Scientometrics
https://doi.org/10.1007/s11192-023-04776-5

Identifying interdisciplinary topics and their evolution based on BERTopic

Zhongyi Wang[1] · **Jing Chen**[1] · **Jiangping Chen**[2] · **Haihua Chen**[2]

Received: 4 February 2023 / Accepted: 12 June 2023

Abstract
Interdisciplinary topic reflects the knowledge exchange and integration between different disciplines. Analyzing its evolutionary path is beneficial for interdisciplinary research in identifying potential cooperative research direction and promoting the cross-integration of different disciplines. However, current studies on the evolution of interdisciplinary topics mainly focus on identifying interdisciplinary topics at the macro level. More analysis of the evolution process of interdisciplinary topics at the micro level is still needed. This paper proposes a framework for interdisciplinary topic identification and evolutionary analysis based on BERTopic to bridge the gap. The framework consists of four steps: (1) Extract the topics from the dataset using the BERTopic model. (2) Filter out the invalid global topics and stage topics based on lexical distribution and further filter out the invalid stage topics based on topic correlation. (3) Identify interdisciplinary topics based on disciplinary diversity and disciplinary cohesion. (4) Analyze the interdisciplinary topic evolution by inspecting the intensity and content in the evolution, and visualize the evolution using Sankey diagrams. Finally, We conduct an empirical study on a dataset collected from the Web of Science (WoS) in Library & Information Science (LIS) to evaluate the validity of the framework. From the dataset, we have identified two distinct types of interdisciplinary topics in LIS. Our findings suggest that the growth points of LIS mainly exist in the interdisciplinary research topics. Additionally, our analysis reveals that more and more interdisciplinary knowledge needs to be integrated to solve more complex problems. Mature interdisciplinary topics mainly formed from the internal core knowledge in LIS stimulated by external disciplinary knowledge, while promising interdisciplinary topics are still at the stage of internalizing and absorbing the knowledge of other disciplines. The dataset, the code for implementing the algorithms, and the complete experiment results will be released on GitHub at: https://github.com/haihua0913/IITE-BERT.

Keywords Academic literature · Interdisciplinary research · Interdisciplinary topics · Topic modeling · Topic evolution · BERTopic

This study is supported by National Social Science Foundation of China (Grant Number 22BTQ102).

✉ Haihua Chen
haihua.chen@unt.edu

[1] School of Information Management, Central China Normal University, Wuhan 430079, China

[2] Department of Information Science, University of North Texas, Denton, TX 76203, USA

Published online: 03 July 2023

Introduction

As globalization continues to bring new challenges for individuals, interdisciplinary research has become increasingly important for addressing complex problems that require expertise from multiple disciplines. Interdisciplinary research not only accelerates scientific discoveries but also promotes scientific developments and innovations (Wu & Zhang, 2019). Interdisciplinarity exists at different levels, from researchers and research institutions to articles and journals (Callon et al., 1983). To quantify interdisciplinarity, scholars have proposed various indicators and measurement methods: diversity (Leydesdorff et al., 2019), betweenness centrality (Leydesdorff & Hellsten, 2006) and information entropy (Loet & Ismael, 2011), or a combination of these indicators. For example, Dong et al. (2018) proposed a multi-dimensional interdisciplinary research (IDR) topic identification method that combines the co-word analysis, IDR feature word analysis, outlier analysis, and burst word monitoring to identify hot and potential interdisciplinary topics. All the above indicators for interdisciplinarity measurement are developed from Rao (Alvargonzález, 2011) and Stirling (Derrick et al., 2011). However, the above-discussed indicators are mostly used to measure the interdisciplinarity of the literature at a macro level. In the case of micro-IDR topics, these indicators do not fully reflect their inner characteristics. To bridge the gap, we focus on the interdisciplinarity of the research topic by measuring the disciplinary diversity and disciplinary cohesion of the topics at a micro level.

The study of topic evolution focuses on tracking the development and change of topic over time. These changes include how their survive, such as whether their importance decreases or increases; and the interactive process among topics, such as merging and splitting between topics (Chen et al., 2017). Topic evolution studies can help researchers, policy makers and funding agencies to understand the full picture of scientific disciplines with complex knowledge structures more effectively and efficiently, especially in interdisciplinary fields (Qian et al., 2020). Therefore, upon identification of interdisciplinary topics, it is necessary to conduct a thorough analysis of their evolution in the interdisciplinary process at a micro level. The existing interdisciplinary topic evolution researches mainly focus on exploring how the knowledge of an interdisciplinary field changes over time (Song et al., 2014) or the formation process of an interdisciplinary field from the perspective of knowledge diffusion (Xu et al., 2018).

The current studies mainly adopt an interdisciplinary perspective, utilizing topic evolution as a means to elucidate the interdisciplinary field's development or to gain a deeper understanding of the factors contributing to its formation. However, the existing researches neglect the degree of interdisciplinarity of topics during their evolution process analysis, they can not reveal the unique patterns of different types of topics, which is essential for understanding the transfer of knowledge between fields and the formation of interdisciplinary topics. To solve this problem, this paper first distinguishes the degree of interdisciplinarity of topics in the interdisciplinary research field, and then analyzes their unique evolution patterns of these interdisciplinary topics.

To sum up, the main contributions of this study are as follows:

(1) We propose a framework for identifying interdisciplinary topics and analyzing their evolution. The framework can be used to explore the degree of interdisciplinary knowledge fusion at a micro level, and present the internal features of interdisciplinary behavior.

(2) We conduct an empirical study on a dataset in library & Information Science (LIS) collected from Web of Science to evaluate the effectiveness of the proposed framework.
(3) Based on the WoS dataset, we discover the trend of each interdisciplinary topic intensity, identify their inheritance, splitting and merging paths, and analyze the knowledge transfer between disciplines under each interdisciplinary topic.

Related works

Interdisciplinary topic and other relevant concepts

Interdisciplinary research is the dynamic process of cross-integrating information, methods, techniques, tools, and theories from more than one disciplines or knowledge communities, which aims to solve or deepen the understanding of problems that transcend the scope of a single discipline (Derrick et al., 2011). "Interdisciplinarity" was first proposed by R.S. Woodworth from Columbia University at Social Science Research Council in 1926, which indicates research activities that exceed the scope of one discipline (Zhang & Wu, 2017). Some scholars think that interdisciplinarity refers to the interdisciplinary characteristics of interdisciplinary research, including the breadth and intensity of interdisciplinary knowledge and the characteristics of interdisciplinary distribution and diffusion of knowledge across disciplines (Li, 2014). Therefore, it helps us to determine the interdisciplinary topics from a disciplinary knowledge fusion perspective.

Based on the definition of interdisciplinarity, this paper defines interdisciplinary topic as a joint research topic at the interdisciplinary intersection of two or more disciplines. The higher the degree of interaction between disciplines, the higher the interdisciplinarity of the topic, which can be determined by measuring both the disciplinary diversity and disciplinary cohesion of the topic (Rafols & Meyer, 2010). Where disciplinary diversity indicates the breadth of the knowledge and disciplinary cohesion reflects the novelty of knowledge integration (Rafols & Meyer, 2010). In our study, these two metrics are measured by the number of disciplines published on the topic and the extent to which publications from different disciplines are cited together respectively. In summary, interdisciplinary topics are important hubs for the intersection and integration of knowledge from different disciplines and often become the frontier of disciplines or new disciplinary growth points.

Interdisciplinary topic identification

Interdisciplinary topic identification has attracted increasing interest in various fields, which aims to discover the specific intersections representing the convergence of different research fields. There are two different kinds of interdisciplinary topic identification processes: first, identify interdisciplinary literature, then identify topics from them, and these are considered interdisciplinary topics. The second is to identify the topics directly from the literature, measure the interdisciplinarity of the topics, and select those above a certain threshold as interdisciplinary topics. Based on the above ideas, different interdisciplinary topics identification methods have been proposed, including co-word analysis, citation analysis, and text mining.

Co-word analysis-based methods for interdisciplinary topic identification are considered a valuable and objective methods (Trotta & Garengo, 2017), which take the frequency of keyword co-occurrence across different disciplines as a proxy for the level of

interdisciplinary research. This metric is commonly referred to as the degree of interdisciplinarity. Specifically, the set of high-frequency interdisciplinary words represents a hot interdisciplinary topic, while the group of low-frequency interdisciplinary words may indicate potential interdisciplinary topics. For example, Ling et al. (2015) constructed several weak co-occurrence networks to analyze research topics and their interdisciplinarity. In addition, some scholars have established the connection between high-frequency keywords and burst words in co-word analysis and combined the two into the same model to identify hot spots and new topics simultaneously (Li, 2017).

Citation-based interdisciplinary topic identification methods can be divided into co-citation analysis, coupled analysis, and direct citation analysis. Adams and Light (2014) constructed a literature coupling network for papers in AIDS research, combined with community detection algorithms to identify the communities among them, and identified topics that spanned multiple disciplinary communities as interdisciplinary topics. Small (2010) processed the co-citation data among journal papers by the clustering algorithm based on the co-citation relationship, thus obtaining interdisciplinarity and similarity among disciplines.

