ARTICLE Computer Science

OPEN ACCESS

Insights into knowledge evolution based on semantic representation and dynamic visual analytics

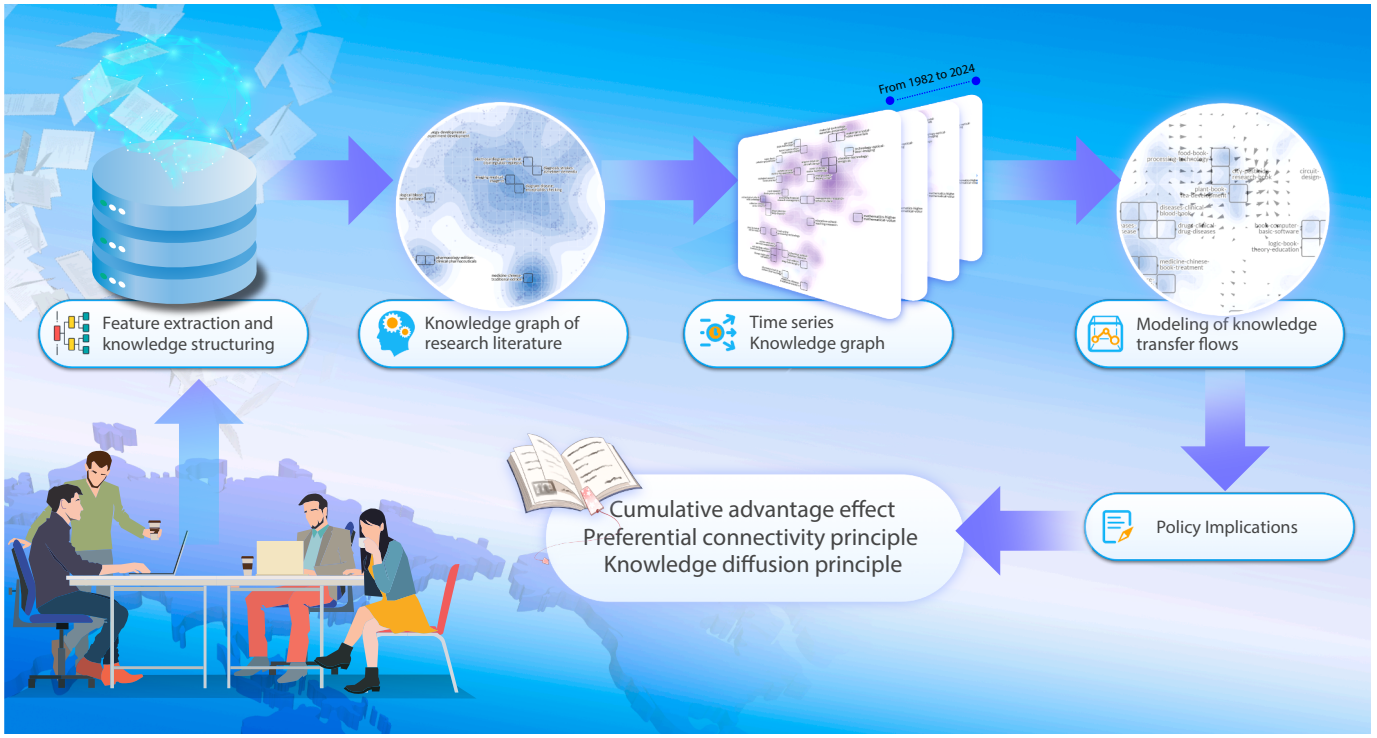
Jieyang Peng,[1,2] Jianing Li,[1] Zhibin Niu,[3] Youzheng Wang,[1] Xiaoming Tao,[1,*] Jivka Ovtcharova,[2] and Jianhua Lu[1]

*Correspondence: taoxm@tsinghua.edu.cn

Received: May 20, 2025; Accepted: November 5, 2025; Published Online: November 21, 2025; https://doi.org/10.1016/j.xinn.2025.101179

www.the-innovation.org

GRAPHICAL ABSTRACT



PUBLIC SUMMARY

- Maps 40k books in semantic space, revealing hotspot migration.
- Interactive KnowFlowViz displays ideas traveling across disciplines.
- Semantic embeddings expose hidden links beyond citation networks.
- Knowledge transfer flow charts rising and fading research themes.

OPEN ACCESS

Computer Science ARTICLE

The Innovation

Insights into knowledge evolution based on semantic representation and dynamic visual analytics

Jieyang Peng,[1,2] Jianing Li,[1] Zhibin Niu,[3] Youzheng Wang,[1] Xiaoming Tao,[1,*] Jivka Ovtcharova,[2] and Jianhua Lu[1]

[1]Department of Electronic Engineering, Tsinghua University, Beijing 100084, P.R. China

[2]Institute for Information Management in Engineering, Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany

[3]College of Intelligence and Computing, Tianjin University, Tianjin 300072, P.R. China

*Correspondence: taoxm@tsinghua.edu.cn

Received: May 20, 2025; Accepted: November 5, 2025; Published Online: November 21, 2025; https://doi.org/10.1016/j.xinn.2025.101179

Citation: Peng J., Li J., Niu Z., et al., (2026). Insights into knowledge evolution based on semantic representation and dynamic visual analytics. The Innovation **7(3)**, 101179.

In the field of knowledge science, understanding the structure and dynamic evolution of knowledge is essential for advancing disciplinary development and anticipating research trends. However, current methodologies lack a unified semantic framework for the structured representation of knowledge, which impedes the quantitative analysis of its evolution and limits the ability to uncover complex relationships among knowledge entities. To bridge these gaps, we propose a structured knowledge representation method based on semantic embedding, enabling a deeper and more consistent understanding of semantic relationships within knowledge units. Building on this foundation, we introduce the concept of knowledge transfer flow to quantitatively analyze and visualize the dynamic evolution of knowledge hotspots over time, revealing the underlying mechanisms that drive knowledge transformation. Furthermore, we develop the KnowFlowViz system, which leverages interactive visual analytics to uncover intricate structural patterns and evolutionary dynamics within knowledge systems, thereby supporting decision-making and guiding future research directions. Our study reveals that established knowledge domains (such as long-standing disciplines) tend to maintain their dominant positions, while newly emerging knowledge entities often preferentially connect with these domains to form interdisciplinary linkages. This phenomenon of advantage accumulation and preferential attachment accelerates the growth and recognition of newcomers. The findings underscore the importance of fostering a more equitable and inclusive knowledge network, and they support the development of policies that nurture emerging disciplines and sustain a diverse, vibrant knowledge ecosystem.

INTRODUCTION

In the era of information explosion, the volume of knowledge generated and accumulated within academic disciplines has reached unprecedented levels. As the boundaries between fields become increasingly blurred and interdisciplinary research flourishes, the challenge of organizing, navigating, and comprehending this vast knowledge landscape poses a significant obstacle for scholars and practitioners.[1] Visualization of structured knowledge, positioned at the intersection of information science, data mining, and human-computer interaction, has emerged as an effective means to represent and interpret complex information in an intuitive manner.[2] By translating intricate knowledge systems into accessible visual forms, researchers can identify patterns, trace formation processes, and uncover migration trajectories of ideas.

Traditional approaches to knowledge representation, typically based on citation networks and bibliometric methods, offer valuable snapshots of knowledge structures but are limited in capturing the dynamic processes of knowledge creation and transformation.[3] Static representations illustrate relationships and hierarchies at a single point in time, yet they fail to convey the temporal dimension of intellectual progress.[4] Understanding the dynamic evolution of knowledge is essential, as it reveals how concepts emerge, diffuse, and interact with existing paradigms. However, citations themselves are not always reliable indicators of intellectual influence, since they may be shaped by author preferences, journal policies, or disciplinary norms.[5] Furthermore, due to the overwhelming volume of publications, no author can cite every relevant work, which often leaves gaps in citation networks and obscures significant intellectual connections, as illustrated in Figure 1A.

To address these challenges, recent studies have explored structured and semantic approaches to knowledge modeling. Abu-Salih[8] provided a survey of domain-specific knowledge graphs, while Zhang et al.[9] proposed frameworks for semi-structured data classification. Deep neural networks have been applied to model hierarchical knowledge, as demonstrated by Peng et al.[10] with hyperbolic neural networks. Wang et al.[11] mine latent semantic signals from a refined knowledge graph via fine-grained attention and contrastive embedding, boosting personalized recommendation. Bali et al.[12] proposed and empirically evaluated a hybrid semantic similarity framework that fuses domain knowledge bases, contextual embeddings, and dynamic synonym repositories. These methods improve representation accuracy but often face issues of scalability, generalizability, or language dependence. Other contributions include Bacciu et al.[13] on graph representation learning, and Xiao et al.[14] on semantic component group.

Research on dynamic knowledge analysis has primarily relied on citation-based and keyword-based approaches. Ye et al.[15] and Zou et al.[16] tracked developments in inventory management and business intelligence through bibliometric indicators, while Li et al.[17] conducted a comprehensive bibliometric and visual analysis of 3,986 green-roof articles to map the field's knowledge structure and identify future research directions. Keyword clustering has been used to reveal research trends, such as in supply chain risk,[18] digital transformation,[19] and concentrating solar power.[20] These approaches can highlight thematic clusters and emerging topics, although their effectiveness is constrained by citation bias or keyword consistency.

