Scientometrics (2022) 127:5353–5383
https://doi.org/10.1007/s11192-022-04273-1

Identification of topic evolution: network analytics with piecewise linear representation and word embedding

Lu Huang[1] · Xiang Chen[1] · Yi Zhang[2] · Changtian Wang[1] · Xiaoli Cao[1] · Jiarun Liu[1]

Received: 13 March 2021 / Accepted: 12 January 2022 / Published online: 5 February 2022
© Akadémiai Kiadó, Budapest, Hungary 2022

Abstract

Understanding the evolutionary relationships among scientific topics and learning the evolutionary process of innovations is a crucial issue for strategic decision makers in governments, firms and funding agencies when they carry out forward-looking research activities. However, traditional co-word network analysis on topic identification cannot effectively excavate semantic relationship from the context, and fixed time window method cannot scientifically reflect the evolution process of topics. This study proposes a framework of identifying topic evolutionary pathways based on network analytics: Firstly, keyword networks are constructed, in which a piecewise linear representation method is used for dividing time periods and a Word2Vec mode is used for capturing semantics from the context of titles and abstracts; Secondly, a community detection algorithm is used to identify topics in networks; Finally, evolutionary relationships between topics are represented by measuring the topic similarity between adjacent time periods, and then topic evolutionary pathways are identified and visualized. An empirical study on information science demonstrates the reliability of the methodology, with subsequent empirical validations.

Keywords Bibliometrics · Topic analysis · Network analytics · Topic evolution

Introduction

The evolution of scientific topics could profile how topics change over time, which usually include whether they are developed maturely, import knowledge from other topics, merge or split into others, as well as which topics are gaining importance or dying out (Chen, Tsutsui, et al., 2017; Zhou & Jiang, 2020). Such evolutionary characteristics are meaningful for researchers to better understand scientific domains, to track development process of topics, to recognize relationships between topics, and then grasp scientific frontiers and research trends (Katsurai & Ono, 2019; Wang, Liu, et al., 2014).

✉ Xiang Chen
bjchenxiang@hotmail.com

[1] School of Management and Economics, Beijing Institute of Technology, Beijing, China

[2] Australian Artificial Intelligence Institute, Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, Australia

Recently, bibliometrics and data mining researchers have paid great attention to topic evolution analysis (Ding & Stirling, 2016; Zhu et al., 2015), and many models have been developed to identify topic emergence and content transitions among topics (Chen, Zhang, et al., 2017; Jeong & Min, 2014; Yau et al., 2014). Network-based analytics, as an important method introduced to recent bibliometric studies, has been proven to be effective in predicting emerging technologies (Érdi et al., 2013), revealing hidden technological opportunities (Park & Yoon, 2018), and analyzing the changing trend of topics in an interdisciplinary field (Liu, 2005; You et al., 2017). Among them, the application of co-word network analysis has been providing novel perspectives to discover knowledge structure (Zhang et al., 2021). However, traditional co-word network may tend to ignore the semantic meanings behind keywords, resulting in the sparse distribution of linkage between keywords in networks and inaccurate descriptions of relationships among keywords (Balili et al., 2020; Hu, Qi, et al., 2018), and then reducing the accuracy of topic identification.

Meanwhile, "time" information is a vital factor when exploring the characteristics of topic evolution (Wu et al., 2021), since the results of topic evolution are inevitably sensitive to the selected time slice, which must be quantified accordingly, dependent on documents (Zhou et al., 2017). It is important to capture the hidden trend turning points of the document set, and grasp topic dynamics over time (Chen, Zhang, et al., 2017; Wang, Cheng, et al., 2014). Fixed time window methods, which divide the total research time into fixed-length segments (Wang, Cheng, et al., 2014), have been widely used in the studies of topic evolution (Wang, Cheng, et al., 2014; Zhou et al., 2017). However, fixed methods could not adapt to quasi-periodic data (Sheng et al., 2015), leading to the lack of diversity and integrity of evolutionary pathways (Miao et al., 2020), for example, two or more topics may fall into one time window, or the data in the time window is too short to describe a topic.

To address these concerns, we propose a novel methodology based on network analytics to reveal the evolutionary pathways of scientific topics. The proposed method has three specific functions: (1) keyword networks combined with a Word2Vec model to capture semantic and contextual relationships between keywords; (2) the application of community detection in the constructed keyword networks to identify topics more accurately reflecting knowledge structure; (3) the use of piecewise linear representation method to detect the trend turning points over time and flexibly divide the time periods based on keywords time series.

The rest of this paper is organized as follows: Literature Review section reviews the literature of the related studies; Methodology section presents the theoretical framework and the details of our method; Empirical Study section presents a case study in the field of information science to verify the effectiveness of the method; finally, we get the conclusion from the case study and discuss the practical significance of the study, as well as the limitations and future work.

Literature review

This literature review specifically contains four aspects: topic evolution, time period division, word embedding and community detection.