However, both of the above two kinds of methods have their disadvantages. For example, co-word analysis-based methods ignore the grammatical and semantic information of the text. The citation analysis-based methods also have their limitations, such as citations take a long time to appear and various citation motivations. To bridge the gap, text mining has been widely utilized to identify interdisciplinary topics. Text mining is a knowledge discovery technique that enables researchers to examine unstructured text and extract previously unknown, understandable, potential, and observed patterns or knowledge from a collection of textual data (Zhang et al., 2015). Thus, it can go deep inside the document and effectively reveal the hidden topic of the articles. Several topic models identify topics based on machine learning from extensive document collections. Latent Dirichlet Allocation(LDA) and probabilistic latent semantic analysis(PLSA) are the most widely used methods. However, these traditional topic models rely on the assumption of "Bag-of-Words," which ignores inner semantics relationship between words through lexical bag representation (Zhou & Wakabayashim, 2022). As these representations do not take into account the context of the words in the sentence, word bag may not accurately represent the document (Grootendorst, 2022). BERTopic, as a neural topic model, can represent words as multi-dimensional vectors and capture the context information (Grootendorst, 2022), yielding more accurate and richer features. Grootendorst (2022) demonstrated the method's effectiveness on topic identification. Based on the above analysis, we aim to adopt the BERTopic model to extract interdisciplinary topic more efficiently and effectively.

Interdisciplinary topic evolution

The analysis of interdisciplinary topic evolution aims to reveal the development or change process of interdisciplinary topics, identifying hotspots, frontiers, or future development trends in the research field.

To reveal the underlying dynamics of interdisciplinary research, the existing studies have explored the evolution of topics in interdisciplinary fields from various

perspectives. For example, Song et al. (2014) utilized Markov Random Field(MRF)-based topic clustering and meta-term mapping to analyze the evolution of topics related to bioinformatics over time. This study revealed some distinct topic transition patterns between different time periods. Some studies focus on the correlation between topics during the evolution of topics, such as splitting, merging, and others. Chen et al. (2017) discussed knowledge transfer indicated by splitting and merging activities in Information Retrieval and identified three types of knowledge migration (non-migration, double migration, and multiple migrations). Balili et al. (2020) proposed the TermBall framework to track and predict fine-grained topic evolution from the perspective of evolutionary types of topic evolution, including emergence, growth, shrinkage, survival, merging, splitting, and dissolution, and validated the framework by applying it to 19 million articles in PubMed Central.

Based on a summary of previous work, we find the present analysis of the topic evolution of interdisciplinary fields primarily centers around tracking research trends within a specific interdisciplinary domain or predicting its future development trends. In addition, traditional research regards all research topics in interdisciplinary fields as interdisciplinary topics without distinguishing the degree of interdisciplinarity. Due to the varying degrees of interdisciplinarity, different types of interdisciplinary topics may display unique patterns during the evolution process. The interactive relationship between different types of interdisciplinary topics and the merging and splitting of interdisciplinary topics during the evolution process are yet to be explored. To address this gap, this paper reveals the evolution of different types of interdisciplinary topics and summarizes the law of the interdisciplinary topics evolution process, providing a new perspective for studying interdisciplinary topic evolution.

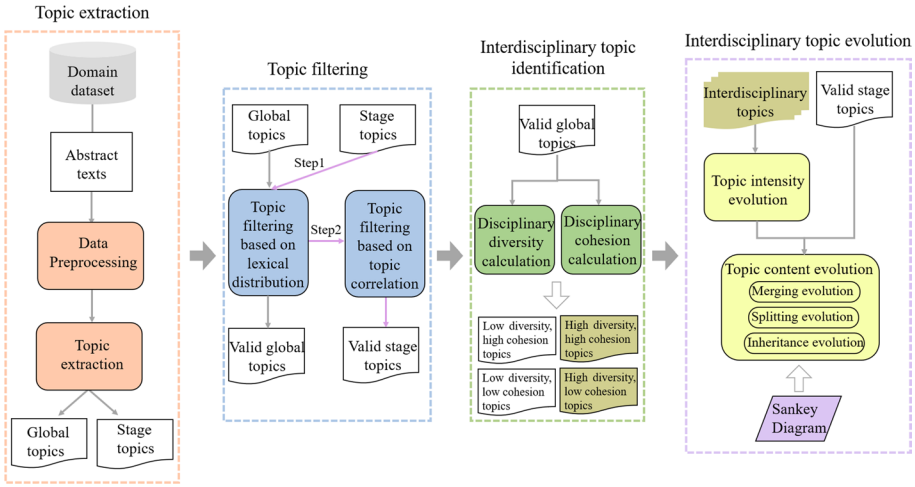

Fig. 1 Framework for interdisciplinary topic identification and evolution analysis

Methodology

In this paper, we proposed a framework for interdisciplinary topic identification and evolution analysis. The framework is shown in Fig. 1. It is a four-step procedure that includes: (1) topic extraction, (2) topic filtering, (3) interdisciplinary topic identification, and (4) interdisciplinary topic evolution.

Topic extraction

In this paper, we apply BERTopic, a state-of-the-art topic modeling technique to extract topics in LIS. The BERTopic model uses BERT embeddings and cluster-based TF-IDF to generate dense clustering, while utilizing the unified manifold approximation and projection (UMAP) technique to reduce the embedding dimension of documents prior to clustering them (Grootendorst, 2022). This approach does not necessitate a predetermined number of topics, which confers an advantage upon BERTopic over LDA. The workflow of the BERTopic model for topic extraction is shown in Fig. 2:

1. Embed documents: papers' abstracts are fed into a pre-trained model, which calculates the word vector of each abstract, and subsequently converts each document into its corresponding embedded representation.
2. Cluster documents into semantically similar clusters: Reduce the embedding dimension by using the UMAP algorithm and cluster the document embeddings with HDBSCAN for document clustering.
3. Create topic representations from clusters: c-TF-IDF (TF-IDF variant) is used to evaluate the importance of each word for each HDBSCAN cluster and we select the representa-

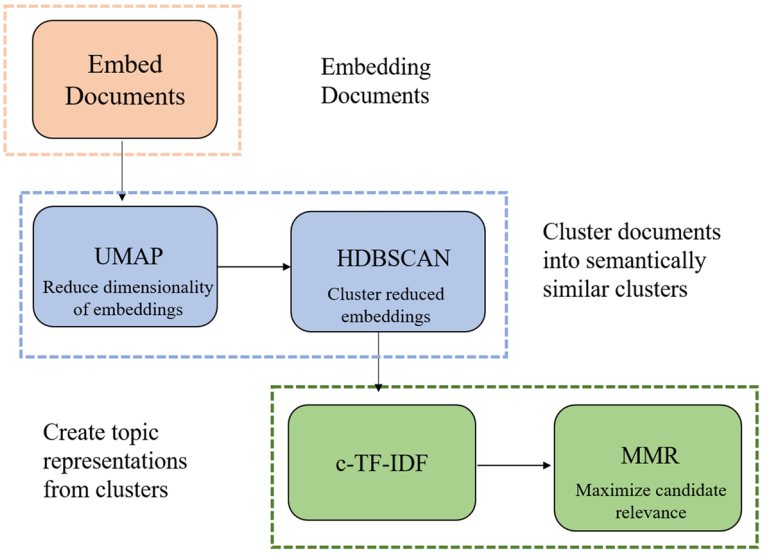

Fig. 2 Topic extraction based on BERTopic model

tive word for each topic for topic representation. In this way, we generate the document-topic distribution matrix and the topic-word distribution matrix.

Topic filtering

Topic filtering based on lexical distribution

To identify the interdisciplinary topics, it is essential to ensure the validity of topics by filtering out meaningless ones. After topic extraction, the "topic-word" distribution matrix is obtained and each topic can be represented as a probability distribution of a set of words. If the probability distribution of words in a topic is even, it means that the meaning of this topic is unclear, we consider this topic is invalid and filter out it. On the other hand, the topic with explicit semantic expression is characterized by an uneven probability distribution of words in the topic. In this case, the high probability words of a topic can explicitly reflect its meaning. Therefore, we consider this topic is valid topic.

Based on the above analysis, as information entropy can be used to quantify the degree of an uneven distribution of the words' probabilities in a topic, we use it to filter the valid topics.The formula for information entropy is given in Eq. (1) (MacKay, 2003).

$$Entropy(T) = -K \sum_{i=1}^{m} P(Wi|T) \ln(P(Wi|T)) \tag{1}$$

Note that K is a constant number and $P(W\ i|T)$ indicates the probability of the ith word of topic T, m is the number of words in a topic. The smaller the information entropy of a topic, the more explicit the meaning of the topic, and it will be more likely regarding as an valid topic.