Visual analytics[21] has further enriched knowledge exploration by transforming abstract data into interpretable visual forms.[22] Time series visualizations[23] enable chronological mapping of research progress,[24] and digital tools such as ResearchRabbit[25] support interactive exploration, as shown in Figure 1B. Spatial visualizations[26] highlight regional patterns and collaborations,[27] while citation network visualizations[28] remain central in many domains,[29] as illustrated in Figure 1C. These approaches uncover structural and temporal patterns[30] but can be restricted by the scope and quality of available data.[31] Wu et al.[32] used network visualization to analyze research frontiers in AI. Additionally, these methods are effective in providing a macro view of research developments[33] and identifying key research clusters,[34] but they may not always capture the nuanced progressions within smaller fields.[35] The mainstream schemes for knowledge visualization are summarized in Figure 1D.

In summary, current knowledge modeling methods can exhibit significant gaps where pertinent connections are missing, leading to an incomplete knowledge representation of the field, thereby obscuring potential intellectual linkages and advancements. This article aims to bridge the above gap by validating a central hypothesis in knowledge science.

Overall hypothesis: although emerging knowledge appears irregular and diverse, its evolution is not entirely random. The emergence of new ideas and the migration of research hotspots follow discernible patterns that can be identified and modeled.

Specifically, the fundamental premise of our work lies in the recognition that knowledge is inherently structured, with concepts, theories, and methodologies interlinked in complex networks. By constructing the underlying structure of knowledge, we cannot only reveal the intricate patterns of knowledge formation but also trace its migration over time, offering insights into the evolution of scholarly discourse and the emergence of new research frontiers. To achieve the above objectives, we propose three progressively in-depth sub-hypotheses.

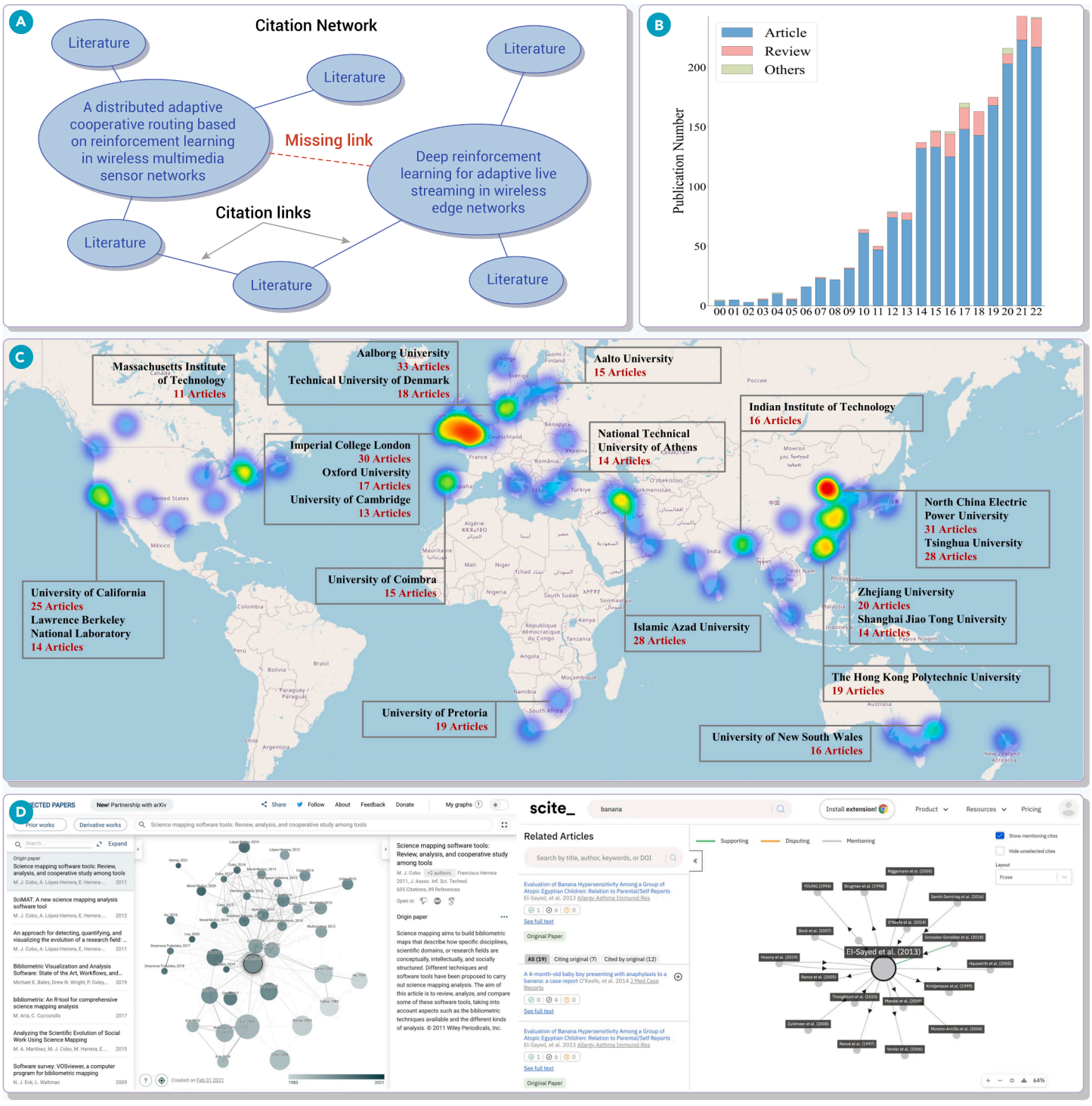


Figure 1. The current visualization method of scientific knowledge networks (A) A typical citation network showing two related publications (2020–2021) that share methodological and topical similarities but lack mutual citations, leading to an incomplete knowledge graph. (B) Time series visualization illustrating the annual publication trends of relevant literature over time. (C) Spatial distribution map highlighting the geographical origins of research outputs. (D) Citation network visualization depicting the interconnections among scientific works based on citation relationships.[6,7]

- Hypothesis 1: irregular and diverse knowledge can be structurally represented. We introduce a semantic-based representation that reveals intrinsic structures beyond citation information.
- Hypothesis 2: the migration of knowledge follows specific patterns. We propose the concept of "knowledge transfer flow" to capture the movement of research hotspots across timescales.
- Hypothesis 3: emerging patterns in knowledge networks are predictable. We conduct empirical studies to examine the mechanisms that govern the integration and growth of new knowledge entities.

MATERIALS AND METHODS

Overall methodology

This study aims to advance our understanding of knowledge dynamics by addressing the structural, temporal, and evaluative dimensions of academic discourse. The research object of this study comprises a diverse range of academic publications, including journal articles, monographs, and other scholarly outputs. The proposed methodology unfolds in three interrelated phases, as illustrated in Figure 2. Leveraging large language models, we begin by extracting two core types of information from the textual content: semantic information, which encompasses key elements such as research subjects, methodologies,

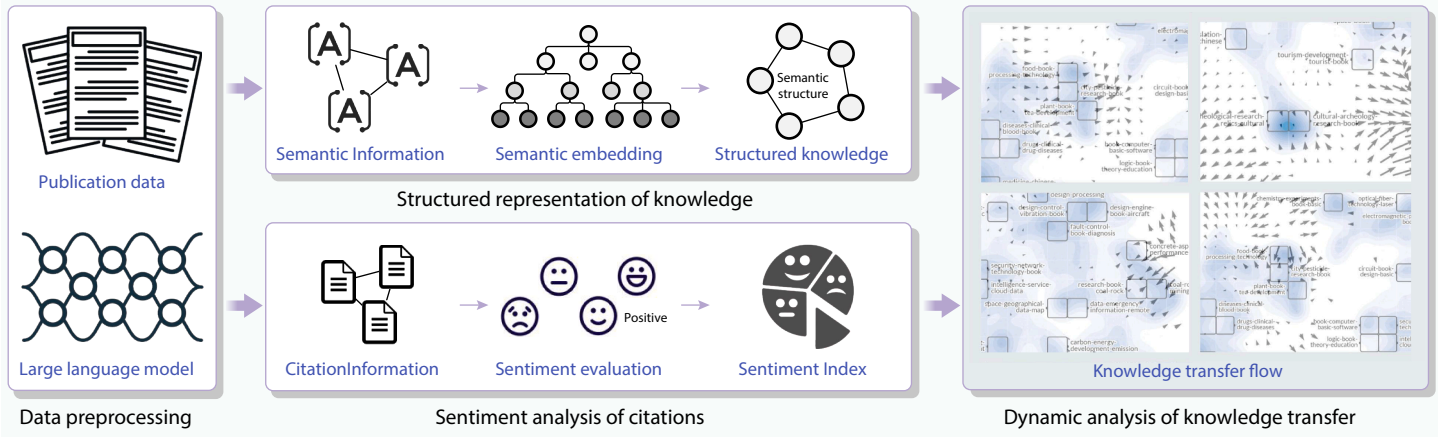


Figure 2. The overall structure of this paper

and intended audiences; and citation information, including both the frequency and contextual usage of citations across temporal scales.