Topic evolution

Understanding topic evolution in a scientific domain can facilitate the promotion of knowledge transfer, and help keep track of innovations and knowledge flows (Chen, Zhang, et al., 2017). Kuhn (1962) studied the structure of the scientific revolution and considered that the development of science included several phases, such as normal, crisis, revolution, and the new normal, and many researchers followed his work and detected the shifts of science through temporal patterns in scientific networks (Börner et al., 2003). Existing studies on the evolution of scientific topics mainly include topic identification (Zhang et al., 2017), content transitions (Chen, Zhang, et al., 2017), keywords overview (Cheng et al., 2020), prominent cluster visualizations (Xu et al., 2019), topic trends detection (Zhou et al., 2006), etc.

In the research of scientometrics, most document sets are organized in the order of time flow, and thus scientific topic content, scientific topic number and other characteristics of corpus are time evolving (Chen, Tsutsui, et al., 2017). In order to capture such evolutionary pathways, topic models that integrate with timestamps have been developed. Existing methods of topic evolution models could be roughly divided into three categories (Zhou et al., 2017): (1) discrete time topic evolution model (The et al., 2006); (2) continuous time topic evolution model (Wang & McCallum, 2006; Wang et al., 2008), where time variable is not divided but instead is continued as a continuous variable during topic evolution, and (3) online topic evolution model (Iwata et al., 2010), in which text corpus time is described by online streaming.

Discrete time topic evolution model has been proved to be effective in capturing the changes of scientific development and scientific topic trends (Blei & Lafferty, 2006; Mei & Zhai, 2005; Zhou et al., 2017). The steps of this model generally include (Zhou et al., 2017): the division of scientific text corpus according to timestamps, topic extraction, and topic evolutionary process identification.

Time period division

Topic evolution is one of the quasi-periodic activities (Silvestrini et al., 2017), which is affected by time (Wang et al., 2016). The results of topic evolution are inevitably sensitive to the selected time slice, which must be quantified accordingly, dependent on documents (Wu et al., 2021; Zhou et al., 2017). Therefore, an important task of understanding topic evolution is seeking out a proper strategy to identify turning points of a research topic and adaptively divide time series into reasonable segments (Chen, Zhang, et al., 2017).

Many studies split the literature of the target field into multiple time span with a fixed window (Wang, Liu, et al., 2014; Zhou et al., 2017), however, fixed window methods could not adapt to quasi-periodic data (Sheng et al., 2015), leading to the lack of diversity and integrity of evolutionary pathways (Miao et al., 2020). For instance, two or more topics may fall into one time window or that the data in the time window is too short to describe a topic. If we use the fixed time window to get the instances, the truncated position of the data will appear in a random position of a period and the cycle number in one time window will be unpredictable (Sheng et al., 2015). Recently, scholars have proposed some flexible partition methods (Wang et al., 2016; Carmona-Poyato et al., 2021). In particular, piecewise linear representation (PLR) has been proved to be one of the adoptive linear fitting methods for trend analysis. PLR is a time series analysis method proposed by Keogh et al.

(2001), which represents the time series approximation of the original data in several compressed sections (Keogh et al., 2004). PLR shows two benefits for topic evolution analysis in the partitioning of the dataset: (1) PLR could divide time series into non-overlapping segments, and approximate each segment with straight lines, reducing the dimension of data compression as well as filter out many local noises (Yan 2015; Huang & Zhou, 2016), allowing us to observe trend turning points more intuitively (Luo et al. 2013); (2) PLR could keep the primary trend of the specific dataset based on time series and captures the dynamics of topic intensity and content (Huang & Zhou, 2016). For example, Chen et al. (2015), Chen, Zhang, et al. (2017)) applied PLR to quantitatively identify trend turning points in patent publication activities. Because of its advantages, it has been utilized in many fields, including stock prediction (Chang et al., 2009; Luo & Chen, 2013), audio signal (Kimura et al., 2008), and socioeconomic analysis (Cruz et al. 2020).

Word embedding

Word embedding is an application of deep learning in natural language processing (NLP). It creates a way to detect underlying semantics in large-scale text by mapping words from vocabularies to vectors of real numbers (Mikolov et al., 2013). Words with similar contexts tend to have similar implications, which is the basic assumption of word embedding (Firth, 1957). The word embedding method can make text quantifiable, which can reduce the human costs of data cleaning methods (Zhang et al., 2018). Compared with other word embedding techniques, neural network algorithms are more effective in discovering word patterns with similar implications (Levy & Goldberg, 2014). Although Bert shows a stronger advantage in performance (Jeong et al., 2020), which is one of the highest performing pre-trained models for NLP learning representation, and requires high computing capacity and large-scale data, we could also highlight three benefits of word embedding compared with Bert in bibliometrics: (1) higher training speed and efficiency; (2) lower demand for computing capacity; (3) stronger versatility for bibliometric tasks. For example, after the emergence of Bert in 2018, there are still multiple researchers applying word embedding in many bibliometric tasks, including recommending patents (Chen et al., 2020), identifying and tracking scientific and technological knowledge memes (Sun & Ding, 2018), and measuring academic entitles' impact (Zhang & Wu, 2021).

