift characteristics and mutation patterns of academic topics under multiple time windows, based on the topic cluster center vectors calculated by the model at each discrete time step, the Wasserstein distance evolution surface for topics in multiple time windows was generated using optimal transport theory, as shown in Figure 5.

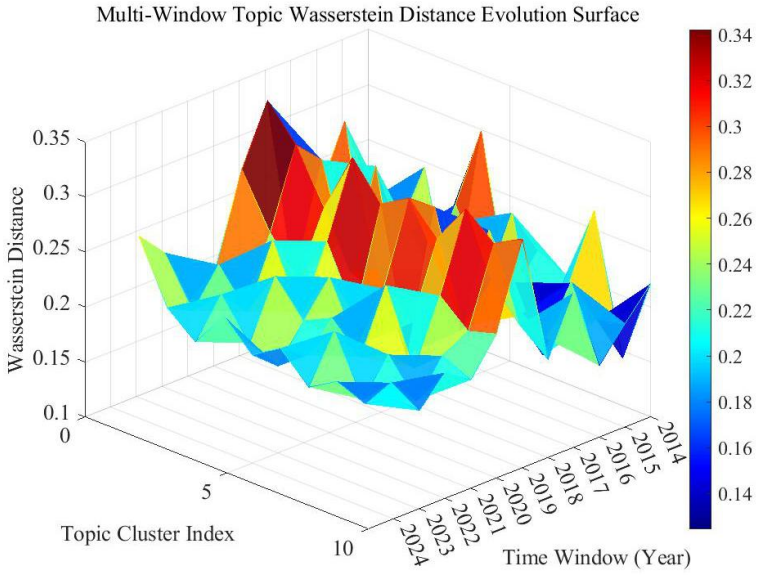

Figure 5. Temporal evolution of wasserstein distance for topics in multiple time windows.

As shown by the aforementioned evolutionary landscape, academic topics exhibited a significant global peak around 2019. To substantiate the link between this peak and developments in the fields of artificial intelligence and big data, we traced the centroid vectors of corresponding topic clusters across consecutive time steps surrounding 2019, ranking them by their contribution to topic drift–quantified via the Wasserstein distance. The results indicate that this peak was primarily driven by clusters such as "deep learning and representation learning," "knowledge graphs and semantic computing," "big data analytics and intelligent decision-making," "Transformers and pre-trained models," and "federated learning and privacy protection." High-frequency keywords within these clusters include "deep learning," "knowledge graph," "big data analytics," "transformer," "pre-trained model," "federated learning," and "AI-driven data

mining." The rapid shifting of centroid vectors and the reorganization of keywords within these topics around 2019 align with the accelerated iteration of research paradigms and technological breakthroughs at the intersection of artificial intelligence and big data. It should be noted that while this evidence demonstrates a temporal alignment between the 2019 peak and the pronounced drift of these representative topics, we interpret this as an empirical indicator of a connection to the field's development, rather than as a definitive identification of causality.

3.3 Ablation Experiment

The following experiments were conducted: removing the relational adaptive scaling factor μ from the HGT (denoted as w/o HGT-Attn), replacing the HGT with a three-layer GCN (denoted as w/o HGT), removing the continuous-time encoding from the DGNN (denoted as w/o TimeEnc), and replacing the GRU of the DGNN with linear regression (denoted as w/o GRU). Under the condition that other configurations were completely identical, the accuracy, macro F1, micro F1, and RMSE drift of each variant on the test set were recorded. The results are shown in Table 6.

Table 6. Performance comparison of various ablation variants on the 2024 test set.