Next, we introduce a hierarchical semantic embedding method to construct the latent structure of knowledge within the corpus. This embedding leverages the previously extracted semantic features to map publications into a structured space that captures deeper conceptual relationships. Simultaneously, we apply a sentiment analysis framework to evaluate the rhetorical stance of citation contexts. This allows us to assess how cited works are received, whether affirmatively, neutrally, or critically, thus complementing citation frequency with qualitative sentiment dynamics. The temporal aggregation of sentiment provides additional insight into how scholarly attitudes toward specific topics evolve over time.

Finally, we design an interactive visualization system "KnowFlowViz" (knowledge flow visualization) to integrate and display the results from the preceding steps. Centered around the concept of "knowledge transfer flow" introduced in hypothesis 2, this interface enables users to explore how research topics migrate, cluster, or dissipate across time. By supporting multi-scale interaction, the system facilitates in-depth exploration of specific knowledge domains, cross-disciplinary comparisons, and the identification of emerging research trends. This visual framework enhances interpretability and offers researchers a dynamic tool for examining the evolution of academic discourse.

Structured representation of knowledge

We collaborated with a major Chinese academic publisher to obtain a dataset of over 40,000 scholarly publications (1982–2024), including monographs, edited volumes, and journal articles.

From each publication, we extracted four semantic features: title, research objective, research method, and target audience (references to supplemental information). Titles capture topical surface information, objectives highlight central problems, methods describe analytical strategies, and audience metadata indicate disciplinary alignment and social reach. Audience classification required a hybrid pipeline combining rule-based heuristics and supervised modeling, assigning publications to researchers, practitioners, or general readers.

Based on Wang et al.'s research,[36] we perform semantic embedding, mapping the four extracted features into distinct low-dimensional spaces where proximity reflects semantic similarity. Each feature type is independently encoded to preserve its semantic nuance, and together they form four complementary semantic landscapes.

Let $X = \{x_1, x_2, \ldots, x_N\}$ be the set of high-dimensional vectors representing the semantic information of the publications. We project these vectors into a 2D space:

$$Y = f_{UMAP}(X) = \{y_1, y_2, \ldots, y_N\} \quad \text{(Equation 1)}$$

where f_{UMAP} is the UMAP function mapping high-dimensional vectors to 2D points. This step reduces the complexity of the data while preserving the global structure, enabling more intuitive visualization and analysis.

Next, we construct a quadtree Q from these 2D points. A quadtree is a tree data structure that recursively partitions a 2D space into four equally sized squares, each represented as a node. Each data point y_i exists in a unique leaf node of the quadtree.

$$Q = quadtree(Y). \quad \text{(Equation 2)}$$

The quadtree structure allows efficient hierarchical partitioning of the data, enabling multi-resolution analysis. To summarize embeddings at different levels of granularity, we traverse the quadtree from the bottom up. In each iteration, we extract summaries of embeddings in each leaf node and merge these leaf nodes with their parent nodes, continuing this process recursively.

Let L_i denote the set of leaf nodes at level i. We define the embedding summary $S(L_i)$ as:

$$S(L_i) = \sum_{j \in L_i} y_j \quad \text{(Equation 3)}$$

where $S(L_i)$ represents the aggregated semantic information at level i. This aggregation step ensures that the semantic relationships are preserved and summarized effectively at various levels of granularity. At each level, we update the summary by aggregating the embeddings from the leaf nodes to their parent nodes:

$$S(L_{i+1}) = \sum_{k \in parent(L_i)} S(L_i) = \sum_{k \in parent(L_i)} \sum_{j \in L_i} y_j. \quad \text{(Equation 4)}$$

This process continues until the entire tree is merged into a single node at the root. This hierarchical aggregation captures the semantic structure of the dataset at multiple levels, providing a comprehensive overview of the knowledge landscape.

For the leaf-level summarization, we adapt the TF-IDF method to a tile-based *TF-IDF*(*t-TF-IDF*) approach. For each leaf node, we merge all documents into a meta-document and compute the TF-IDF scores.

Let D_t be the set of documents in a tile t. The t-TF-IDF score for a term w in tile t is given by:

$$t - TF - IDF(w, t) = tf(w, D_t) \times \log \frac{N}{df(w)} \quad \text{(Equation 5)}$$

where $tf(w, D_t)$ is the term frequency of w in D_t, N is the total number of documents, and $df(w)$ is the document frequency of w across all documents. This hierarchical document merging approach ensures scalability, as the n-gram count matrix is constructed once and updated in each aggregation iteration with a single matrix multiplication.

Finally, we extract keywords with the highest *t-TF-IDF* scores to summarize the embeddings in each leaf node. This process effectively identifies the most representative terms in each semantic region, enhancing the interpretability of the knowledge structure.

Tendency analysis of citations

This section outlines our methodology for analyzing the sentiment of citations[37] to measure the temporal sentiment trajectory of each publication. Our approach involves identifying all the contexts in which a given publication is cited from its year of publication onwards, performing sentiment analysis on these contexts, and calculating the average sentiment per year. By arranging these yearly averages chronologically, we construct a time series representing the sentiment trend for the publication (references to supplemental information).

To perform this sentiment analysis, we utilize the DistilBERT-base-uncased method, which leverages the *DistilBERT* model. *DistilBERT* is designed to be more efficient while maintaining comparable performance.

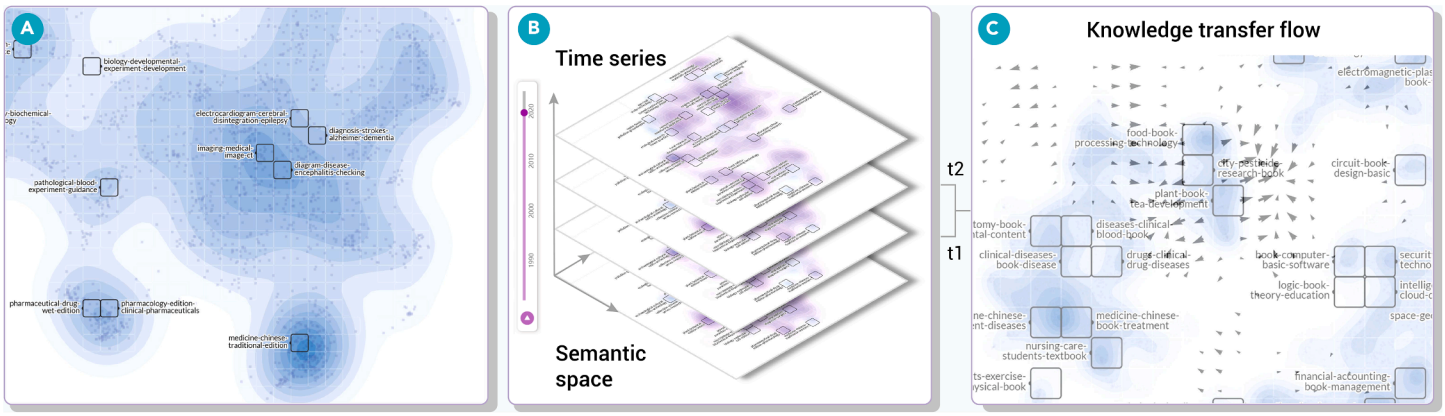


Figure 3. Modeling knowledge dynamics: from semantic landscapes to transfer flows (A) An example of a KDE-based knowledge representation. The points in the figure represent publications. Publications with similar topics are also closer to the location in the semantic space. The contour plots are the results after KDE, and the darker areas represent research hotspots. (B) The principle of knowledge transfer flow: the difference between the potential fields at two specific time periods. (C) Knowledge transfer flow: the direction of the arrow represents the direction of knowledge transfer over time.

First, let $\mathcal{C} = \{c_1, c_2, \ldots, c_M\}$ represent the set of citation contexts for a publication, where each c_i is a text snippet that cites the publication. For each context c_i, we apply the following steps:

DistilBERT is a smaller, faster, and lighter version of BERT, achieved through knowledge distillation. The process involves training a student model $f_{DistilBERT}$ to mimic the behavior of a larger teacher model f_{BERT} by minimizing their prediction differences.[38] The knowledge distillation loss can be expressed as:

$$\mathcal{L}_{distill} = \sum_{i=1}^{M} KL(\sigma(f_{BERT}(c_i)/T) \,\|\, \sigma(f_{DistilBERT}(c_i)/T)) \quad \text{(Equation 6)}$$

where *KL* denotes the Kullback-Leibler divergence, σ represents the softmax function, and *T* is the temperature parameter used to soften the probabilities, which ensures that the distilled model retains the knowledge of the teacher model while being more efficient.