Word2Vec method (Mikolov et al., 2013) is a symbol of neural network algorithms, which can be used for detecting a potential decomposition of the particular point-wise mutual information matrix (Levy & Goldberg, 2014). It is an unsupervised method to extract semantic relationships among words based on their co-occurrence in text documents in a corpus (Onan, 2019), and maps words into vectors in a high dimensional space, exhibiting relatively high efficiency and accuracy in measuring the semantic similarity among words (Hu, Wu, et al., 2018). In the field of bibliometrics, Word2Vec model shows promising predictive performances on topic identification (Onan 2020). It is capable of depicting the semantic associations among different keywords in a certain corpus (Hu, Wu, et al., 2018), containing the public attributes inherited from the words in context, but also keeps the word its own private attributes, reflecting multiple semantic information of keywords. Furthermore, it could dynamically model the semantic meanings behind the low-frequency keywords, and lessen the synonym problems and sparse distribution of linkage between keywords in networks (Hu, Wu, et al., 2018).

Word2Vec consists of two models: a Skip-Gram model and a continuous bag-of-words model (CBOW). The CBOW model predicts a word based on the current context words, while the Skip-Gram model predicts the context words based on the current word (Zhang et al., 2018). Compared with CBOW, Skip-Gram has been proven to have a tiny advantage with bibliometric data (Hu, Wu, et al., 2018; Zhang et al., 2018).

Community detection

Communities in a network are groups of nodes, which are highly connected to each other than to the rest of the nodes in the network (Yang et al., 2010). Compared with traditional clustering methods, community detection methods show two benefits: (1) Community detection methods are easier to find more technical details with lower data volume demand (Huang et al., 2021); (2) Community detection methods can better identify communities with larger semantic structures in network analysis (Rabitz et al., 2021; Sharma et al., 2021). Clustering aims to group similar patterns into certain categories (Zhang et al., 2018), and many clustering models were developed specifically for bibliometric tasks including co-word analysis (Ding & Chen, 2014), citation/co-citation analysis (Klavans & Boyack, 2017), and co-authorship (Li et al., 2014). However, traditional clustering methods have weaknesses including over-reliance on data pre-processing and human intervention, challenges of initial parameter configuration and issues of local optimization issues (Zhang et al., 2018). In contrary, community detection algorithm is a network-based method, which fully considers the semantic structure and topological structure of the network (Huang et al., 2020; Xie et al., 2013). Community detection methods are widely applied in keyword networks, social networks and other complex network analytics (Qi et al., 2021). Network structure can capture the relationships between technical concepts in a research area at an aggregate level (Balili et al., 2020), improving the accuracy of topic identification (Arruda et al., 2016). A community in bibliometric analysis can be represented as a cluster of words that are densely connected to one another (Balili et al., 2020; Ding, 2011). Many researchers adopted community detection algorithms for topic identification. Huang et al. (2021) used community detection Smart Local Moving to identify emerging topics in information science; Kiss et al. (2021) identified the most relevant sport nutrition topics by running a community detection algorithm on the proximity network constructed via network text analysis.

Scholars have proposed many community detection algorithms, which are mainly divided into six categories (Newman and Girvan 2004; Fortunato, 2010; Symeon et al., 2012): (1) divisive algorithm (Newman and Grivan 2004; McCain, 2008), (2) modularity-based method (Rees & Gallagher, 2012), (3) model-based method (Qiu & Lin, 2011), (4) local community detection method (Branting, 2012), (5) feature-assisted method (Wasserman & Faust, 1994), and (6) spectral and clustering method (Mathieu & Gibson, 1993; Symeon et al., 2012).

Modularity-based method is a widely adopted algorithm for community detection (Newman and Girvan 2004; Wang, Liu, et al., 2014). The high modularity means thickset connections within communities but sparse connections between communities in networks (Wang, Liu, et al., 2014). The modularity method promotes the further development of community detection algorithms, and some new methods have been proposed by computer scientists and physicists. Blondel et al. (2008) proposed a Fast Unfolding algorithm, which is a heuristic method based on modularity optimization (Blondel et al., 2008). This method has been found to be robust and effective to identify topics in complex networks, and the results generated from it are useful for studying the structure of science (Blondel et al., 2008; Lancichinetti & Fortunato, 2009).

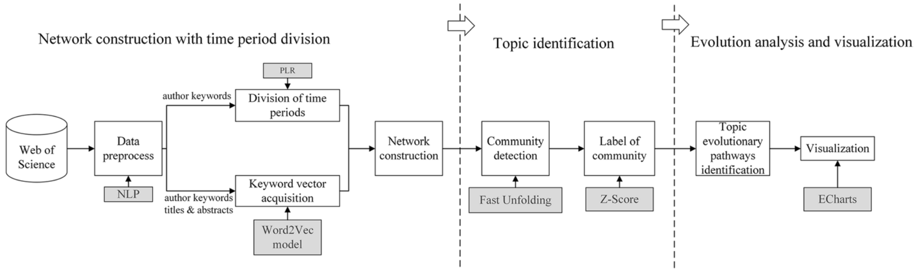


Fig. 1 Research framework for identifying topic evolutionary pathways

Methodology

This paper presents a method based on network analytics for identifying topic evolutionary pathways, which consists of three major steps: (1) Construction of a series of networks with sequential time periods by introducing piecewise linear representation and a Word-2Vec model; (2) Topic recognition based on community detection and (3) Identification and visualization of topic evolutionary pathways. The research framework of the proposed methodology is given in Fig. 1.