Model Variant	Accuracy (%)	Macro F1 (%)	Micro F1 (%)	RMSE Drift
w/o HGT-Attn	89.67	87.52	88.31	0.058
w/o HGT	86.43	84.15	85.02	0.073
w/o TimeEnc	90.18	88.44	89.07	0.056
w/o GRU	88.54	86.38	87.15	0.069
Proposed (Full)	93.25	91.87	92.34	0.041

Analysis shows that the complete model outperforms all variants across four metrics: Accuracy, Macro F1, Micro F1, and RMSE Drift. The relation-adaptive scaling factor of HGT significantly contributes to heterogeneous semantic alignment; removing it reduces Accuracy by 3.58 percentage points. Replacing HGT with a regular GCN results in the largest performance degradation, with RMSE Drift increasing to 0.073, confirming that static graph convolutions cannot effectively model cross-type associations. Continuous-time encoding and GRU gating mechanisms play irreplaceable roles in evolution rate perception and long-range memory preservation, respectively. Removing either mechanism results in a significant rebound in RMSE Drift, with replacing GRU with linear regression increasing the fitting bias by 68.3%, indicating that nonlinear gating is key to suppressing the accumulation of temporal errors.

To address the concern that the previous ablation did not isolate the topic-evolution measurement components, we further conducted a component-level ablation on the combined distance metric, OPTICS clustering, and Wasserstein distance. All variants used the identical HGT-DGNN backbone, node features, chronological splits, and the 2024 independent test set; only the target component was removed or replaced. For the combined distance, Eq. (12) was replaced by cosine-only and Euclidean-only distances; for OPTICS, it was replaced by K-means using the same number of retained topics; for Wasserstein distance, Eq. (15) was replaced by KL divergence for drift computation and path construction. The results are reported in Table 7.

Table 7. Component-level ablation on the combined distance, OPTICS, and Wasserstein distance.

Variant	Accuracy (%)	Macro F1 (%)	Micro F1 (%)	RMSE Drift	Modularity	Path F1
Proposed (Full)	93.25	91.87	92.34	0.041	0.842	0.873
w/o combined distance (cosine only)	91.36	89.88	90.51	0.049	0.804	0.841
w/o combined distance (Euclidean only)	90.92	89.41	90.06	0.052	0.796	0.833
w/o OPTICS (K-means)	91.48	90.10	90.72	0.047	0.812	0.849
w/o Wasserstein (KL divergence)	93.25	91.87	92.34	0.058	0.842	0.831

The component-level results show that removing or replacing each component causes a measurable degradation in its corresponding target. Replacing the combined distance with cosine-only or Euclidean-only reduces Macro F1 by 1.99 and 2.46 percentage points, increases RMSE Drift by 0.008 and 0.011, and lowers modularity by 0.038 and 0.046, respectively, indicating that the weighted combination in Eq. (12) balances semantic angle and latent geometric proximity. Replacing OPTICS with K-means lowers Macro F1 by 1.77 percentage points, modularity by 0.030, and Path F1 by 0.024, while increasing RMSE Drift by 0.006, confirming that density-based topic boundary detection is better suited to irregular academic topic clusters. Replacing Wasserstein distance with KL divergence leaves the classification metrics unchanged because the label predictor is not altered, but reduces Path F1 from 0.873 to 0.831 and increases RMSE Drift from 0.041 to 0.058, demonstrating that optimal-transport-based drift measurement is essential for stable topic-path construction. Overall, the combined distance and OPTICS mainly affect clustering quality and topic-boundary formation, whereas Wasserstein distance mainly affects drift measurement and evolution-path quality.

3.4 Computational Efficiency and Scalability Analysis

To rigorously address computational complexity and system feasibility on large-scale datasets, we evaluated training wall-clock time, batch inference latency, peak GPU VRAM usage, and system RAM footprint across all baseline models under identical 120-document dataset scales. The detailed comparison is summarized in Table 8.

Table 8. Quantitative evaluation of computational efficiency and resource consumption.