Each citation context c_i undergoes tokenization using the tokenizer from DistilBERT-base-uncased, converting the text into tokens $t = \{t_1, t_2, \ldots, t_n\}$ and adding special tokens [*CLS*] and [*SEP*]:

$$t = tokenizer(c_i). \quad \text{(Equation 7)}$$

Since it is an uncased model, all tokens are converted to lowercase to ensure uniformity in the text data, facilitating consistent processing.

Each token t_j is then mapped to a dense vector through the embedding layer *E*:

$$e_j = E(t_j). \quad \text{(Equation 8)}$$

Positional embeddings p_j are added to the token embeddings to include positional information:

$$h_j^0 = e_j + p_j. \quad \text{(Equation 9)}$$

This encoding process allows the model to capture both the semantic meaning and the position of each token in the context. The token embeddings are passed through multiple transformer layers. For layer *l*, the output h_j^l is:

$$h_j^l = LayerNorm\left(h_j^{l-1} + FFN\left(MultiHeadAtt\left(h_j^{l-1}\right)\right)\right) \quad \text{(Equation 10)}$$

where *MultiHeadAttn* computes the self-attention scores, *FFN* is a feedforward network, and *LayerNorm* is layer normalization. This hierarchical processing captures complex interactions between tokens.

We use the final hidden state of the [*CLS*] token $h_{[CLS]}$ as the representation of the entire sequence:

$$h_{[CLS]} = h_{[CLS]}^L. \quad \text{(Equation 11)}$$

A classification layer is added, producing sentiment scores:

$$y_{sent} = softmax(Wh_{[CLS]} + b). \quad \text{(Equation 12)}$$

The model is fine-tuned on a sentiment analysis dataset to minimize the cross-entropy loss:

$$\mathcal{L}_{sent} = -\sum_{i=1}^{N} y_i \log(y_{sent,i}). \quad \text{(Equation 13)}$$

This fine-tuning adapts Saha et al.,[39] the pre-trained model to the specific task of sentiment analysis, ensuring accurate sentiment predictions.

The sentiment score for each context c_i, denoted s_i, is extracted, where $s_i \in [-1,1]$. We then compute the average sentiment for each year *y*:

$$\bar{s}_y = \frac{\sum_{c_i \in C_y} s_i}{|C_y|} \quad \text{(Equation 14)}$$

where C_y is the set of citation contexts in year *y*. This averaging process aggregates the sentiment scores, providing a yearly sentiment measure.

The resulting time series $\{\bar{s}_y\}$ represents the sentiment trajectory of the publication over time, providing insights into its evolving perception within the academic community. By leveraging the DistilBERT-base-uncased method, we achieve an efficient yet powerful sentiment analysis framework that captures nuanced changes in scholarly sentiment.

Dynamic analysis of knowledge transfer

In the above section, we constructed a semantic structural system[40] of knowledge, i.e., mapping publications into a normalized semantic space. In this section, we quantify the evolution of the knowledge system over time.

When we regard the intensity of the knowledge as the spatial density, the density map can be used to visualize the spatial distribution of publications. In this paper, the density map is obtained by using a kernel density estimation (KDE) method, so that the discrete knowledge can be transformed into a continuous function for further analysis. The method is formalized in the following.

Let $x_1, x_2, \ldots, x_n$ be the discrete geographical locations of publications in semantic space, x_i denoted as a vector $(lon_i, lat_i)^\tau$ of longitude and latitude. For time *t*, the density of a position of x in a density map is defined as follows:

$$\hat{f}_t(x)|_t = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\frac{x - x_i}{h} \quad \text{(Equation 15)}$$

where *n* is the sample size, *h* is the bandwidth of the kernel K_h, which is symmetrical and positive definite. Since the Gaussian kernel can capture data sensitivity of a large spatial area, and it has a lower computation complexity compared with other kernels with exponential functions, it is here used for the implementation. An example of a KDE-based knowledge representation is shown in Figure 3A.

With this method, we can model the distribution of knowledge at different time periods. We observed, that the geospatial knowledge in semantic space fluctuate continuously over time and, thus, the transfer of knowledge is a continuum occupying a simply connected region in the time dimension with an irrotational characteristic. Therefore, the potential flow method[41] can be used to visualize the spatial shift of knowledge flow, which implies the flow between spatially different areas.

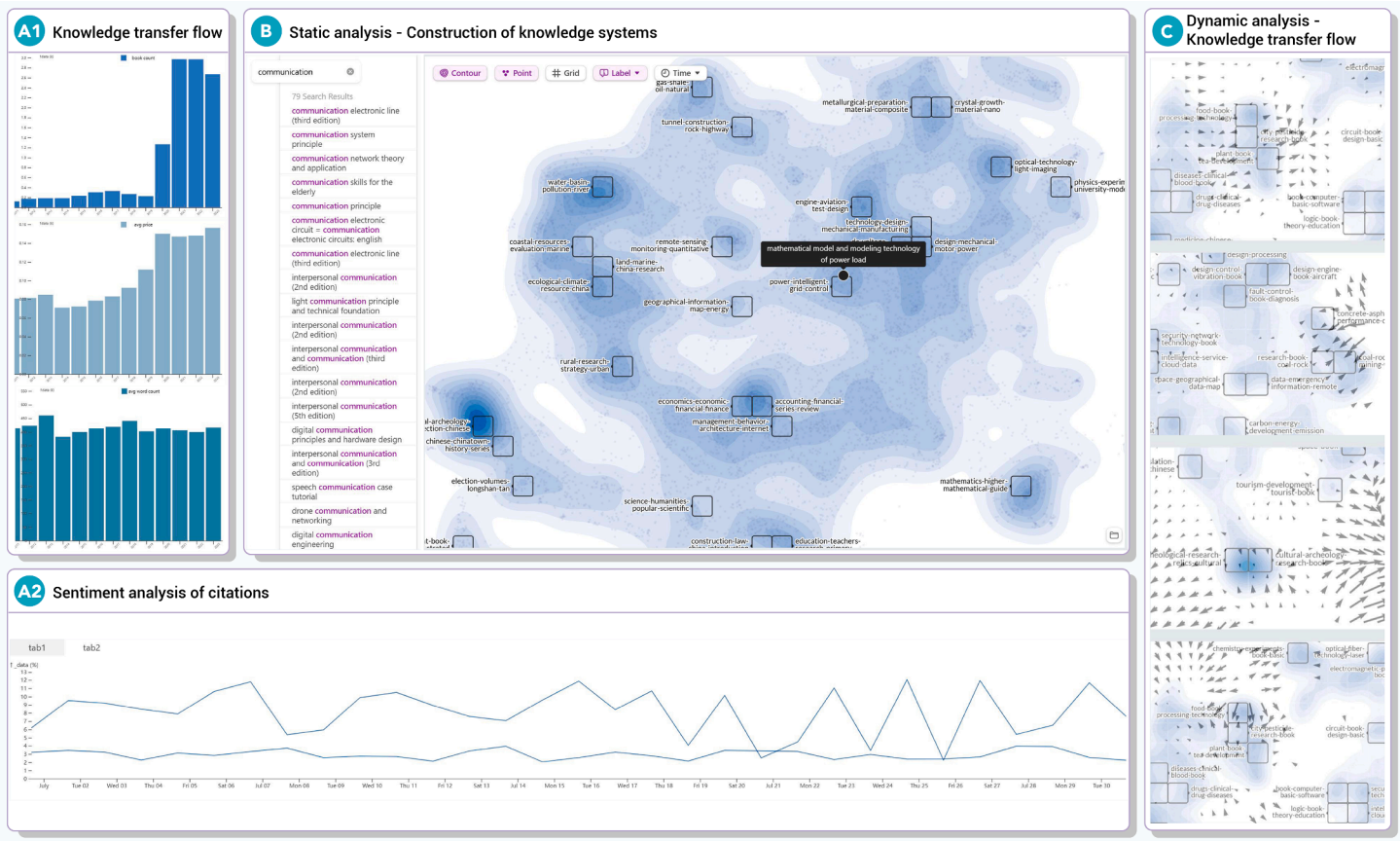


Figure 4. Main interface of the KnowFlowViz system The interface is mainly divided into four modules: (A1) statistical analysis, (A2) sentiment analysis, (B) static analysis,[36] and (C) dynamic analysis.

Formally, the spatiotemporal knowledge-shift is defined as $\nabla shift(x)|_{t_1,t_2}$, which is the gradient of the velocity potential $shift(x)$ that refers to the spatial knowledge fluctuation at the selected moments or periods.

$$shift(x)|_{t_1,t_2} = \widehat{f}_t(x)|_{t_2} - \widehat{f}_t(x)|_{t_1} \quad \text{(Equation 16)}$$

where $\widehat{f}_t(x)$ is the result of KDE in Equation 15, namely the knowledge distribution. x denoted as a vector $(lon, lat)^T$ of longitude and latitude. $t1$ and $t2$ represent two different time periods, respectively. The visual representation of Equation 16 is shown in Figure 3B.

The vector field map in Figure 3C represents the spatiotemporal knowledge-shift model between two specific time periods. The vectors in the vector field map point to the direction of the shift of knowledge, and the magnitude of a vector represents the intensity of the shift. Therefore, this potential flow field can be defined as knowledge transfer flow from the perspective of dynamic knowledge. The significance of the energy demand flow lies in the spatial and temporal distribution of knowledge fluctuations (note: not the distribution of knowledge).