Topic filtering based on topic correlation

The underlying assumption of filtering stage topic is that the set of stage topics extracted from the stage corpus must be related with those extracted global topics from the global corpus. To ensure the validity of stage topics, the similarity between stage and global topics must exceed a specific threshold. Different similarity measures can be utilized to compute the similarity score, including *KL* divergence, cosine distance, and Manhattan distance. As each topic's word distribution approximates a vector, cosine similarity performs well in distinguishing correlations between such vectors by emphasizing the contribution of the top words with a high probability and suppressing the noise produced by words with low probability (Chen et al., 2017). Therefore, we adopted a cosine similarity algorithm to compute the similarity scores between stage and global topics. the cosine similarity can be computed by Eq. (2):

$$CS(T_1, T_2) = \frac{T_1 \cdot T_2}{||T_1|| \cdot ||T_2||} = \frac{\sum_{i=1}^{n} \left(X_i|T_1) \cdot (Y_i|T_2)\right)}{\sqrt{\sum_{i=1}^{n} (X_i|T_1)^2} \cdot \sqrt{\sum_{i=1}^{n} (Y_i|T_2)^2}} \tag{2}$$

Equation (2) represents a similarity metric which indicates how much two topics *T1* and *T2* are similar. Where X are the keywords involved in topic *T1*, and Y is the keywords involved in topic *T2*.

Utilizing a high threshold in topic selection may lead to the exclusion of significant topics, whereas a lower threshold may not effectively exclude some invalid topics. Thus, selecting an appropriate threshold for this purpose is crucial. In this paper, we employ Eq. (3) for determining the threshold value (Jiang et al., 2022).

$$threshold = \frac{Z_t}{\sum_{i=1}^{m} Z_i} \tag{3}$$

where Z_t represents the number of global topics and Z_i represents the number of stage topics.

Interdisciplinary topic identification

This study combines disciplinary diversity and disciplinary cohesion for identifying interdisciplinary research topics. Disciplinary diversity is utilized to assess the disciplinary extent of the knowledge under a topic, and disciplinary cohesion is employed to measure the interdisciplinary correlation degree of disciplinary knowledge under a topic. The former dimension mainly assesses the degree of interdisciplinary integration based on the diversity of disciplines, and the latter measures the extent of interdisciplinary integration based on the disciplinary co-occurrence network analysis.

Indexes of disciplinary diversity

To measure the characteristics of interdisciplinary topics, Xu et al. (2016) proposed the topic terms interdisciplinarity (TI) index. This index can calculate disciplinary diversity of topics at a micro level, so we adopt this index to measure the disciplinary diversity of topics. The formula for calculating the TI value is shown in Eq. (4).

$$TI = d \times \log tf \tag{4}$$

where d denotes the number of disciplines where the topic is distributed; tf represents the frequency of the topic. TI is an indicator to measure the disciplinary diversity of the topic. The higher the TI value, the higher the disciplinary diversity of the topic.

Indexes of disciplinary cohesion

In complex network systems, network density serves as a metric for evaluating the closeness of the connections between nodes, which can indicate the level of cohesion within the network (Rafols & Meyer, 2010). Specifically, a higher network density of the disciplinary co-occurrence network is indicative of a stronger disciplinary cohesion. Hence, this paper employed the network density index of the disciplinary co-occurrence network to compute the disciplinary cohesion of a topic, as expressed in Eq. (5).

$$Network\ Density = \frac{2L}{N(N-1)} \tag{5}$$

here N represents the number of nodes in the network, that is, the number of disciplines contained in the topic, and L represents the number of connected edges in the network, that is, the number of disciplinary pairs with co-occurrence relationships.

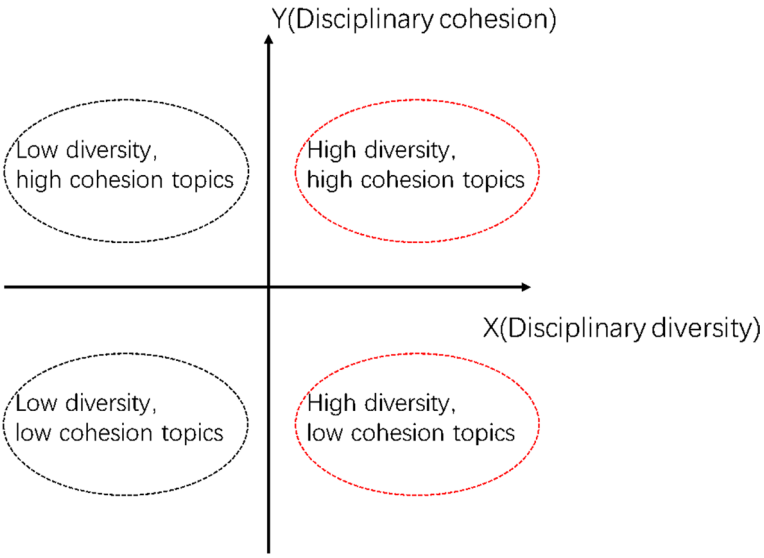

Fig. 3 Multidimensional scaling analysis of interdisciplinary topics

The values of the above two metrics can be used to divide the topics into four quadrants, as shown in Fig. 3.

The topics falling within the four quadrants can be explained as follows:

1. *Low diversity–low cohesion*, indicating that the topic integrates knowledge belonging to very few disciplines and the disciplinary knowledge under this topic is dispersed from each other, signifying a typical promising single-disciplinary research topic;
2. *Low diversity–high cohesion*, indicating that the topic is an integration of knowledge belonging to very limited disciplines and the disciplinary knowledge is closely related to each other, signifying a typical mature single-disciplinary research topic;
3. *High diversity–high cohesion*, indicating that the topic integrates knowledge derived from different disciplines, but the distance between these diverse disciplines is relatively close. Such topics fall within the scope of interdisciplinary research topics. Typically, these are mature interdisciplinary research topics;
4. *High diversity–low cohesion*, indicating that the study integrates knowledge from multiple different disciplines, wherein the knowledge itself is far apart. Such research represents the most significant and promising interdisciplinary research topic.

This paper has opted to examine topics falling within high diversity-high cohesion and high diversity-low cohesion quadrants as interdisciplinary research topics. The evolution process of these topics' characteristics is further explored in Sect. Interdisciplinary topic evolution analysis.

Interdisciplinary topic evolution

Interdisciplinary topic evolution comprises the evolution of topic intensity and topic content. Topic intensity evolution refers to pattern of topic prevalence over time, while topic content evolution relates to the trend of topic content over time.

Interdisciplinary topic intensity evolution

Currently, there are two primary methods for measuring topic intensity. The first method involves mapping each article to a topic and interpreting the topic intensity as the number of articles corresponding to the topic. The second method involves determining the posterior probability of topic words (Hall et al., 2008). Therefore, at the macro level, the intensity of a topic can be measured by means of indicators such as the number of articles published on the topic. From the micro perspective, topic intensity can be defined by the probability of its feature words. This paper proposes a methodology that combines the probability of a topic's feature words with the number of publications under the topic to derive the topic intensity calculation index. Let $W_k = \{W_1, W_2, ..., W_m\}$ denote the set of feature words of interdisciplinary topic t, where W_k represents the mth feature word of interdisciplinary topic t. The intensity of interdisciplinary topic t at time y can be calculated as Eq. (6):

$$\theta_y^t = \sum_{k=1}^{m} Num_y(t) \times \theta_t^{W_k} \tag{6}$$

θ_y^t is the intensity value of the interdisciplinary topic t at time y, $Num_y(t)$ represents the number of articles containing interdisciplinary topic t at time y, and $\theta_t^{W_k}$ represents the contribution of W_k to interdisciplinary topic t, that is, the probability of the feature words corresponding to the interdisciplinary topic t.

Interdisciplinary topic content evolution

The steps to identify the content evolution paths of the interdisciplinary topics are as follows:

1. *Identify the related stage topics of each interdisciplinary global topic* As not all stage topics are related to their interdisciplinary global topic, this paper calculates the similarity between each stage topic and its target interdisciplinary global topic to identify the related stage topics to the interdisciplinary topic. A stage topic whose similarity to an interdisciplinary global topic exceeds a predefined threshold is identified as its related stage topic. The similarity between them indicates the extent to which interdisciplinary global topics have developed at this stage.
2. *Identify the content evolution path of each interdisciplinary global topic* The cosine similarity between the related stage topics across adjacent stages are calculated. The degree of similarity between two stage topics is indicative of their correlation, with a higher cosine

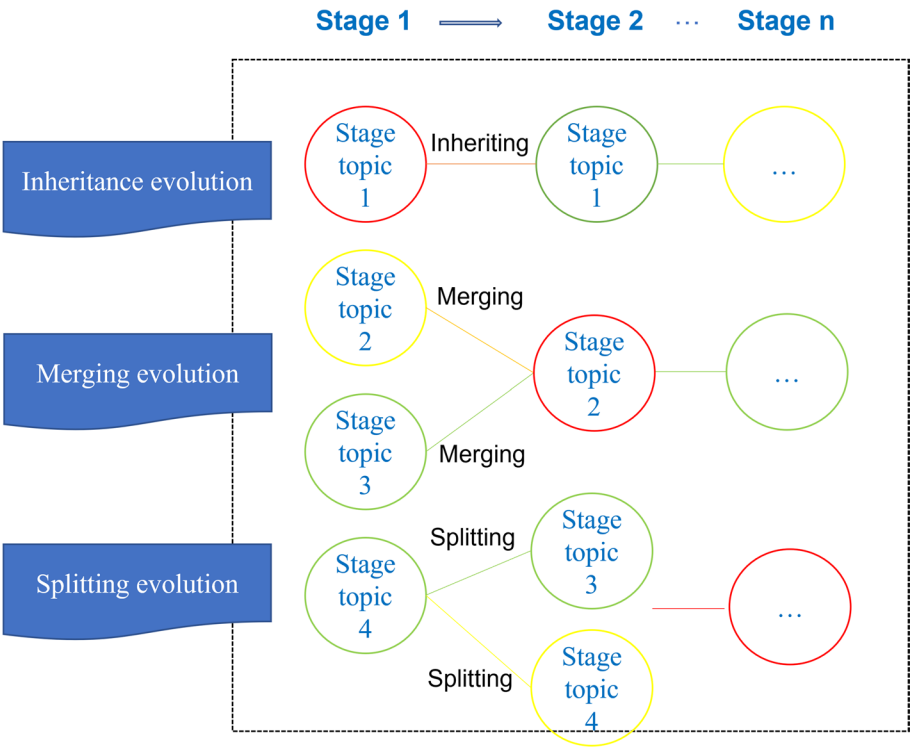

Fig. 4 Evolution path of topic content

similarity value indicating a stronger correlation. If the similarity between two stage topics exceeds a predefined threshold, they are considered to have an evolutionary path.