Network construction with time period division

Data collection and pre-processing

The data used in this paper are articles in a particular field gathered from Web of Science (WoS). A natural language processing (NLP) technique is first used to retrieve author keywords, titles, abstracts and published years from these articles, and then a term clumping process is carried out, which includes removing the garbled code in the title and abstract, removing the author keywords with garbled code and XML label in the keywords, etc. (Zhang et al., 2014).

Division of time periods based on piecewise linear representation

In this section we segment the keyword sequence based on the piecewise linear representation. First, the number of keywords in the same time period (month, season, year, etc.) is counted to form a keyword quantity sequence $K = \{k_1, k_2, \ldots, k_t, \ldots, k_l\}$, where k_t represents the number of keywords in the research field within the time period t ($1 \leq t \leq l$), as shown in Fig. 2a. In order to remove noise and prevent keywords with the same semantics from being repeated counted, this paper uses term clumping (Zhang et al., 2014) to clean keywords.

After that, the piecewise linear representation is used to fit K into a piecewise linear structure K_{PLR} with connection of start and end, as the red line shown in Fig. 2b.

Here, K_{PLR} can be expressed as:

$$K_{\mathrm{PLR}} = \left\{ L_1\left(k_1, k_2, \ldots \ldots, k_{t_1}\right), L_2\left(k_{t_1+1}, k_{t_1+2}, \ldots \ldots, k_{t_2}\right), \ldots \\ L_i\left(k_{t_{i-1}+1}, k_{t_{i-1}+2}, \ldots \ldots, k_{t_i}\right) \ldots, L_s\left(k_{t_{s-1}+1}, k_{t_{s-1}+2}, \ldots \ldots, k_l\right) \right\} \tag{1}$$

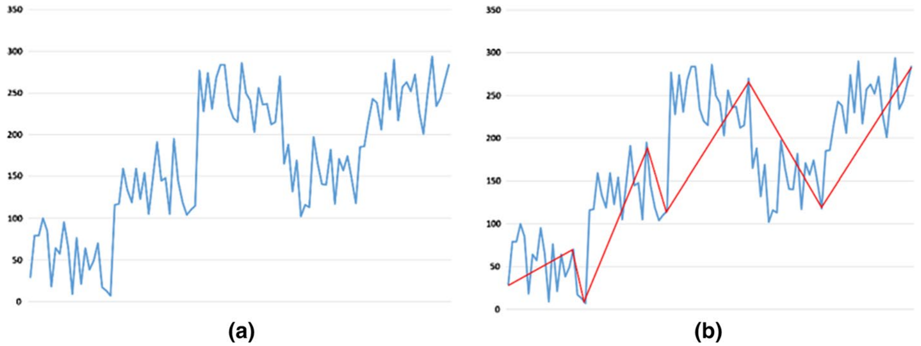


Fig. 2 Schematic diagram of piecewise linear representation

In this formula $L_i\left(k_{t_{i-1}+1}, k_{t_{i-1}+2}, \dots\dots, k_{t_i}\right)$ represents the i $(1 \le i \le s)th$ line segment ofK_{PLR}, which starts at $t_{i-1}+1$ and ends att_i. The connection points of these line segments are Trend Turning Points(TTP), which could be expressed as:

$$\mathrm{TTP} = \left\{1, t_1, t_2, \cdots t_i \cdots, t_n\right\} \tag{2}$$

where t_i denotes the time point at the end of the i $(1 \le i \le n)th$ segment. These time pints represent the beginning of a new turning trend, and time periods can be divided based on these time points. The time periods could be shown as:

$$T = \left\{T_1, T_2, \cdots T_i \dots, T_n\right\} \tag{3}$$

where T_1 represents the time period in which the start time is 1 and the end time is t_1, $T_\mathrm{i}(2 \le i \le n)$ represents the time period with a start time of $t_{\mathrm{i}-1}+1$ and an end time of t_i.

In the process of piecewise linear representation, the setting of parameter s, which represents the number of segments, is very critical. The smaller parameter s is, the larger the overall fitting error will be; while the bigger parameter s is, the more noise will be introduced. In order to find out the best parameter s, this paper refers to the study of Chen et al. (2015).

Firstly, the range set of parameter s is determined, after that, the root mean square error (RMSE) of each value in this set is calculated and stored in a sequence, in which RMSE decreases with the increase of parameter s. RMSE is an index that can intuitively show the fitting effect of the observed value to the true value. In this paper, it is used to measure the error between the curve after piecewise linear fitting and the original curve, and its formula is as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(k_t - plr_t\right)^2} \tag{4}$$

where k_t and plr_t are the data points respectively on the original curve and the curve after piecewise linear fitting at time node t,N represents the total number of data points.

Then, the value of parameter s corresponding to the maximum value of the derivative of RMSE sequence is obtained, which represents the optimal number of segments, recorded as s_{AD}:

$$s_{\mathrm{AD}} = \max\left|\frac{\Delta \mathrm{RMSE}}{\Delta s}\right| \tag{5}$$

Word vector acquisition based on Word2Vec model

This section aims to generate a vector representation for each keyword by using a Word2Vec model (Mikolov et al., 2013). In essence, Word2Vec model is a neural network model, which transforms each word into an N-dimensional numeric vector (Ding et al., 2020).