Visual encoding for the KnowFlowViz system

Beyond quantitative indicators such as annual publication volume, cost, and distribution across publishers,[42] it is equally important to reveal the qualitative and semantic connections that shape the evolution of knowledge. In this section, we develop KnowFlowViz with reference to the visualization tool developed by Wang et al.,[36] a visualization system that integrates four interlinked interfaces—statistical analysis, static analysis, dynamic analysis, and sentiment analysis. The main interface of the developed KnowFlowViz system is shown in Figure 4.

Statistical and sentiment analysis. In the statistical analysis interface of KnowFlowViz, we utilize three bar charts to display the annual publication counts, average word counts, and average prices of books published by our collaborating publishers. These bar charts are designed to be interactive, allowing users to control the x axis scale with the mouse wheel and to drag the axis left or right. This design enables users to zoom in on specific years of interest, making it easier to analyze trends and identify significant changes over time.

In the sentiment analysis interface, we use line charts to depict the temporal changes in the sentiment associated with each publication. The sentiment data are derived from comments citing the publication and quantified using the tendency analysis algorithm described in Section "tendency analysis of citations". Each publication is represented by a line, with the x axis showing the time series and the sentiment index ranging from −1 to 1, where 1 indicates a positive sentiment and −1 indicates a negative sentiment. In our sentiment framework, a negative sentiment score—closer to −1 typically reflects language associated with critical remarks. This may include explicit disagreement with findings, concerns about methodological limitations, or broader critiques of the work's assumptions or conclusions. Conversely, scores approaching +1 indicate favorable references, such as support, endorsement, or positive recognition of a publication's contribution.

This visual representation helps users track the reception of publications over time, providing insights into how the perception of specific works evolves. Notably, the sentiment analysis interface is linked with the static analysis interface; when a point is selected in the static analysis view, the corresponding sentiment curve is highlighted in the aentiment analysis view. The integration of these interfaces facilitates a comprehensive analysis, enabling users to correlate semantic proximity with sentiment trends, thereby offering a deeper understanding of the impact and reception of academic works.

Static analysis. The static analysis section of our platform showcases a dual-pane interface, elegantly divided into a search bar panel on the left and a semantic space panel on the right, as depicted in Figures 5A and 5B.

The static analysis interface integrates textual search with visual exploration. In Figure 5A, a search bar allows users to input keywords, instantly retrieving relevant document titles. These documents are simultaneously highlighted in the semantic space on the right, linking queries with their spatial representation. This seamless transition enables users to quickly locate and assess the contextual relevance of results.

The semantic space provides a visual canvas to examine relationships between documents, as presented in Figure 5B. A toolbar in the upper-left corner offers options such as heatmaps, scatterplots, and thematic vocabularies, allowing users to tailor the visualization and uncover clustering, distribution, and thematic trends. An icon in the lower-right corner gives access to semantic dimensions such as document titles, research objects,

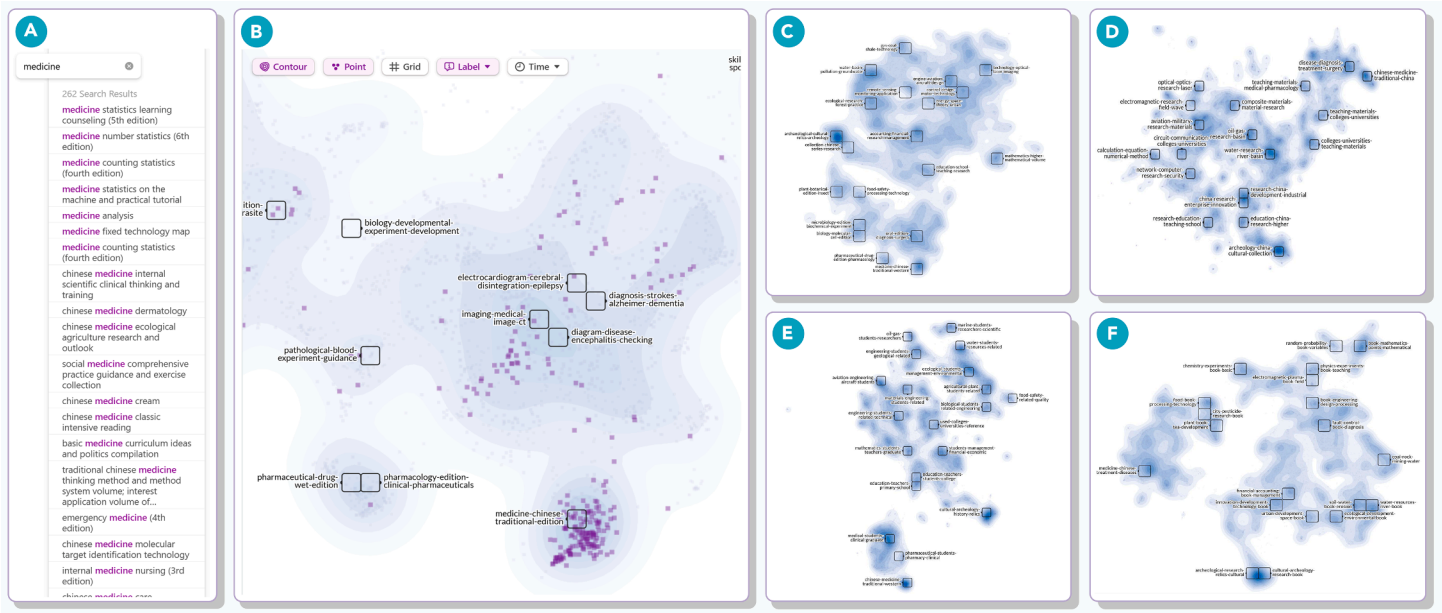


Figure 5. Interactive semantic exploration of literature through multidimensional embeddings (A) Search interface and semantic linking: the left panel features a search bar for keyword-based queries, with results displayed below and simultaneously highlighted in the semantic space (right), bridging textual search and visual exploration. (B) Customizable semantic space: the right-side visualization offers interactive tools to switch between heatmaps, scatterplots, and thematic vocabularies. A dimension selector (bottom right) allows users to explore documents by title, research object, methodology, or audience. (C) Embedding by book name: semantic clustering based on book titles reveals topical groupings and naming conventions across the corpus. (D) Embedding by book subject: documents are grouped by subject matter, highlighting disciplinary clusters and thematic overlaps. (E) Embedding by book outline: structural similarities in content organization drive the spatial layout. (F) Embedding by book reader: user engagement patterns shape the embedding.

methodologies, and target audiences, enabling flexible shifts in analytical perspective. When a different semantic dimension is selected from the corpus of word embeddings, the visualization in Figure 5B dynamically updates to present the corresponding results, as illustrated in Figures 5C–5F.

Dynamic analysis. The dynamic analysis interface visualizes the evolution and migration of knowledge across four 2-year periods from 2020 to 2024. Users can interactively pan and zoom to control detail, while the interface remains linked to static analysis: selections in the semantic space automatically update the migration flows shown here, ensuring a coherent analytical perspective. The flow field presented in this interface is a direct visualization of the knowledge transfer flow defined in Section "dynamic analysis of knowledge transfer", where the flow direction indicates the movement of knowledge within the semantic space, and the flow intensity represents the strength of this migration.

By presenting knowledge flows dynamically, the system reveals patterns, trends, and potential future directions, helping researchers anticipate emerging topics and methodologies. Its interactive design encourages exploration of different scenarios and hypotheses, fostering engagement and deeper understanding of the mechanisms driving knowledge evolution.

RESULTS

Statistical analysis of macro publication data

This section focuses on the publishing landscape of a prominent publisher spanning two and a half decades, from 2000 to 2024. The data comprises three key indicators: the annual number of books published, the average price per book, and the average word count per publication, as shown in Table 1 and Figure 4A.

Firstly, the early years exhibited a gradual increase in the number of books published, reflecting the publisher's growing commitment to expanding its catalog. However, this growth was not linear and showed marked fluctuations. The most dramatic surge occurred between 2013 and 2014, with the annual output rising sharply from 36 to 1,272 titles. This expansion coincided with the intensification of global academic competition for funding, grants, and academic appointments. From the early 2010s onward, scholarly output, especially in peer-reviewed journals and academic monographs, became a primary metric for evaluating researchers and institutions. In response, universities and research bodies began to actively encourage publication, leading to a rapid increase in research outputs. This pressure extended to book publishing, especially in fields such as science, technology, engineering, mathematics, and the social sciences.

Following this period, output declined but surged again during 2022–2023, indicating a cyclical pattern shaped by a range of institutional and societal factors. The recent increase can, in part, be attributed to the effects of the COVID-19 pandemic, which transformed research workflows. The shift to remote work and digital collaboration tools enabled researchers to maintain or even accelerate their output. Simultaneously, publishers adopted more flexible digital production models, further contributing to the observed fluctuations.