There are three different kinds of content evolution paths of an interdisciplinary global topic, including the inheritance, merging and splitting evolution paths, as shown in Fig. 4.

(1) *Inheritance evolution* Two topics in adjacent stages have a high semantic correlation, which means the latter topic inherits the information of the previous topic, forming an inherited evolutionary relationship.
(2) *Merging evolution* Suppose there is a high similarity between two or more topics at the previous stage and one topic at the next stage, and the topic at the later stage is new. In this case, these topics at the previous stage are merged into a new topic, forming a merged evolutionary relationship.
(3) *Splitting evolution* Suppose the topic at the previous stage has a high similarity with more than two topics at the next stage, and the latter topics are new. In this case, one topic at the previous stage is split into several new topics, forming a split evolutionary relationship.

Experiments and results

Data collection and exploration

Considering that Information Science & Library Science (LIS) is a typical interdiscipline, we collected the research papers in the LIS field from Web of Science (WoS) as the experimental data in this study. We first retrieved the articles from 2005 to 2022 based on the search strategy WC= "Information Science & Library Science" on November 11, 2022, and obtained 70,384 papers. Then, we removed duplicate articles and filtered out articles without titles, publication years, or abstracts. Finally, we obtained 57,994 articles as our experimental data.

Figure 5 presents the distribution of disciplines over time, as examined by counting the number of articles and disciplinary categories in different years. From Fig. 5 we can see that the number of articles exhibited a steady upward trend between 2005 and 2022. In the meanwhile, the number of disciplinary categories increased slightly and remained relatively stable over the same period. Specifically, the number of disciplinary categories showed an upward trend from 2005 to 2014, reaching a peak in 2014, when LIS research intersected with 25 disciplines. Since then, the number of disciplines has remained stable, indicating that research in LIS has continued to attract the attention of various disciplines and maintained interdisciplinary research collaborations.

To explore the horizontal distribution of disciplines, we analyzed the total number of articles in each discipline. All disciplines and their corresponding number of articles are reported in Table 1. From Table 1, we observe that the top three disciplines are "Computer Science, Information Systems" (19,092, 32.9%), "Computer Science, Interdisciplinary Applications" (9751, 16.8%), and "Management" (7969,13.7%). Table 1 indicates that research in LIS has mainly attracted the attention of scholars in Computer Science, Management, and Communication and is also interacting with many other fields.

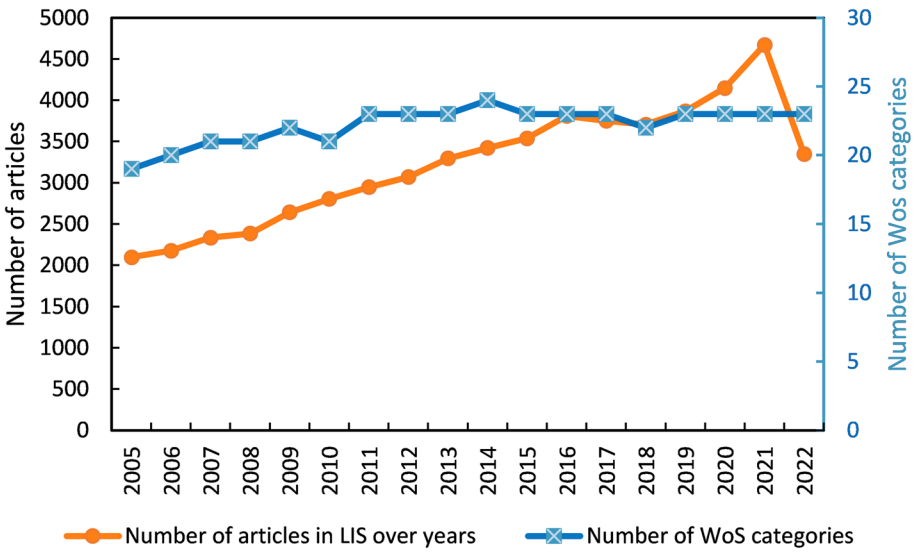

Fig. 5 Number of articles and WoS categories

Table 1 Number of research articles of each WoS categories

WoS categories	Total number
Information Science & Library Science	57,994
Computer Science, Information Systems	19,092
Computer Science, Interdisciplinary Applications	9751
Management	7969
Communication	5396
Social Sciences, Interdisciplinary	3733
Geography	3312
Health Care Sciences & Services	2745
Medical Informatics	2745
Public, Environmental & Occupational Health	2547
Social Sciences, Biomedical	2547
Geography, Physical	1656
Telecommunications	1220
Philosophy	441
Ethics	428
History	398
Development Studies	364
Law	332
Education & Educational Research	308
Humanities, Multidisciplinary	297
History Of Social Sciences	199
Multidisciplinary Sciences	189
Computer Science, Theory & Methods	129
History & Philosophy Of Science	15
Medical Ethics	1

Topic extraction

In this paper, topics are categorized into global topics and stage topics. Therefore, the topic extraction process is also divided into two steps: global topic extraction and stage topic extraction.

Global topic extraction

Topics derived by the BERTopic model from the global corpus are referred to as global topics. During the global topic extraction phase, the method proposed in Sect. "Topic extraction" was utilized to extract global topics. Specifically, BERTopic's preprocessed_text() is first used to preprocess the text of global corpus, and then the default embedding model, all-MiniLM-L6-v2 was employed for text representation. Considering the sample size, we set the min_cluster_size parameter to 50 during the HDBSCAN clustering process. When instantiating the BERTopic model, we set calculate_probabilities to True, diversity to 0.5, and min_topic_size to 50. Additionally, we used an n-gram range of (2,3) and selected the top feature words for output, while the

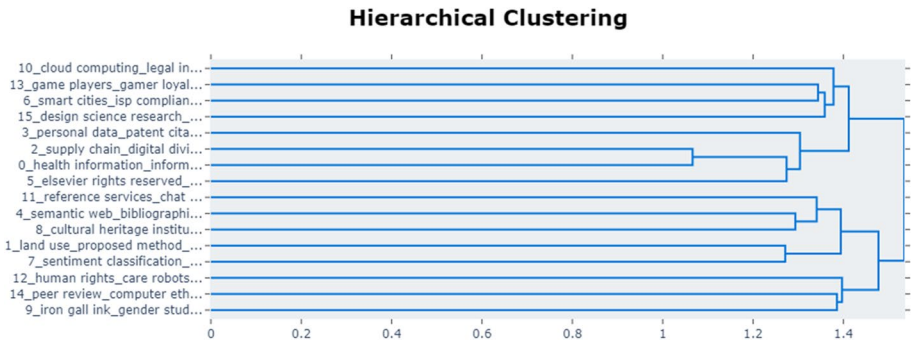

Fig. 6 Result for hierarchical clustering (part)

Table 2 Global topics and their feature words (part)

Topic id	Feature words
0	Health information; information literacy; health literacy; knowledge sharing; open access; elsevier rights reserved; research limitations implications; purpose paper; public libraries; semi structured interviews
1	Land use; proposed method; data reuse; road network; study area; cellular automata; urban growth; data set; geographic information systems; trajectory data
2	Supply chain; digital divide; software development; rights reserved; economic growth; elsevier bv rights; erp implementation; communication technology ict; dynamic capabilities; purpose paper
3	Personal data; patent citations; information disclosure; privacy calculus; data protection; united states; emerging technologies; rights reserved; trademark office; privacy issues
4	Semantic web; bibliographic records; data model; purpose paper; controlled vocabularies; metadata elements; research limitation simplications; knowledge organization systems; linked open data; functional requirements
5	Elsevier rights reserved; universal service; internet access; rural areas; net neutrality; network operators; broadband services; mobile network; united states; regulatory framework

remaining parameters were set to their default values. As a result of conducting BERTopic, a total of 86 global topics were extracted.