The training procedure based on the Word2Vec model is as follow: First, text acquired from abstract and title in each time period are segmented into sentences and words; and then the segmented sentence sequences are input into the Word2Vec model, using skip-gram model to be trained as a corpus because of its higher accuracy; finally, through the well-trained Word2Vec model, each keyword obtained from section "Data collection and pre-processing" is mapped as a vector originating from a point in a multi-dimensional semantic space.

Network construction

With the well-segmented time periods and keyword vectors, the objective of this section is to construct a keyword network based on semantic similarity. First, we assume $W^{T_i} = \left\{ w_1^{T_i}, w_2^{T_i}, \ldots, w_i^{T_i}, \ldots, w_n^{T_i} \right\}$ as the term-clumped keyword set in time period T_i, where $w_i^{T_i}$ represents the *i th* keyword belonging to the keyword set W^{T_i}, so that the keyword network in time period T_i can be defined as $G_{T_i} = \{V, E\}$, where V and E represent the vertices and edges in G_{T_i}, which are consist of the keyword in W^{T_i} and the cosine value between keyword vectors(weight) respectively.

Finally, the keyword networks in all time periods constitute the semantic network G, shown as:

$$G = \left\{ G_{T_1}, G_{T_2}, \ldots, G_{T_i}, \ldots, G_{T_n} \right\} \tag{6}$$

where G_{T_i} is the keyword network in time period T_i.

Topic identification based on community detection

After constructing the network, the purpose of this part is to recognize the topics based on community discovery, which includes two sections: (1) Community detection based on Fast Unfolding and (2) Label of community based on Z-Score.

Community detection based on Fast Unfolding

Since community detection algorithm could fully consider the semantic structure and topological structure of the network, it is selected for the keyword network constructed in our study. In this section we use Fast Unfolding algorithm to identify communities in each keyword network. Fast Unfolding is a community detection algorithm based on maximization of the modularity, which has achieved better performance in large complex networks (Blondel et al., 2008). Modularity is an index, which can measure the tightness of connections within communities and the sparsity of connections between communities. The higher the modularity is, the better the result of community detection is, that is, the internal connection is closer and the connection between communities is sparse (Newman, 2012). It can be recorded as R:

$$R = \frac{1}{2A}\sum_{i,j}\left[A_{ij} - \frac{N^i N^j}{2A}\right]\delta\left(M_i, M_j\right) \tag{7}$$

where in each keyword network A is the sum of the weights of all edges, A_{ij} represents the weight of the edge between node i and node j, N^i is the sum of the weights of all edges connected to node i, N^j is the sum of the weights of all edges connected to node j, $\delta\left(M_i, M_j\right)$ is used to indicate whether node i and node j are in the same community: if it's true, the value is 1; otherwise, it is 0. In the part of case study, we will use R index to measure the performance of community detection.

In order to highlight the community structure and avoid the introduction of noise from weak-related and negative-related key pairs (key pairs with negative vector cosine values), referring to the study of Zeng et al. (2019), edges between some weak-related pairs of keywords are removed by using a threshold δ,which is increased from zero to 0.5 (step size 0.05). A corresponding modularity for each time period will be calculated, and that δ with the greatest modularity is used as the pruning threshold in this paper.

Label of community based on Z-Score

After identifying the communities in the network, we need to set a label for each community. Although some studies selected hot keywords (i.e., high-frequency keywords) to represent one community (Chen et al., 2016; Su et al., 2015), they may be too general or ambiguous to indicate a specific area or concept (Tseng et al., 2007; Zhou et al., 2019) and neglect low-frequency terms. Referring to the methods of Wang, Liu, et al. (2014), Z-Score index is used to rank the internal nodes of each community, since it could achieve good performance at the regional level rather than the global level (Guimerà et al., 2007). The node with the highest Z-Score value is selected as the label of the community, and the community could be treated as the topic finally. The formula are as follows:

$$z_i = \frac{N^i_M - B/M^o}{\sqrt{Q/M^o - \left(B/M^o\right)^2}} \tag{8}$$

$$B = \sum\nolimits_{(j\in M)} N^j_M \tag{9}$$

$$Q = \sum_{j\in M}\left(N^j_M\right)^2 \tag{10}$$

where z_i is the Z-Score value of the ith node in community M, N^i_M represents the sum of the weight of the edges between the ith node and other nodes in community M, M^o represents the number of nodes in community M. The higher the Z-Score of a node is, the closer the relationship between the node and other nodes in the community is, and the more representative the node is. Referring to the Study of Guimerà et al. (2007), nodes with Z-Score greater than or equal to 2.5 can serve as the core node of the community.

Identification and Visualization of topic evolutionary pathways

Once the topics are indicated by network communities, the evolution analysis of a research field turns into the analysis of changes of communities (topics). In this section we analysis evolution of research topics by identifying and visualizing evolutionary relationships among topics.