In parallel with output trends, the average price per book has shown a steady upward trajectory. This reflects rising production costs as well as the increasing prevalence of specialized academic works, which typically command higher prices due to their focused content and niche readership. The most pronounced price increases have occurred since 2020, likely driven by the broader economic impacts of the pandemic. Inflation, combined with growing expectations for digital formats and enriched content, has contributed to this upward trend.

Interestingly, the average word count per book has remained relatively stable across the observation period, suggesting a consistent editorial approach to content length. It may indicate a balancing act between content depth and production efficiency, ensuring that books remain comprehensive yet manageable in scope.

Revealing patterns between research areas and target audiences

This section explores associations within the research corpus through semantic embedding of scientific literature. The structured knowledge derived from embeddings based on research fields and target audiences is illustrated in Figures 5C and 5F.

As shown in Figures 5C and 5D, research topics are distributed in distinct clusters. Medical and nursing domains, including disease diagnosis, clinical treatment, surgical techniques, and traditional Chinese medicine, are concentrated in the upper right. Educational publications, such as new physics, chemistry, and vocational training, occupy the upper and central areas. Engineering and technology, including electromagnetic studies, aviation materials, and turbine design, dominate the central-left region. Environmental and earth sciences, including water management and geological research, appear on the middle-right, while clusters of social sciences and humanities, such as psychology, economics, history, and cultural studies, emerge at the bottom.

The visualization highlights major hotspots, such as medical and healthcare (deep blue clusters in the upper right) and engineering and technology (central

Table 1. Book publishing data from 2000 to March 2024

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
No. of books	3	9	12	16	20	26	36	47	53	91	152	129	179
Average price	65	69	71	100	89	62	71	67	72	83	75	80	81
Average word count	419,667	473,111	450,417	440,438	515,750	390,342	457,222	431,609	446,547	461,915	416,303	414,287	423,866
Year	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024 (by March)	
No. of books	192	193	239	311	335	274	233	1272	2976	2975	2676	262	
Average price	85	71	72	78	83	92	112	150	147	148	156	133	
Average word count	461,556	383,311	401,101	413,811	419,998	441,737	403,324	413,518	407,262	401,006	417,146	393,615	

deep blue regions). It also reveals connections across domains: electromagnetic engineering, turbine design, and water studies overlap, suggesting technological-environmental linkages, while social sciences and humanities cluster closely, reflecting methodological convergence. Network security, data processing, and computer science form coherent clusters, underscoring shared challenges.

The right panel of Figures 5E and 5F links research outputs to audiences. For example, the upper-left quadrant contains petroleum engineering and aerospace publications for students and professionals, while the center emphasizes education-focused materials for mathematics and teaching audiences. Dense clusters of medical students and clinical trainees highlight the scale of health science education. Marine engineering and water management readers appear in the upper right, while ecological management and agricultural students cluster together in the upper middle. Humanities audiences, such as history and archaeology, form overlapping groups in the lower right.

Based on the structured knowledge obtained from semantic embedding, the following conclusions can be drawn.

- Alignment between research themes and target audiences: semantic embedding reveals a strong correspondence between research content and its intended readers. This alignment enhances the practical relevance of academic publications and educational materials by ensuring that they are developed in accordance with audience needs.
- Identification of emerging interdisciplinary fields: the visualizations indicate the growing presence of interdisciplinary research areas, such as environmental studies and computer science. These fields attract a wide range of readers, suggesting increasing integration across traditional academic boundaries.
- Emergence of data-centric research: the concentration of topics including network security, data processing, and computational methods highlights the expanding role of data-driven approaches in scientific research. This shift reflects a broader transition toward quantitative analysis and algorithmic modeling in addressing complex research challenges.

Revealing potential patterns of knowledge evolution

This section analyzes the knowledge transfer flow over time to identify evolutionary patterns. Figures 6A and 6B displays eight flow maps across four consecutive biennial periods (2020–2024), with the top and bottom rows representing flows based on research content and target audience embeddings, respectively.

An analysis of knowledge transfer between 2020 and 2021 reveals several distinct local features, as shown in Figures 6A and 6B. One of the most notable developments is the rise of urban studies in 2021, reflecting heightened global concern with sustainable urbanization and its associated challenges. Another emerging focus is optical fiber research, driven by advances in optical communication technologies and the growing demand for high-bandwidth data transmission. The expansion of this field illustrates the broader technological trajectory and the importance of fiber optics in modern communication infrastructure. Clinical research remains highly active but has shifted its focus toward specialized topics such as drug-disease interactions and blood-related disorders, signaling a refinement of priorities and a stronger emphasis on targeted medical challenges.

Beyond the sciences, Figure 6C shows the migration of knowledge in the humanities, particularly in archaeology and cultural heritage. Between 2021 and 2022, this area received strong academic attention, partly stimulated by national policies such as China's "Cultural Heritage Protection Project." From 2022 to 2023, research in this domain continued to expand but at a slower pace, as academic resources were increasingly redirected to fields prioritized in the "14th Five-Year Plan," including technological innovation and public health.

From a broader perspective, the heatmap and streamline visualizations spanning 2020 to 2024 reveal consistent patterns of knowledge migration across domains and audiences. In the period 2020 to 2021, traditional disciplines such as mathematics, physics, and computer science acted as central hubs, with knowledge increasingly flowing into artificial intelligence and machine learning. By 2021 to 2022, bioinformatics and healthcare gained prominence, reflecting stronger integration of computational methods with biological and medical research. The trend extended into 2022 to 2023, when sustainability and climate-related research became central, indicating a growing role of science in addressing environmental and societal challenges. In 2023 to 2024, sustainability and climate science maintained their influence, while education and the social sciences attracted greater attention, demonstrating a shift toward more comprehensive responses to multifaceted societal problems.

Parallel to these disciplinary shifts, the target audiences of academic research also evolved. In 2020 to 2021, scholarly communication was directed primarily toward researchers and academics, consistent with traditional incentive structures centered on citations and journal metrics. By 2021 to 2022, publications increasingly addressed professionals outside academia, illustrating a growing interest in translational research and practical applications. The years 2022–2023 witnessed further expansion of audiences to include the general public and media, marking a greater emphasis on science communication and public engagement. By 2023 to 2024, students and educators had also become important recipients of knowledge transfer, reflecting the rising importance of education and the dissemination of research to younger generations.

Building on the above empirical findings, we distilled several underlying principles that characterize the mechanisms driving knowledge evolution. These principles emerge directly from the observed dynamics in publication patterns, network structures, and diffusion pathways.

- Cumulative advantage effect: this principle extends the concept of the Matthew effect to the context of knowledge evolution. It suggests that research domains or topics with an early advantage in resources or attention are more likely to attract continued interest and investment. Such areas benefit from existing infrastructure, institutional support, and greater visibility, which reinforce their leading position. This self-reinforcing dynamic contributes to the uneven growth of research domains and the persistence of structural inequalities within the academic landscape.
- Preferential connectivity principle: originating from network theory, this principle explains the tendency of new research topics or communities to establish connections with already prominent domains. In knowledge networks, emerging disciplines often seek associations