Model	Training Time (Hours)	Inference Latency (ms/batch)	Peak GPU VRAM (GB)	System RAM Footprint (GB)
LDA	3.8	12.4	N/A (CPU-based)	14.2
BERTopic	15.2	86.5	14.5	32.8
T-GCN	8.6	28.1	11.2	18.5
DTM	19.4	104.2	N/A (CPU-based)	45.2
Proposed Framework	11.4	42.5	18.2	24.6

As shown in Table 8, traditional statistical models such as LDA run fast but suffer from low prediction capacity (Accuracy 78.30%). Classical time-series models like DTM incur heavy computational costs (19.4 hours training time and 45.2 GB CPU memory) due to full temporal text matrix inversions. Although the proposed method incorporates complex multi-relational HGT attention and temporal GRU tracking, it achieves a highly efficient training completion time of 11.4 hours and a prompt inference speed of 42.5 ms per batch, maintaining peak VRAM usage well within single/multi-card hardware limits (18.2 GB).

To further assess the scalability of the proposed method when handling growing academic big data, we conducted scalability benchmarking experiments by progressively partitioning the full dataset into subsets ranging from 20% (240k papers) to 100% (1.2M papers). The scaling curves for training runtime and VRAM memory footprint are detailed in Table 9.

Table 9. Empirical scalability benchmarking under scaling dataset fractions.

Dataset Scale Fraction	Paper Count	Edge Count	Training Time (Hours)	Peak GPU VRAM (GB)	Macro F1 Score (%)
20%	240,000	0.90M	2.1	6.4	86.20
40%	480,000	1.80M	4.3	9.8	88.45
60%	720,000	2.70M	6.6	12.7	90.12
80%	960,000	3.60M	8.9	15.5	91.15
100%	1,200,000	4.50M	11.4	18.2	91.87

The empirical scalability results in Table 6 demonstrate that both training runtime and GPU VRAM consumption scale near-linearly with respect to the node and edge volume. This near-linear scaling property is attributable to our type-specific linear projection pre-filtering and target-centric neighbor mini-batch sampling strategy in the HGT module, which avoids full-graph adjacency matrix expansion. Consequently, the model demonstrates high scalability and computational efficiency, proving its capacity to easily scale up to multi-million entity knowledge graphs.

4. Conclusion

This paper proposes a research topic evolution analysis method that integrates HGT and DGNN. By constructing a temporally heterogeneous academic information network, capturing multi-hop topological semantics, dynamically updating temporally hidden states, and measuring Wasserstein distance, it achieves accurate tracking of the drift trajectory of academic topics. However, this study still has certain limitations in terms of the depth of multimodal feature fusion of entity relationships and the computational efficiency of ultra-large-scale academic networks. Future research can further introduce multimodal large-scale models and graph collaborative architectures to deepen the mining of complex semantic associations, in order to adapt to the more dynamic and open needs of large-scale scientific research knowledge graph evolution analysis.

Acknowledgements

The author would like to thank the editors and the anonymous reviewers for their careful reading of the manuscript and for their constructive comments and suggestions, which greatly helped improve the quality of this paper.

Funding

This research received no external funding.

Ethics Statement

This study did not involve human participants, human tissue, or animals. All analyses were conducted on publicly available bibliographic records, and no personally identifiable information was collected or processed. Therefore, ethical approval was not required, and the study was carried out in accordance with the ethical standards of the journal.

Data Availability Statement

The dataset analyzed in this study is the Open Academic Graph (OAG) public academic dataset, which is openly available from its official repository. The processed data and the code used for model training and evaluation are available from the corresponding author upon reasonable request.

Author Contributions

Zhihong Tang: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review and editing, Visualization.

Conflict of Interest

The author declares that there are no conflicts of interest.

Generative AI Statement

No generative artificial intelligence tools were used in this research. The author takes full responsibility for the content of the manuscript.