Topic evolution pathways identification

This section identifies the relationships based on the similarity between the corresponding communities. Core nodes are the most representative nodes in a community, and they are also the key to the development and change of a community (Wang, Liu, et al., 2014). Therefore, this paper uses core nodes to measure the similarity between communities. We assume M_{t+1} and M_t are two topics in the time periods t + 1 and t respectively, and the similarity $\mathrm{HS}\left(M_t, M_{t+1}\right)$ between M_{t+1} and M_t can be calculated as:

$$\mathrm{HS} = (M_t, M_{t+1}) = \mathrm{Sim}\left(H\left(M_t\right), H\left(M_{t+1}\right)\right) \tag{11}$$

where $\mathrm{H}\left(M_t\right)$ and $\mathrm{H}\left(M_{t+1}\right)$ are the set of core nodes in M_t and M_{t+1}.

Considering that the similarity between topics largely depends on the semantic similarity, this paper measures $Sim\left(H\left(M_t\right), H\left(M_{t+1}\right)\right)$ by adopting the weighted average of vector cosine value, based on the word vector of core nodes and the given weight of the corresponding Z-Score value.

In order to unify dimensions, value of Z-Score of each community is standardized. Taking community M_t as an example, W_t is assumed as a keyword node in set $\mathrm{H}\left(M_t\right)$, and Min–max normalization (Isler & Kuntalp, 2010) is used to standardize the value of Z-Score corresponding toW_t, which is shown as follows:

$$Z'_{W_t} = \frac{Z_{W_t} - Z_{\min}}{Z_{\max} - Z_{\min}} \tag{12}$$

where Z'_{W_t} is the value of Z-Score corresponding to keyword W_t after standardization, and Z_{W_t} is the value of Z-Score corresponding to W_t, $Z_{\max}$ and $Z_{\min}$ represent the maximum and minimum of the value of Z-Score corresponding to the nodes in $\mathrm{H}\left(M_t\right)$ before standardization, respectively.

After that, the calculation process of $\mathrm{Sim}\left(H\left(M_t\right), H\left(M_{t+1}\right)\right)$ can be expressed as follows:

$$\mathrm{Sim}\left(H\left(M_t\right), H\left(M_{t+1}\right)\right) = \frac{1}{\mathrm{sum}\left(Z'_t * Z'_{t+1}\right)} \sum_{\substack{W_t \in H\left(M_t\right) \\ W_{t+1} \in H\left(M_{t+1}\right)}} Z'_{W_t} \cdot Z'_{W_{t+1}} \cdot \cos\left(v_{W_t}, v_{W_{t+1}}\right) \tag{13}$$

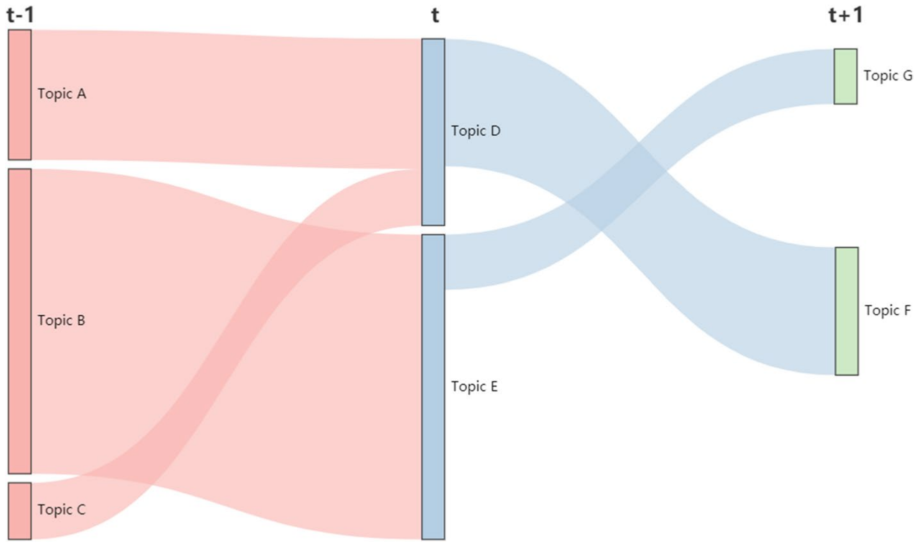


Fig. 3 Visualization design of topic evolutionary pathways

$$\mathrm{sum}\left(Z_t^{'} * Z_{t+1}^{'}\right) = \sum_{\substack{z_i \in Z_t^{'} \\ z_j \in Z_{t+1}^{'}}} z_i \cdot z_j \tag{14}$$

where $Z_t^{'}$ and $Z_{t+1}^{'}$ represent the standardized Z-Score value set corresponding to the core keyword set $\mathrm{H}\left(M_t\right)$ and $\mathrm{H}\left(M_{t+1}\right)$ respectively; z_i and z_j represent the elements of $Z_t^{'}$ and $Z_{t+1}^{'}$ respectively; and v_{W_t} is the word vector obtained from W_t based on the Word2Vec model.

Based on the value of similarity between topics, the evolutionary relationship among topics in adjacent time periods can be analyzed, that is, the precursor and successor of each topic can be determined. It is assumed that there exists an evolutionary relationship if the similarity between two topics in adjacent time periods is greater than a given threshold δ. Taking a community M_{t+1} in time period t + 1 as an example, the predecessor Pre $\left(M_{t+1}\right)$ can be expressed as:

$$\mathrm{Pre}\left(M_{t+1}\right) = \left\{M_t | M_t \in G_t, HS\left(M_t, M_{t+1}\right) > \delta\right\} \tag{15}$$

where G_t represents the keyword network in the time period t, and M_t represents a community in G_t.