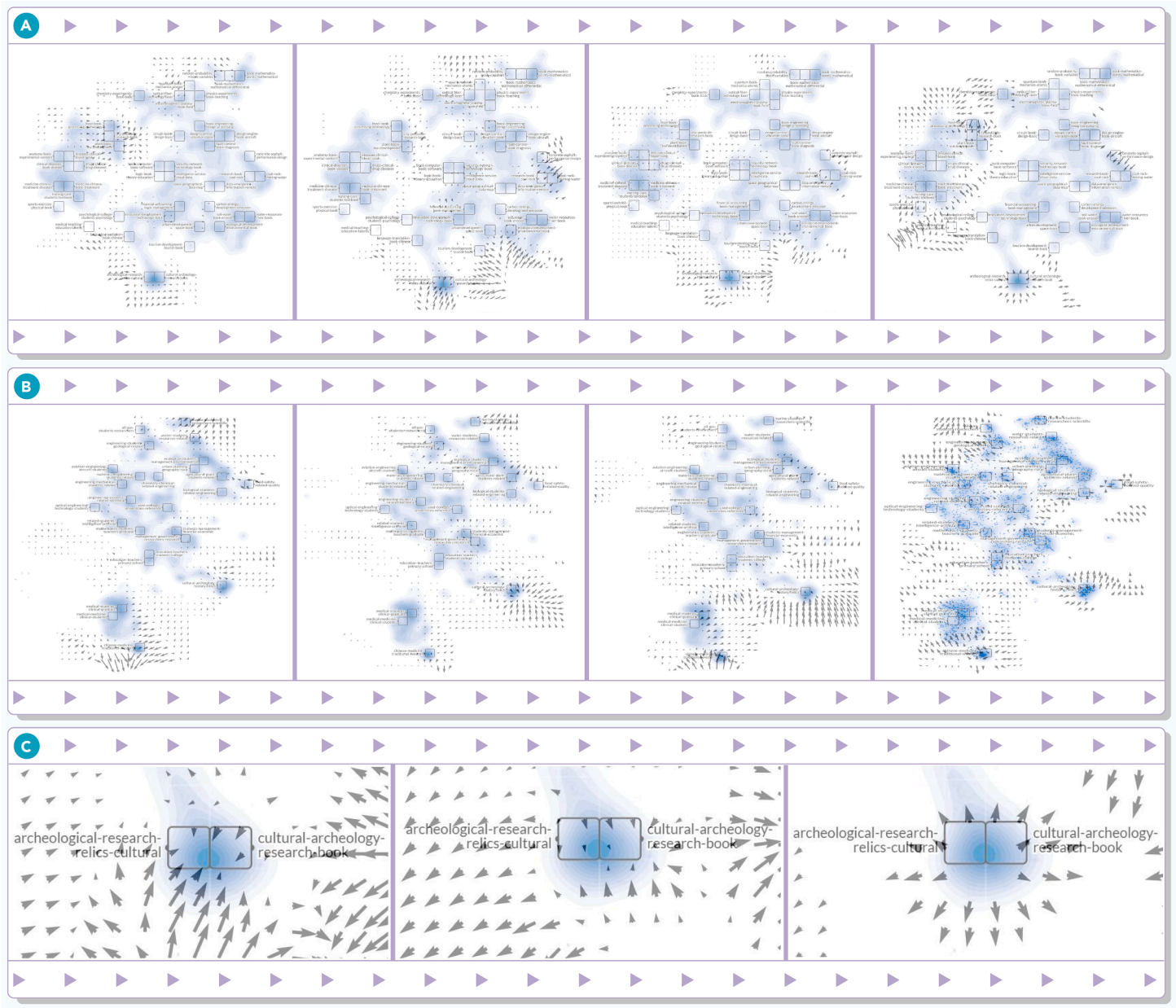


Figure 6. Empirical visualization of knowledge transfer flows (A) Evolution of research content over time, highlighting thematic shifts and emerging topics. (B) Evolution of target readership over time, illustrating how audience focus has changed across periods. (C) Knowledge flow trends in the humanities field from 2021 to 2024, showing the direction and intensity of idea diffusion.

with well-established fields in order to gain credibility and accelerate recognition. This behavior results in a network structure where highly connected nodes become increasingly central, reinforcing existing hierarchies and influencing the trajectory of knowledge diffusion.

- Knowledge diffusion principle: the spread of knowledge across domains is governed by a range of factors, including communication channels, institutional structures, and the configuration of knowledge networks. Diffusion is rarely linear or uniform. Instead, it follows dynamic and context-specific patterns, shaped by both internal drivers—such as methodological innovations—and external pressures—such as policy shifts or societal demands. These complex dynamics are reflected in the observed migration flows across different fields and audiences over time.

DISCUSSION

The findings of this study reveal a close correspondence between research themes and their intended audiences. This observation suggests that knowledge production is not only shaped by disciplinary interests but also by the expectations of specific readership groups. For example, in medicine and engineering, the alignment between scholarly outputs and practitioners' needs enhances the translational value of research. This demonstrates that publishing patterns can be understood as adaptive responses to user demands, which in turn affect the direction of future research. For policy makers, this implies that introducing audience-oriented metadata in publishing workflows could provide a systematic basis for such recognition, while research agencies might incorporate indicators of audience translation into funding assessments.

Building on this, the rise of interdisciplinary domains such as environmental science, data science, and sustainability illustrates the permeability of disciplinary boundaries. The convergence of these fields responds directly to global challenges such as climate change and digital transformation. Their growth highlights how pressing societal issues function as external drivers that restructure knowledge networks and attract diverse audiences. This process underscores the role of interdisciplinarity as a mechanism for bridging knowledge systems and enabling innovation. For research governance, this

underlines the importance of targeted support for early-stage interdisciplinary collaborations. Seed funding schemes and flexible peer review procedures can help to reduce entry barriers for emerging domains, thereby enabling them to establish credibility and accelerate their integration into the broader academic system.

The diversification of audiences over time further confirms that scholarly communication has moved from a narrow academic model toward a broader ecosystem that includes policymakers, professionals, students, and the public. This shift reflects an increasing recognition of the societal role of science. It also reveals how institutional incentives, such as funding programs requiring societal impact, create feedback loops that encourage researchers to target wider audiences. Such transformations highlight the importance of science communication policies that promote accessibility and inclusiveness. At the same time, balanced funding strategies are needed to ensure that emphasis on short-term societal impact does not crowd out long-term investments in fundamental research.

The temporal dynamics also point to the influence of global and national agendas in directing scholarly attention. For instance, sustainability and climate research gained prominence in parallel with policy initiatives such as the Paris Agreement and national carbon neutrality plans, while cultural heritage research experienced fluctuations in alignment with state-led projects. These patterns demonstrate that policy frameworks and funding schemes exert strong structuring effects on the academic landscape. Consequently, policymakers play a decisive role in either reinforcing existing centers of knowledge or fostering new areas of growth through deliberate interventions.

Finally, the distilled principles of cumulative advantage, preferential connectivity, and knowledge diffusion provide a theoretical framework for understanding the empirical patterns observed. Cumulative advantage explains the persistence of early leaders who consolidate their positions through accumulated resources and visibility, while preferential connectivity accounts for the tendency of new topics to associate with established hubs in order to gain credibility. Knowledge diffusion, in turn, highlights the uneven and context-specific spread of ideas, shaped by both internal methodological innovations and external policy shifts. Together, these mechanisms suggest that knowledge evolution is not random but path dependent, with systemic inequalities that may perpetuate existing hierarchies. Recognizing these dynamics carries direct implications for policy and evaluation: field-normalized metrics, equitable funding strategies, and support for underrepresented domains are crucial for maintaining a balanced and inclusive research ecosystem. By integrating such measures into science governance, policymakers can mitigate structural imbalances and promote knowledge development that is both academically robust and societally responsive.

CONCLUSION

This study proposes a semantic-based framework for analyzing the formation and migration of knowledge. By moving beyond citation-based approaches, we demonstrate that semantic embeddings capture the structural organization of heterogeneous knowledge. The concept of knowledge transfer flow further enables us to trace the evolution of research hotspots, while the integration of sentiment analysis reveals how perception and interpretation shape the reception of ideas.

The empirical validation through the KnowFlowViz system illustrates the applicability of this framework. The system supports interactive exploration of static knowledge structures as well as dynamic migration flows, offering a tool for trend forecasting, research evaluation, and knowledge governance.

Future work will refine the semantic representation model by incorporating more diverse data sources and advanced natural language processing. We will also extend the concept of knowledge transfer flow to include interdisciplinary collaborations and the influence of technological drivers. Furthermore, improvements in sentiment analysis will allow for the detection of more nuanced expressions of evaluation in scientific discourse. Finally, the development of a time slider for KnowFlowViz will enable continuous exploration of knowledge shifts across the full temporal range, overcoming the limitations of snapshot-based visualization.

RESOURCE AVAILABILITY

Materials availability

The web-based platform developed in this study has been made publicly accessible and can be accessed at https://star-kiwi-grossly.ngrok-free.app/. All necessary resources for reproducing the functionalities of the platform are included.

Data and code availability

All datasets and code utilized or generated in this study are available from the corresponding author upon reasonable request.

FUNDING AND ACKNOWLEDGMENTS

The research is partially supported by EU H2020 Research and Innovation Program under the Marie Sklodowska-Curie Grant Agreement (Project-DEEP, grant no. 101109045), the National Natural Science Foundation of China (nos. NSFC 62401334 and 62442106). Additionally, the research is funded by the German Federal Ministry of Education and Research (BMBF) (project AITT, AI-assisted Technology Transfer, no. 03LB3058B), the Program of Jiangsu Province under grant NTACT-2024-Z-001. The authors also wish to express their sincere gratitude to Mr. Liang Chen, Mr. Shengli Ren, and Mr. Jun Qian from China Science Publishing & Media Ltd. for providing the essential data and for sharing their expert insights into the field of scholarly publishing, which were invaluable to this research.

AUTHOR CONTRIBUTIONS

Formal analysis, J.P.; investigation, J.P.; methodology, J.P.; software, J.P. and J.L.; validation, J.P.; visualization, J.P.; writing – original draft, J.P.; writing – review & editing, J.O.; conceptualization, Z.N.; data curation, Y.W.; funding acquisition, X.T.; supervision, J.O. and J.L. All authors contributed to the manuscript and approved the final version.

DECLARATION OF INTERESTS

The authors declare no competing interests.

SUPPLEMENTAL INFORMATION

It can be found online at https://doi.org/10.1016/j.xinn.2025.101179.