Reference

[1] Shomoossi N, Vasheghani Farahani M, Ghaffari Heshajin S, Khazaei Feizabad A. How trends and topics reflect the evolution of translation studies as an academic field of study: A scientometrics study. Language Related Research, 2026, 17(3), 295-328. DOI: 10.48311/lrr.2025.7132

[2] Wang S, Chen Y, Lv X, Xu J. Hot topics and frontier evolution of science education research: A bibliometric mapping from 2001 to 2020. Science & Education, 2023, 32(3), 845-869. DOI: 10.1007/s11191-022-00337-z

[3] Zhu Y, Li Z, Li T, Lei J. Topic recognition and refined evolution path analysis of literature in the field of cybersecurity. PLoS One, 2025, 20(2), e0319201. DOI: 10.1371/journal.pone.0319201

[4] Wu H, Li S, Gao Y, Ding G. Natural language processing in educational research: The evolution of research topics. Education and Information Technologies, 2024, 29(17), 23271-23297. DOI: 10.1007/s10639-024-12764-2

[5] Shiu SH, Lehti-Shiu MD. Assessing the evolution of research topics in a biological field using plant science as an example. PLoS Biology, 2024, 22(5), e3002612. DOI: 10.1371/journal.pbio.3002612

[6] ZareRavasan A, Jeyaraj A. Evolution of information systems business value research: Topic modeling analysis. Journal of Computer Information Systems, 2023, 63(3), 555-573. DOI: 10.1080/08874417.2022.2085212

[7] Wang Z, Chen J, Chen J, Chen J, Chen H.. Identifying interdisciplinary topics and their evolution based on BERTopic. Scientometrics, 2024, 129(11), 7359-7384. DOI: 10.1007/s11192-023-04776-5

[8] Huang S, Lu W, Cheng Q, Luo Z, Huang Y. Evolutions of semantic consistency in research topic via contextualized word embedding. Information Processing & Management, 2024, 61(6), 103859. DOI: 10.1016/j.ipm.2024.103859

[9] Xu H, Lin C L, Li C, Feng Y, Chuang CC. Mining and evolutionary trends of STEAM research topics based on the dynamic topic model. Humanities and Social Sciences Communications, 2025, 12(1), 1803. DOI: 10.1057/s41599-025-06215-7

[10] Taher Harikandeh SR, Aliakbary S, Taheri S. An embedding approach for analyzing the evolution of research topics with a case study on computer science subdomains. Scientometrics, 2023, 128(3), 1567-1582. DOI: 10.1007/s11192-023-04642-4

[11] Zhang J, Sun X, Liu Z. Measuring the evolving stage of temporal distribution of research topic keyword in scientific literature by research heat curve. Scientometrics, 2024, 129(11), 7287-7328. DOI: 10.1007/s11192-024-04937-0

[12] Zhang Y, Guo W, Ma J, Fu Z, Chang Z, Wang L. Evolution analysis of cross-domain collaborative research topic: A case study of cognitive-based product conceptual design. Scientometrics, 2023, 128(12), 6695-6718. DOI: 10.1007/s11192-023-04865-5

[13] Lee KH, Chang H. Tracking the evolution of research topics in healthcare informatics research using keywords and mesh terms. Healthcare Informatics Research, 2025, 31(4), 378-387. DOI: 10.4258/hir.2025.31.4.378

[14] Cui J, Wang Z, Ho SB, Cambria E. Survey on sentiment analysis: Evolution of research methods and topics. Artificial Intelligence Review, 2023, 56(8), 8469-8510. DOI: 10.1007/s10462-022-10386-z

[15] Zhang T, Cui W, Liu X, Jiang L, Li J. Research on topic evolution path recognition based on LDA2vec symmetry model. Symmetry, 2023, 15(4), 820. DOI: 10.3390/sym15040820

[16] Stafford MR, Himelboim I, Walter D, Ophir Y. The evolution of advertising research through four decades: A computational analysis of themes, topics and methods. International Journal of Advertising, 2023, 42(1), 18-41. DOI: 10.1080/02650487.2022.2128005