Visualization of topic evolutionary pathways

It is important to visualize the topic evolutionary pathways for understanding the dynamics of the corresponding research field. Sankey Diagram has been employed in the studies of topic evolution (Nguyen et al., 2018; Pépin et al., 2017; Ren et al., 2018). Figure 3 provides an example of Sankey Diagram in our study, which is realized by ECharts.js.[1] Referring to Pépin et al. (2017), the visualization design of topic evolutionary pathways is as follows:

[1] https://github.com/apache/echarts.

(1) Rectangles represent topics, and the size of the rectangle for a topic is proportional to the sum of the similarity between the topic and other related topics. In each time period, the topic is arranged from top to bottom according to the number of keywords it contains. The more keywords the topic contains, the higher the position.
(2) The connecting lines represent the evolutionary relationships between topics at two consecutive periods, and the width of the line is proportional to the similarity between the topics at both ends. Following the study of Schwartz et al. (2015), we set the threshold δ as 0.7 and assume that there is an evolutionary relationship between two topics when the similarity between them is greater than 0.7.

Palla et al. (2007) proposed six types of activities in the evolution of communities. Referring to this study, we also classify the evolution relationship of topics into six forms: Birth, Growth, Merging, Contraction, Splitting and Death.

- Birth: topics that do not exist in the time period of t, and emerge during the time period of t + 1;
- Growth: topics exist in the time period of t, and will exist in the time period of t + 1 with a larger scale;
- Merging: two or more topics in the time period of t, and merge into a new topic in the time period of t + 1;
- Contraction: topics exist in the time period of t, and will exist in the time period of t + 1 with a smaller scale;
- Splitting: topics in the time period of t and are split into two or more new topics in the time period of t + 1;
- Death: topics that exist in the time period of t, and do not exist during the time period of t + 1.

Empirical study

As a relatively mature field, IS ensures enough training data for the model. Further, IS is a typical interdisciplinary discipline (Holland and George 2008) that evolves frequently and complexly. Therefore, we chose Information Science (IS) as the field to conduct an in-depth analysis and examined our framework further.

Network construction

Data collection and pre-processing

Following Hou et al. (2018)'s study, we selected ten leading journals in IS and downloaded 10,135 papers published in these journals between 2010 and 2019 from the Web of Science (WoS) as our inputs. The description of the dataset is provided in Table 1[2].

[2] JASIST changed its name from Journal of the American Society for Information Science and Technology to Journal of the Association for Information Science and Technology in 2014.

Table 1 Sources of journal data

Journal Name	No. of Papers
Scientometrics	3222
Information Research an International Electronic Journal	1083
Journal of Informetrics	879
Information Processing & Management	801
Journal of Documentation	660
Journal of Information Science	592
Library & Information Science Research	406
Research Evaluation	371
Journal of the Association for Information Science and Technology	2121
Total	10,135

Table 2 The fitting results of three methods

Methods	s	RMSE	Standardized s	Standardized RMSE	AVE
Sliding window	6	64.2653	0.3333	0.1269	0.2301
Top-down	8	62.9382	1	0	0.5
Bottom-up	5	73.3984	0	1	0.5

AVE denotes the average of the two indexes (s and RMSE) after standardization

The natural language processing function within VantagePoint[3] (VP) was applied to the entire dataset. Following the studies by Hu, Wu, et al. (2018) and Chae et al. (2020), 31,523 author keywords were extracted, and a small number of cases without author keywords were deleted from the entire dataset. According to the study of Zhang et al. (2014), the term clumping process was then used to remove the keywords with garbled code and XML labels in the author keywords, and a total of 31,276 valid author keywords remained. In order to segment the keyword sequence and divide time periods, the valid keywords were divided into 120 slices in month.

Division of time periods

Three common methods of piecewise linear representation (Keogh et al., 2001) (sliding window algorithm, top-down algorithm, and bottom-up algorithm) were applied to identify trend turning points, and the method with the best performance was selected to divide time periods at last.

[3] VantagePoint is a text mining visualization software for bibliometric data (such as scientific paper patents and academic project applications). Please visit the website for detail: www.thevantagepoint.com.