REFERENCES

1. Walsh, I. and Rowe, F. (2023). Bibgt: combining bibliometrics and grounded theory to conduct a literature review. *Eur. J. Inf. Syst.* **32**:653–674. DOI:10.1080/0960085X.2022.2039563
2. Donthu, N., Kumar, S., Mukherjee, D. et al. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *J. Bus. Res.* **133**:285–296. DOI:10.1016/j.jbusres.2021.04.070
3. Ji, S., Pan, S., Cambria, E. et al. (2022). A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Trans. Neural Netw. Learn. Syst.* **33**:494–514. DOI:10.1109/TNNLS.2021.3070843
4. Abualqumboz, M., Chan, P.W., Bamford, D. et al. (2021). Temporal dimensions of knowledge exchanges in horizontal knowledge networks. *J. Knowl. Manag.* **25**:899–919. DOI:10.1108/JKM-05-2020-0346
5. Krishen, A.S., Dwivedi, Y.K., Bindu, N. et al. (2021). A broad overview of interactive digital marketing: A bibliometric network analysis. *J. Bus. Res.* **131**:183–195. DOI:10.1016/j.jbusres.2021.03.061
6. Connected Papers (2025). Connected papers: Explore connected work in your field. https://www.connectedpapers.com
7. Scite (2025). Scite: Smart citations for better research. https://scite.ai
8. Abu-Salih, B. (2021). Domain-specific knowledge graphs: A survey. *J. Netw. Comput. Appl.* **185**:103076. DOI:10.1016/j.jnca.2021.103076
9. Zhang, L., Li, N. and Li, Z. (2021). An overview on supervised semi-structured data classification. In IEEE 8th International Conference on Data Science and Advanced Analytics (IEEE), pp. 1–10. DOI:10.1109/DSAA53316.2021.9564205
10. Peng, W., Varanka, T., Mostafa, A. et al. (2022). Hyperbolic deep neural networks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**:10023–10044. DOI:10.1109/TPAMI.2021.3136921
11. Wang, W., Shen, X., Yi, B. et al. (2024). Knowledge-aware fine-grained attention networks with refined knowledge graph embedding for personalized recommendation. *Expert Syst. Appl.* **249**:123710. DOI:10.1016/j.eswa.2024.123710
12. Bali, A., Bhagwat, A., Bhise, A. et al. (2024). Semantic similarity detection and analysis for text documents. In 2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE) (IEEE), pp. 1–9. DOI:10.1109/icETITE58242.2024.10493834
13. Bacciu, D., Errica, F., Micheli, A. et al. (2020). A gentle introduction to deep learning for graphs. *Neural Netw.* **129**:203–221. DOI:10.1016/j.neunet.2020.06.006
14. Xiao, S., Chen, Y., Song, Y. et al. (2024). Ui semantic component group detection: Grouping ui elements with similar semantics in mobile graphical user interface. *Displays* **83**:102679. DOI:10.1016/j.displa.2024.102679
15. Ye, Y. and Ge, Y. (2019). A bibliometric analysis of inventory management research based on knowledge mapping. *Electron. Libr.* **37**:127–154. DOI:10.1108/EL-11-2017-0241

16. Zou, Z., Cheng, J., Huang, K. et al. (2019). Research on the developments of business intelligence and its enlightenment based on bibliometric statistics and knowledge map analysis. *J. Phys, Conf. Ser.* **1176**:042089. DOI:10.1088/1742-6596/1176/4/042089
17. Li, H., Xiang, Y., Yang, W. et al. (2024). Green roof development knowledge map: A review of visual analysis using citespace and vosviewer. *Heliyon* **10**:e24958. DOI:10.1016/j.heliyon.2024.e24958
18. Sun, L., Xu, X., Yang, Y. et al. (2020). Knowledge mapping of supply chain risk research based on citespace. *Comput. Intell.* **36**:1686–1703. DOI:10.1111/coin.12306
19. Chawla, R.N. and Goyal, P. (2022). Emerging trends in digital transformation: a bibliometric analysis. *Benchmarking: An Int. J.* **29**:1069–1112. DOI:10.1108/BIJ-01-2021-0009
20. Chen, Q. and Wang, Y. (2020). Research status and development trend of concentrating solar power. In 2020 9th Int. Conf. Renewable Energy Research and Application (ICRERA) (IEEE), pp. 390–393. DOI:10.1109/ICRERA49962.2020.9242893
21. Peng, J., Kimmig, A., Wang, D. et al. (2023). A systematic review of data-driven approaches to fault diagnosis and early warning. *J. Intell. Manuf.* **34**:3277–3304. DOI:10.1007/s10845-022-02020-0
22. Ruan, W., Hou, H. and Hu, Z. (2017). Detecting dynamics of hot topics with alluvial diagrams: A timeline visualization. *J. Data Inf. Sci.* **2**:37–48. DOI:10.1515/jdis-2017-0013
23. Chotisarn, N., Merino, L., Zheng, X. et al. (2020). A systematic literature review of modern software visualization. *J. Vis.* **23**:539–558. DOI:10.1007/s12650-020-00647-w
24. Korkut, E.H. and Surer, E. (2023). Visualization in virtual reality: a systematic review. *Virtual Real.* **27**:1447–1480. DOI:10.1007/s10055-023-00753-8
25. Sharma, R., Gulati, S., Kaur, A. et al. (2022). Research discovery and visualization using researchrabbit: A use case of ai in libraries. *Collnet J. Scientometrics Inf. Manag.* **16**:215–237. DOI:10.1080/09737766.2022.2106167
26. Pan, L., Xu, Z. and Skare, M. (2023). Sustainable business model innovation literature: a bibliometrics analysis. *Rev. Manag. Sci.* **17**:757–785. DOI:10.1007/s11846-022-00548-2
27. Meng, F., Lu, Z., Li, X. et al. (2024). Demand-side energy management reimagined: A comprehensive literature analysis leveraging large language models. *Energy* **291**:130303. DOI:10.1016/j.energy.2024.130303
28. Ji, H. and Gan, W. (2020). Data visualization for making sense of scientific literature. In 2020 Int. Conf. Intell. Transp., Big Data & Smart City (ICITBS), pp. 870–873. DOI:10.1109/ICITBS49701.2020.00191
29. Chen, X. and Liu, Y. (2020). Visualization analysis of high-speed railway research based on citespace. *Transp. Policy* **85**:1–17. DOI:10.1016/j.tranpol.2019.10.004
30. Aljohani, N.R., Fayoumi, A. and Hassan, S.U. (2021). A novel deep neural network-based approach to measure scholarly research dissemination using citations network. *Appl. Sci.* **11**:10970. DOI:10.3390/app112210970
31. Dong, S., Mei, F., Li, J.J. et al. (2023). Global cluster analysis and network visualization in prosthetic joint infection: a scientometric mapping. *Orthop. Surg.* **15**:1165–1178. DOI:10.1111/os.13681
32. Wu, T., Duan, Y., Zhang, T. et al. (2022). Research trends in the application of artificial intelligence in oncology: A bibliometric and network visualization study. *Front. Biosci.* **27**:254. DOI:10.31083/j.fbl2709254
33. Das, R., Diaz, J., Avissar, P. et al. (2020). 4051 assessing outcomes of miami ctsi's mentored career development kl2 program: Using bibliometric and network visualization approaches to complement traditional outcome metrics. *J. Clin. Transl. Sci.* **4**:70. DOI:10.1017/cts.2020.230
34. Chang, C.Y., Gau, M.L., Tang, K.Y. et al. (2021). Directions of the 100 most cited nursing student education research: A bibliometric and co-citation network analysis. *Nurse Educ. Today* **96**:104645. DOI:10.1016/j.nedt.2020.104645
35. Mahadevan, K. and Joshi, S. (2022). Omnichannel retailing: a bibliometric and network visualization analysis. *Benchmarking: An Int. J.* **29**:1113–1136. DOI:10.1108/BIJ-12-2020-0622
36. Wang, Z.J., Hohman, F. and Chau, D.H. (2023). Wizmap: Scalable interactive visualization for exploring large machine learning embeddings. Preprint at *arXiv*. DOI:10.48550/arXiv.2306.09328
37. Nandwani, P. and Verma, R. (2021). A review on sentiment analysis and emotion detection from text. *Soc. Netw. Anal. Min.* **11**:81. DOI:10.1007/s13278-021-00776-6
38. Siino, M. (2024). Badrock at semeval-2024 task 8: Distilbert to detect multigenerator, multidomain and multilingual black-box machine-generated text. In Proc. 18th Int. Workshop Semantic Evaluation (SemEval-2024), pp. 239–245. DOI:10.18653/v1/2024.semeval-1.37
39. Saha, U., Mahmud, M.S., Keya, M. et al. (2022). Exploring public attitude towards children by leveraging emoji to track out sentiment using distil-bert a fine-tuned model. In Int. Conf. Image Process. Capsule Networks (Springer), pp. 332–346. DOI:10.1007/978-3-031-12413-6_26
40. Salatino, A., Aggarwal, T., Mannocci, A. et al. (2025). A survey of knowledge organization systems of research fields: Resources and challenges. *Quant. Sci. Stud.* **6**:567–610. DOI:10.1162/qss_a_00363
41. Peng, J., Kimmig, A., Niu, Z. et al. (2021). A flexible potential-flow model based high resolution spatiotemporal energy demand forecasting framework. *Appl. Energy* **299**:117321. DOI:10.1016/j.apenergy.2021.117321
42. Van Kranenburg, H., Hagedoorn, J. and Pennings, J. (2023). Measurement of international and product diversification in the publishing industry. *J. Media Econ.* **17**:87–104. DOI:10.1207/s15327736me1702_2