To ensure the coherence of each topic, we applied BERTopic's visualize_hierarchy() method to perform hierarchical clustering of the 86 topics, as illustrated in Fig. 6. Accordingly, in this study, we merged topics that were clustered together for the first time, and as a result, we obtained a total of 40 global topics. The resulting topics, along with their representative "feature words" are listed in Table 2 (part). Figure 7 shows a bar chart of the first six topics, depicting some of the feature words that contribute the most to each topic according to c-TF-IDF scores.

Fig. 7 Top five c-TF-IDF scores in six topics

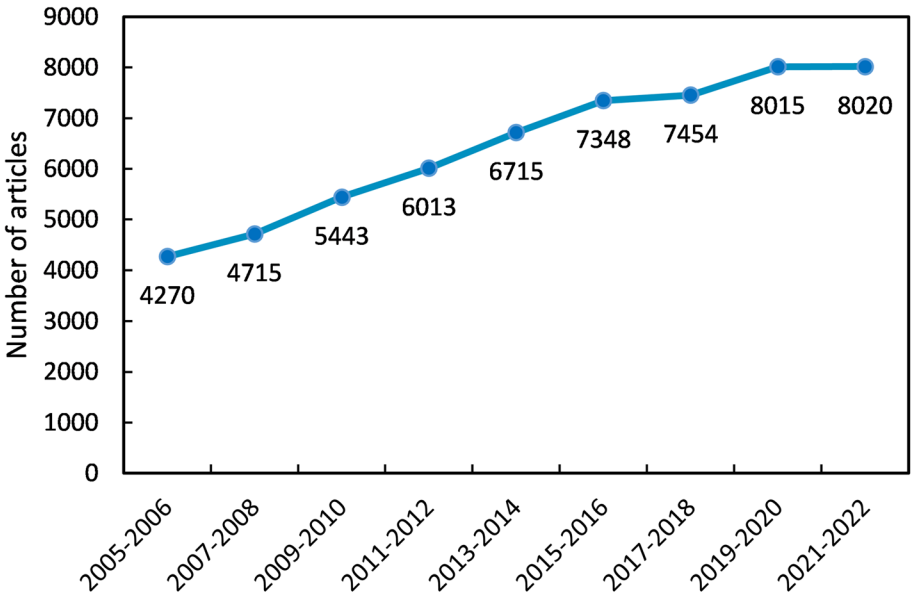

Fig. 8 Number of articles in different time slices

Stage topic extraction

The global corpus was further segmented into nine sub-stages with a two-year interval: 2005–2006, 2007–2008, 2009–2010, 2011–2012, 2013–2014, 2015–2016, 2017–2018, 2019–2020, and 2021–2022. The division of stages was based on the availability of

Table 3 Number of valid stage topics

Stage	Yearly slices	Number of stage topic	Number of valid stage topic	Percentage of invalid topic (%)
1	2005–2006	47	25	46.81
2	2007–2008	64	32	50.00
3	2009–2010	41	24	41.46
4	2011–2012	49	32	34.69
5	2013–2014	61	37	39.34
6	2015–2016	111	70	36.94
7	2017–2018	121	73	39.67
8	2019–2020	64	36	43.75
9	2021–2022	75	42	44.00

sufficient textual data for each stage. Accordingly, two years have been selected for each stage to provide a more detailed analysis of the evolution process. The distribution of the number of articles in different time spans after the division is depicted in Fig. 8.

As indicated in Fig. 8, a total of nine stages were identified. In this paper, the topics generated from the stage corpus are referred to as stage topics. Similarly, the extraction of stage topics was performed using the BERTopic model. Given the limited size of the stage corpus, the standard parameters were employed in the HDBSCAN clustering process. Furthermore, the parameters were set identically to the aforementioned process during the initialization of the BERTopic model. In addition, The identification results of the stage topics were also optimized through hierarchical clustering method.The number of stage topics associated with each stage is presented in Table 3.

Topic filtering

Global topic filtering

In this study,we employed the methodology proposed in Sect. "Topic filtering based on lexical distribution" to eliminate meaningless global topics. For the 40 global topics identified between 2005 and 2022, we computed the information entropy of each topic individually, and the average entropy value of each topic was 0.2442, which was utilized to filter out meaningless topics with an information entropy greater than 0.2442. At last we obtained 20 meaningful global topics.

Stage topic filtering

First, to eliminate the meaningless stage topics at each stage, We adopted the same method proposed in Sect. "Topic filtering based on lexical distribution". Second, the methodology proposed in Sect. "Topic filtering based on topic correlation" was employed to establish the correlation between stage topics and global topic, which can help us further filter stage topics based on topic correlation. Based on the Eq. (3), a threshold value of 0.0631 was obtained. The find_topics() function of the BERTopic model was used to compute the

cosine similarity between the stage topics and each global topic. A stage topic was considered valid if its similarity with the global topic is greater than the threshold. The filtering results are shown in Table 3.

Interdisciplinary topic identification

Interdisciplinary feature analysis based on diversity dimension

To reveal the interdisciplinary diversity characteristics of research topics in the field of LIS, the TI of each research topic was computed using Eq. (4). The results are reported in Table 4. The observations regarding research topics in the LIS field are as follows:

(1) Some research topics have a high value of TI despite having a small number of papers. Such topics are likely to intersect with multiple disciplines, such as topic13 (game loyalty) and topic16 (data analytics).
(2) Some research topics have a large number of articles but a low TI value. These topics may represent the core research topics of LIS or the common theories or methods of the discipline, such as topic8 (digital archiving) and topic9 (paper conservation).
(3) Some research topics have both a high value of TI and a large number of articles, representing hot research topics interested to all disciplines, such as topic0(health information) and topic1(data reuse).

Figure 9 illustrates the TI values of global topics at each stage. As shown in Fig. 9, the span of disciplinary diversity of most topics in LIS presents an upward trend. Notably, topic6 (smart cities) demonstrates a particularly significant increase in disciplinary diversity. Furthermore, topic8 (digital archiving), topic9 (paper conservation), and topic12 (artificial moral) display a steady increase in disciplinary diversity. The three topics with the highest degree of TI value: topic0 (health information), topic1 (data reuse), and topic2 (supply chain) also show a steady increase in disciplinary diversity. The increasing degree of disciplinary diversity indicates that these topics have received wide attention from scholars in various fields. This trend is suggestive of these topics being potential interdisciplinary hot and frontier topics. Although topic13 (game loyalty) presents a decreasing trend in

Table 4 TI of Global topics

Topic number	Record number	TI	Topic number	Record number	TI
0	20,777	107.94	36	82	26.793
1	1798	61.841	22	141	25.791
2	1780	48.756	17	141	24.591
6	429	47.382	8	410	23.515
3	719	39.994	18	171	22.33
13	219	37.447	26	123	18.809
7	428	34.209	12	240	16.662
16	178	33.7563	9	303	14.889
5	469	29.383	27	119	12.453
4	699	28.449	20	161	8.827

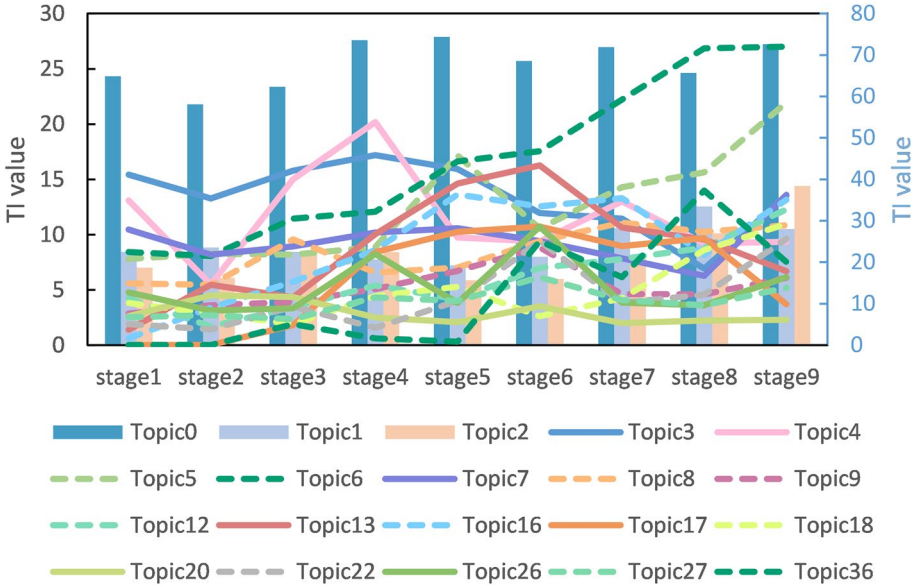

Fig. 9 The interdisciplinary trend of global topics at each stage

interdisciplinary diversity, it still maintains a high degree of disciplinary diversity. Additionally, there are topics, such as topic20 (spectrum sharing), which exhibit stable disciplinary diversity, suggesting that research in this area has relatively matured.