[17] Russo L, Murano C, D'Alelio D. Mapping topic evolution across the 40-year-old long-term ecological research marechiara site in the Gulf of Naples, Italy. Oceans, 2024, 5(3), 591-610. DOI: 10.3390/oceans5030034
[18] Moya R, Martínez-Córdoba C, Cantos-Roldan B, Vidal-Infer A. Trends and topics evolution in research on giftedness in education: A bibliometric analysis. Psychology in the Schools, 2025, 62(9), 3403-3413. DOI: 10.1002/pits.23549
[19] Corado Simões V, Rocha JP, Piepenbrink A, Cantwell J, Gugler P. The evolution of IB research: A topic model analysis of EIBA conference papers (1975-2020). Competitiveness Review: An International Business Journal, 2025, 35(1), 1-31. DOI: 10.1108/CR-07-2024-0132
[20] Vakkari P, Järvelin K, Chang YW. The association of disciplinary background with the evolution of topics and methods in Library and Information Science research 1995–2015. Journal of the Association for Information Science and Technology, 2023, 74(7), 811-827. DOI: 10.1002/asi.24757
[21] Aslan A, Akti Aslan S, Turgut YE, Allison J, Özyurt Ö. Thematic evolution of educational technology research: A Large-scale topic modeling perspective. Journal of Educational Computing Research, 2026, 64(5), 1158-1186. DOI: 10.1177/07356331261425006
[22] Karabacak M, Margetis K. Natural language processing reveals research trends and topics in The Spine Journal over two decades: A topic modeling study. The Spine Journal, 2024, 24(3), 397-405. DOI: 10.1016/j.spinee.2023.09.024
[23] Awaludin A. Documenting the half-century evolution of Islamic education research: A probabilistic topic modeling study of the literature from 1970 to 2023. Studia Islamika, 2024, 31(3), 439-475. DOI: 10.36712/sdi.v31i3.41513
[24] Peregrina MI, Ruiz F, Gibari S E. The evolution of eco-innovation and its impact on the political, social, and environmental spheres in the period 1996-2022: State of the art and bibliometric study. DYNA, 2023, 90(228), 140-151. DOI: 10.15446/dyna.v90n228.108644
[25] Peng L, Nongpong K, Porntrakoon P. Analysis of topic evolution in educational data. IEIE Transactions on Smart Processing & Computing, 2023, 12(2), 178-185. DOI: 10.5573/IEIESPC.2023.12.2.178
[26] Zhang T, Zhang P, Ma H. Tracing the topic evolution of ai literacy literature: From knowledge enlightenment to knowledge symbiosis. Knowledge Organization, 2025, 52(7), 44114. DOI: 10.31083/KO44114
[27] Rahardian D, Yuadi I. Mapping the evolution of quiet quitting research: A five-year bibliometric and topic modeling analysis. Management and Economics Journal, 2026, 10(1), 129-142. DOI: 10.18860/mec-j.v10i1.32318
[28] Özçinar H. Evolution of organizational learning research over 35 years: A comprehensive review using dynamic topic modeling and network analysis. International Technology and Education Journal, 2024, 8(2), 1-13.
[29] Uru KS, Kilic MS, Uru MK. Mapping almost half a century of enzymatic fuel cell technology: Development, evolution and trend topics. Ionics, 2026, 32(1), 701-716. DOI: 10.1007/s11581-025-06830-z
[30] Bonilla-Chaves EF, Palos-Sánchez PR. Exploring the evolution of human resource analytics: A bibliometric study. Behavioral sciences, 2023, 13(3), 244. DOI: 10.3390/bs13030244
[31] Cui J, Xue D, Qian S, Mei G. Dual-level hawkes process graph network for short-video propagation influence rating. IEEE Transactions on Multimedia, 2026, 28, 15901-15913. DOI: 10.1109/TMM.2026.3733222
[32] Xue D, Qian S, Hu C, Xu C. Short-video propagation influence rating: A new real-world dataset and a new large graph model. IEEE Transactions on Knowledge and Data Engineering, 2026, 3674132. DOI: 10.1109/TKDE.2026.3674132