Interdisciplinary feature analysis based on cohesion dimension

We computed the disciplinary cohesion of topics utilizing the method proposed in Sect. Indexes of disciplinary cohesion. Specifically,we imported the disciplinary co-occurrence matrix of each topic into Gephi software to derive the network density. Our findings indicate that topic0 displayed a low level of disciplinary cohesion, with a value of 0.463, while simultaneously exhibiting the highest TI value. In addition, topic1 possesses the second highest TI value, indicating a strong degree of interdisciplinarity. However, its disciplinary cohesion does not surpass 0.5. This observation suggests that topic0 and topic1 can be classified as interdisciplinary topics characterized by high diversity and low cohesion. The multi-dimensional analysis of the remaining topics is shown in Fig. 10.

Figure 10 yielded several noteworthy findings. First, it reveals three distinct types of topics in LIS. Specifically, topic4 and topic7 were identified as high diversity and high cohesion topics, which have already evolved into relatively mature interdisciplinary research areas. In contrast, topic2, topic3, topic5, topic6, topic13, topic16, topic22 and topic36 were found to be high diversity and low cohesion topics, suggesting that they have the potential to become key areas for interdisciplinary knowledge fusion. Additionally, the majority of topics displayed low diversity but high cohesion, indicating that most researches in LIS tends to integrate knowledge within the same discipline. Secondly, there were no low-diversity and low-cohesive topics in this field. Therefore, topic1, topic2, topic3, topic4,

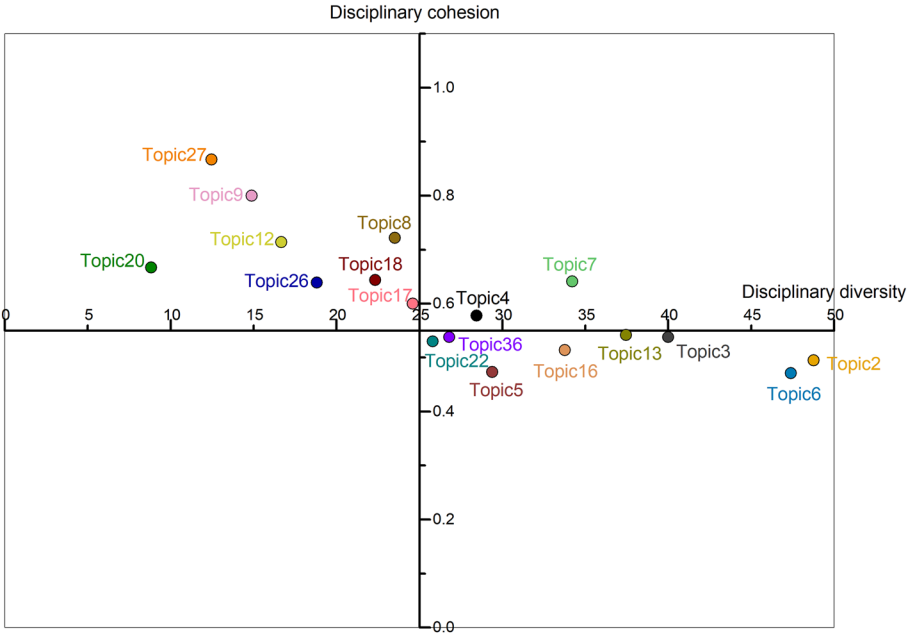

Fig. 10 Multidimensional scaling analysis of interdisciplinary topics(topic0 and topic1 not included)

topic5, topic6, topic7, topic13, topic16, topic22 and topic36 were identified as the interdisciplinary topics in LIS field.

Interdisciplinary topic evolution analysis

Topic intensity evolution

Tracking the changes in intensity of interdisciplinary topics over time can help to describe the popularity trend of them. In this paper, the intensity of each interdisciplinary topic at each stage is calculated based on Eq. (6). The intensity values for each interdisciplinary topic are presented in Fig. 11.

The results, depicted in Fig. 11, show that topic3 (personal data) and topic4 (semantic web) exhibit a declining trend, while topic2 (supply chain), topic5 (elsevier rights reserved), topic6 (smart cities), topic16 (virtual worlds), topic22 (subject headings), and topic36 (science technology) are ascending topics. Specifically, topic2 exhibits the highest intensity and the most prominent upward trend, while topic36 has gradually received the attention of scholars from stage 3. Additionally, topic1 (health information and knowledge sharing) shows a steady increase in intensity, while the intensity of topic7 (sentiment classification) maintains a steady level, but showed a slow downward trend from the fifth stage.

Topic content evolution

To reflect the evolution of topic content, we first employed the method proposed in Sect. "Interdisciplinary topic content evolution" to capture the relevant stage topics for

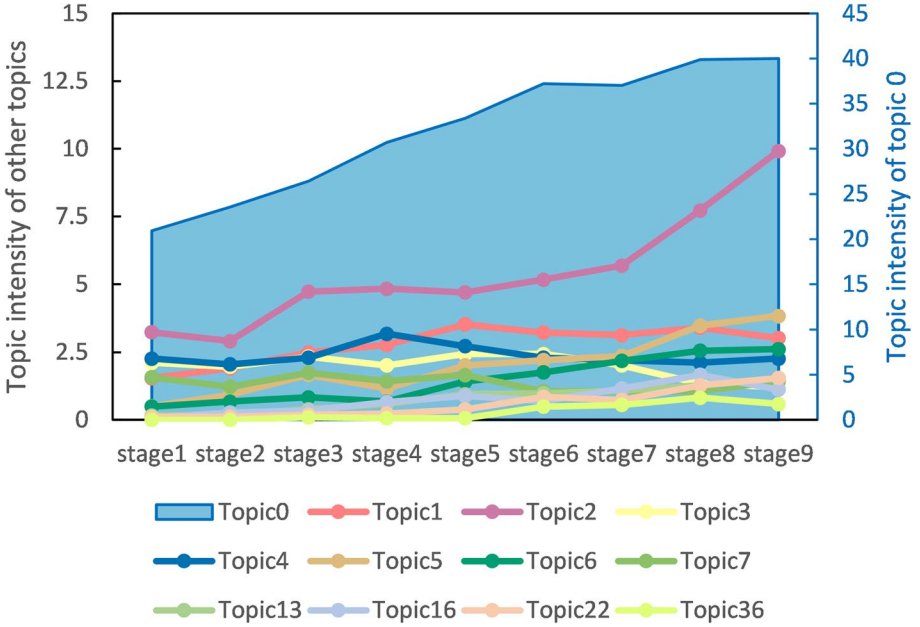

Fig. 11 Intensity evolution of interdisciplinary topics

interdisciplinary topics. We believe that the degree of similarity between interdisciplinary topics and their related stage topics can serve as an indicator of their respective developmental states, including adjusted medium maturity, medium maturity, and high maturity, so We divided 0.5 by 3 and rounded to 0.15 to determine the relevance interval (Chen et al., 2017). Since the similarity between most interdisciplinary topics and stage topics is less than 0.75, we have set 0.75 as the right endpoint of topic maturity state. Thus, the similarity threshold of interdisciplinary topics and their related stage topics is 0.3.

We then investigates the content evolution of interdisciplinary topics by examining the cosine similarity between their related stage topics in adjacent stages, where only those with a similarity score exceeding 0.3 are considered. The evolution paths of topic content are subsequently constructed, and the Sankey diagram is employed to present a visual

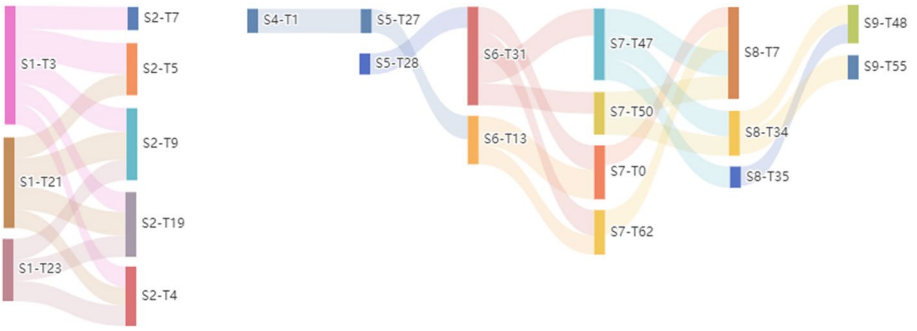

Fig. 12 Content evolution path diagram of topic2

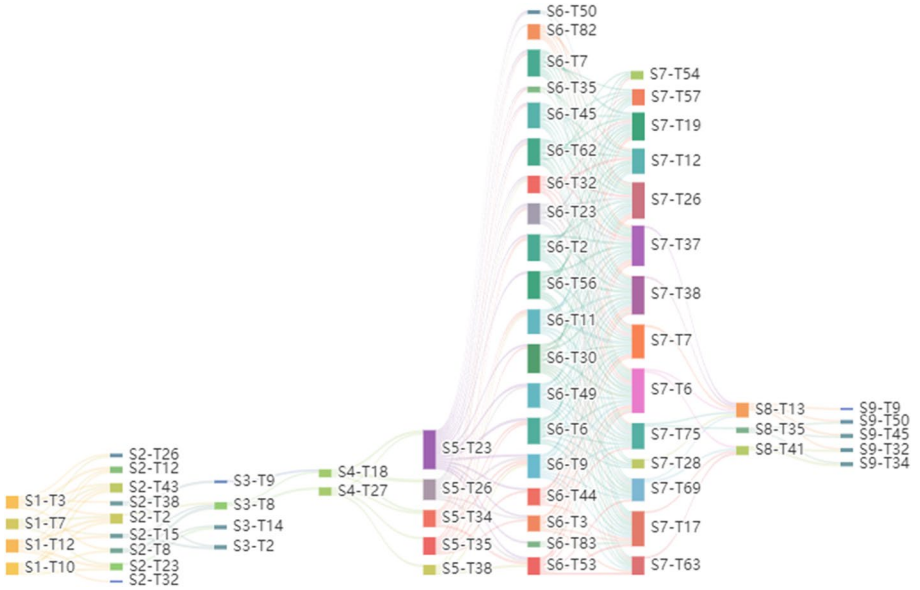

Fig. 13 Content evolution path diagram of topic4

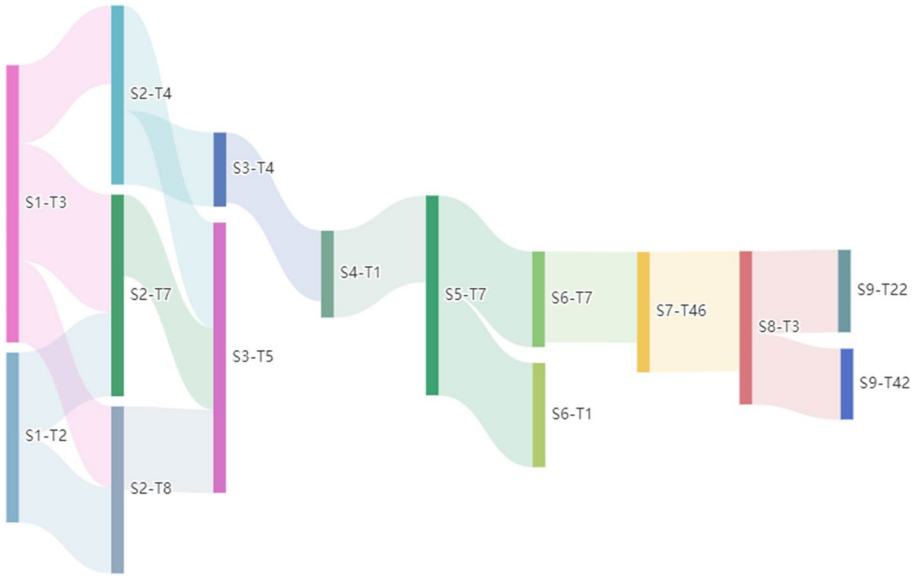

Fig. 14 Content evolution path diagram of topic5

representation of this process. The horizontal connecting lines in the diagrams describe the type and process of topic evolution at each stage, and the thickness of the connecting line indicates the magnitude of cosine similarity. The vertical element block shows the

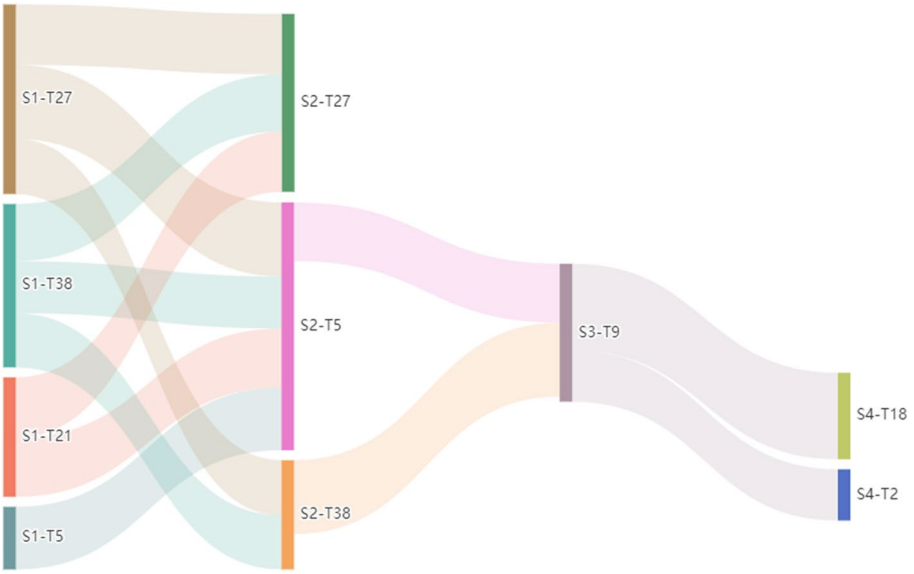

Fig. 15 Content evolution path diagram of topic7

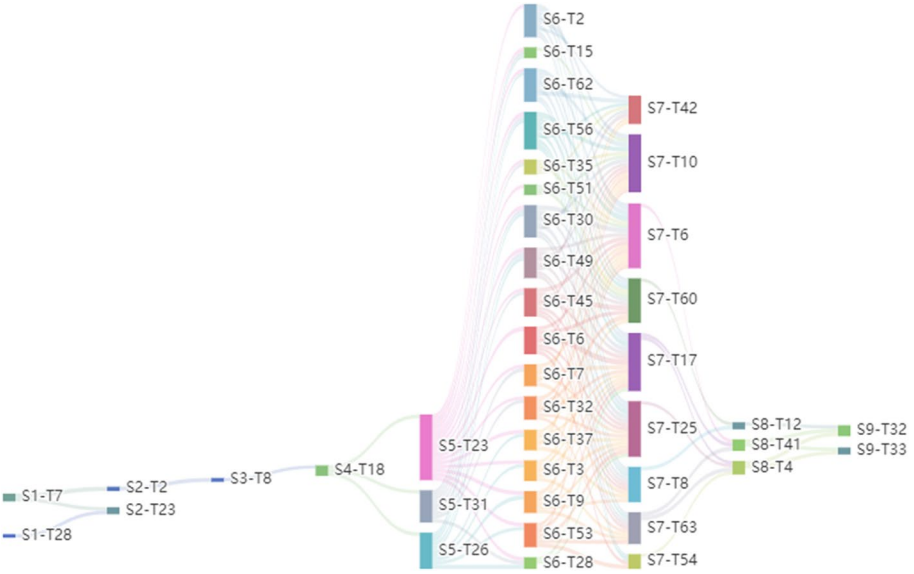

Fig. 16 Content evolution path diagram of topic22

distribution and importance of topics at each stage. We present the evolutionary path of each interdisciplinary topic separately, with S1 representing Stage 1 (i.e., 2005–2006) and so on.

To investigate the laws of the topic content evolution process of different interdisciplinary topics, we selected topic4 and topic7, which have high disciplinary cohesion and diversity, as well as topic2, which has high diversity, and topic5 and topic22, which have low diversity in the fourth quadrant. The results of this analysis, as presented in Fig. 12, 13, 14, 15 and 16, reveals that, interdisciplinary topics with high diversity and low cohesion are relatively in a more unstable development state and have diversified evolution paths. For instance, topic5 and topic22 have underwent three evolution paths of topic inheritance, merging and splitting during the development process, while the mature interdisciplinary topics (topic4 and topic7) evolve mainly through merging and splitting paths. In addition, different interdisciplinary topics are active at different stages. Specifically, topic5 underwent merges and splitting of topics in the first three stages, and inherited and split in subsequent stages. In contrast, topic7 merged and split in the first four stages, with no significant content evolution occurring in subsequent stages.

Notably, although topic7 did not shift in topic content since the fourth stage, there are new topics in the later stage, which means it maintained a relatively stable state of development. Specifically, the research direction at the fifth stage was S5-T13 (similarity of 0.508), which also remained stable afterward, such as S6-T62 (similarity of 0.452), S7-T116 (similarity of 0.449), S8-T6 (similarity of 0.435) and S9-T33 (similarity of 0.485). Topic4 showed a decreasing trend in topic intensity after the fourth stage, but from the fifth stage onwards, the splitting and merging among this topic became more active and even peaked at stage6.

Findings and discussions

1. The growth points of LIS mainly exist in the interdisciplinary research topics, which indicates that LIS is developing towards the broad and pluralistic direction of interdisciplinary knowledge integration.

From Fig. 10, we find that the promising research topics characterized by low cohesion in LIS mainly fall in the fourth quadrant, remarkably, our study did not identify any research topics in the third quadrant. which means that growth points in LIS mainly comes from the knowledge fusion between LIS and other disciplines, such as topic2 (supply chain) and topic3 (personal data), and there is no new growth point has been formed within the LIS.

2. The interdisciplinary intensity of research topics in LIS shows an increasing trend, indicating that with the development of society, the research questions faced by LIS are becoming more complex. The more complex problems people are faced with, the more interdisciplinary knowledge needs to be integrated, resulting in greater interdisciplinary knowledge.

Through an examination of the interdisciplinary topic intensity evolution, as illustrated in Fig. 11, we find that the intensities of most interdisciplinary topics show a steady upward trend, which means that the disciplinary diversity of interdisciplinary topics is increasing, for example, the topic2, topic5, topic6, topic16, topic22, and topic36. What this means is that the inherent complexity of problems encountered by LIS necessitates an integration of knowledge from various disciplines to effectively solve the multifaceted problems. It is hence evident that interdisciplinary fusion plays a vital role in addressing intricate issues.

3. Mature interdisciplinary topics mainly formed from the internal core knowledge in LIS stimulated by external disciplinary knowledge, which are characterized by long research time and stable topic intensity.

Mature Interdisciplinary topics with high cohesion and high diversity are often produced by the collaboration of core professional knowledge in LIS field such as "bibliographic records","knowledge organization systems", "paper propose" with interdisciplinary input knowledge such as "community detection" and "semantic web", "sentiment classification" and "social media". This type of interdisciplinary topics exist at all stages in our study and have received continuous attention. This kind of interdisciplinary topics have accumulated certain research achievements in the process of growth, and these researches have been relatively mature, which are the main driving forces for the development of the LIS field.

4. The promising interdisciplinary topics are at the stage of internalization and absorption of knowledge from other disciplines, and have not been deeply integrated with them.

The semantic relationship between the disciplinary knowledge under the promising interdisciplinary topic has not been found, new interdisciplinary knowledge has not been formed, for example, topic22 ("open government data", "data initiatives"), topic5 ("internet access ", "net neutrality"), topic2 ("supply chain ", "digital divide"). Knowledge of different disciplines jointly promote the innovation and development of the promising interdisciplinary topics and stimulate the research in LIS field.

5. The promising interdisciplinary topics are in an unstable development process with more diversified evolution paths. These topics are in the growing stage, different disciplinary knowledge under this kind of topics collides with each other.

The promising interdisciplinary topics with high diversity and low cohesion are in an unstable development process, and their evolution paths are more diversified. For example, topic5 and topic22 have experienced three evolution paths of topic inheritance, merging and splitting in the development process. However, the highly mature interdisciplinary topics (topic4 and topic7) are mainly merged and split in the development process. Most of the promising interdisciplinary topics with high diversity and low cohesion have formed a continuous evolutionary relationship since the fourth stage, indicating that the interdisciplinary knowledge collaboration of such topics is gradually close, and they are expected to develop into mature interdisciplinary knowledge growth points.

Conclusion

Interdisciplinary research is increasingly regarded as the key to tackle contemporary complex societal challenges and to stimulate scientific innovation. To explore the internal mechanism of interdisciplinary research, more and more researchers proposed different methods to identify the interdisciplinary topics and their evolution process from various perspectives. However, the existing researches neglect the degree of interdisciplinarity of topics during their evolution process analysis, they can not reveal the unique patterns of different types of topics, which is essential for understanding the transfer of knowledge

between fields and the formation of interdisciplinary topics. To solve this problem, this paper first applied BERTopic to identify interdisciplinary topics from large-scale academic literature and then conducted fine-grained evolution analysis on these extracted topics. Specifically, we first identified the interdisciplinary topics in LIS from two perspectives: disciplinary diversity and disciplinary cohesion. Then, we investigated the trend of topic intensity and content evolution and analyzed the evolutionary path of interdisciplinary topics in this field.

However, this study has its limitations. First, this approach considers only journal literature, regardless of other literature, such as conference papers, books, dissertations, and others. In the future, we will do further studies to consider all kinds of literature to improve the performance of the proposed framework for interdisciplinary topic identification and evolution. Second, this paper limits to comprehensively understand the evolution process of interdisciplinary topics, however, unfortunately, this study lacks to predict the future trend of each interdisciplinary research topic. Therefore, in the future, we further attempt to deploy link prediction on each interdisciplinary topic to predict their future trends.

Acknowledgements The paper is a substantially extended version of the article "Interdisciplinary Topics Extraction and Evolution Analysis" presented in the 3rd Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents 2022 (EEKE 2022) at JCDL 2022. The authors are grateful to all the anonymous reviewers for their precious comments and suggestions.

Author contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Zhongyi Wang, Jing Chen, Jiangping Chen, and Haihua Chen. The first draft of the manuscript was written by Zhongyi Wang and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

References

Adams, J., & Light, R. (2014). Mapping interdisciplinary fields: Efficiencies, gaps and redundancies in HIV/AIDS research. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0115092

Alvargonzález, D. (2011). Multidisciplinarity, interdisciplinarity, transdisciplinarity, and the sciences. *International Studies in the Philosophy of Science, 25*(4), 387–403.

Balili, C., Lee, U., Segev, A., Kim, J., & Ko, M. (2020). Termball: Tracking and predicting evolution types of research topics by using knowledge structures in scholarly big data. *IEEE Access, 8*, 108514–108529.

Callon, M., Courtial, J. P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information, 22*(2), 191–235.

Chen, B., Tsutsui, S., Ding, Y., & Ma, F. (2017). Understanding the topic evolution in a scientific domain: An exploratory study for the field of information retrieval. *Journal of Informetrics, 11*(4), 1175–1189.

Derrick, E. G., Falk-Krzesinski, H. J., Roberts, M. R., & Olson, S. (2011). Facilitating interdisciplinary research and education: A practical guide. In *Report from the "Science on FIRE: Facilitating interdisciplinary research and education" workshop of the American Association for the advancement of science*.

Dong, K., Xu, H., Luo, R., Wei, L., & Fang, S. (2018). An integrated method for interdisciplinary topic identification and prediction: A case study on information science and library science. *Scientometrics, 115*, 849–868.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. arXiv preprint arXiv:2203.05794.

Hall, D., Jurafsky, D., & Manning, C. D. (2008). Studying the history of ideas using topic models. In *Proceedings of the 2008 conference on empirical methods in natural language processing* (pp. 363–371).

Jiang, L., Zhang, T., & Huang, T. (2022). Empirical research of hot topic recognition and its evolution path method for scientific and technological literature. *Journal of Advanced Computational Intelligence and Intelligent Informatics, 26*(3), 299–308.

Leydesdorff, L., & Hellsten, I. (2006). Measuring the meaning of words in contexts: An automated analysis of controversies about'monarch butterflies'',frankenfoods',and'stem cells'. *Scientometrics, 67*(2), 231–258.
Leydesdorff, L., & Ismael, R. (2011). Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations. *Journal of Informetrics, 5*(1), 87–100.
Leydesdorff, L., Wagner, C. S., & Bornmann, L. (2019). Interdisciplinarity as diversity in citation patterns among journals: Rao-stirling diversity, relative variety, and the gini coefficient. *Journal of Informetrics, 13*(1), 255–269.
Li, M. (2017). An exploration to visualise the emerging trends of technology foresight based on an improved technique of co-word analysis and relevant literature data of wos. *Technology Analysis & Strategic Management, 29*(6), 655–671.
Li, J. (2014). The concept and measurement of interdisciplinarity. *Documentation, Information & Knowledge, 3*, 87–93.
Ling, W., Haiyun, X., Ting, G., & Shu, F. (2015). Study on the interisciplinary topics of information science based on weak co-occurrence and burst detecting. *Library and Information Service, 59*(21), 105.
MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge University Press.
Qian, Y., Liu, Y., & Sheng, Q. Z. (2020). Understanding hierarchical structural evolution in a scientific discipline: A case study of artificial intelligence. *Journal of Informetrics, 14*(3), 101047.
Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience. *Scientometrics, 82*(2), 263–287.
Small, H. (2010). Maps of science as interdisciplinary discourse: Co-citation contexts and the role of analogy. *Scientometrics, 83*(3), 835–849.
Song, M., Heo, G. E., & Kim, S. Y. (2014). Analyzing topic evolution in bioinformatics: Investigation of dynamics of the field with conference data in dblp. *Scientometrics, 101*, 397–428.
Trotta, D., & Garengo, P. (2017). A co-word analysis on human resource management literature: The role of technological innovation from 2007–2017. In *20th Excellence in services international conference conference proceedings* (Vol. 9, pp. 797–810).
Wu, X., & Zhang, C. (2019). Finding high-impact interdisciplinary users based on friend discipline distribution in academic social networking sites. *Scientometrics, 119*(2), 1017–1035.
Xu, H., Guo, T., Yue, Z., Ru, L., & Fang, S. (2016). Interdisciplinary topics of information science: A study based on the terms interdisciplinarity index series. *Scientometrics, 106*, 583–601.
Xu, J., Bu, Y., Ding, Y., Yang, S., Zhang, H., Yu, C., & Sun, L. (2018). Understanding the formation of interdisciplinary research from the perspective of keyword evolution: A case study on joint attention. *Scientometrics, 117*(2), 973–995.
Zhang, C., & Wu, X. (2017). Review on interdisciplinary research. *Journal of the China Society for Scientific and Technical Information, 36*(05), 523–535.
Zhang, Y., Chen, M., & Liu, L. (2015). A review on text mining. In *2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS)* (pp. 681–685). IEEE.
Zhou, Z., & Wakabayashim, K. (2022). Topic modeling using jointly fine-tuned BERT for phrases and sentences. In *The 14th forum on data engineering and information management*