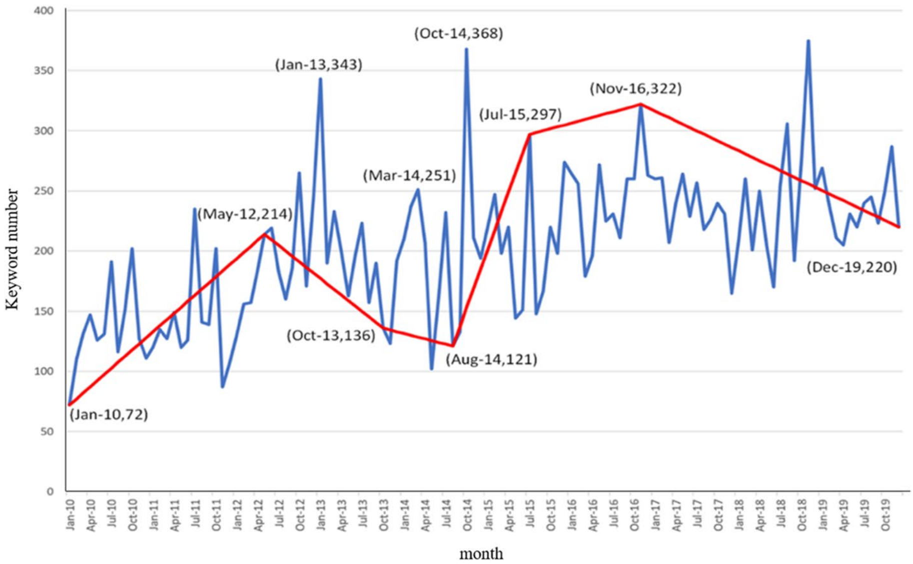


Fig. 4 The piecewise linear fitting results of keyword quantity sequence

Table 3 Time period of keyword quantity sequence

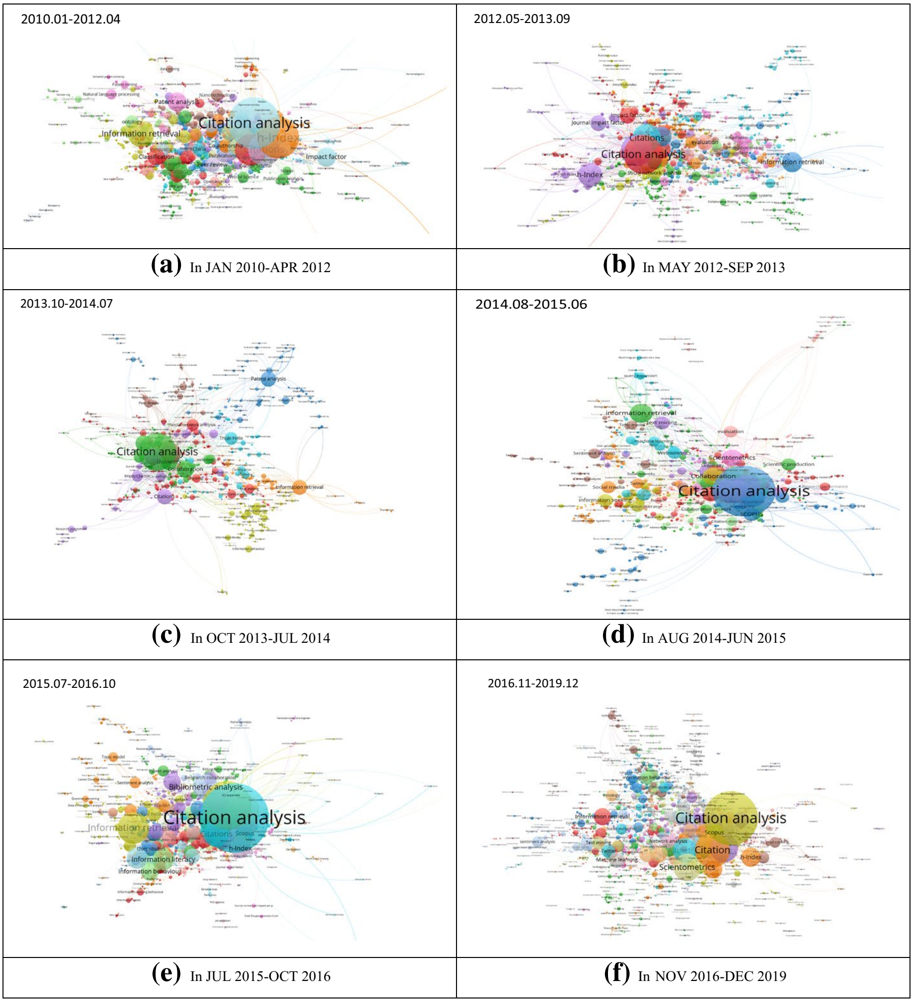
Period number	Starts	Ends
1	JAN 2010	APR 2012
2	MAY 2012	SEP 2013
3	OCT 2013	JUL 2014
4	AUG 2014	JUN 2015
5	JUL 2015	OCT 2013
6	NOV 2016	DEC 2019

Following the study of Chen et al. (2015), we set the parameter s (the optimal number of segments) as a range from 2 to 20. Following the steps in Methodology section, we can obtain parameter s corresponding to the three methods and the corresponding RMSE. In order to unify the dimensions, two indexes, s and RMSE, were standardized by the Min–max normalization. Lastly, we used the average value of the two standardized indexes to evaluate the fitting results of the three methods. The results are shown in Table 2.

The sliding window algorithm with minimum average (0.2301) was selected and the fitting result is shown in Fig. 4, from which we can find that 6 periods were divided. The start and the end of each time period are shown in Table 3.

Table 4 The number of keywords in each time period

Period number	Period	#Keyword
1	JAN 2010-APR 2012	3002
2	MAY 2012-SEP 2013	2608
3	OCT 2013-JUL 2014	2051
4	AUG 2014-JUN 2015	2207
5	JUL 2015-OCT 2016	3229
6	NOV 2016-DEC 2019	4499



(a) In JAN 2010-APR 2012

(b) In MAY 2012-SEP 2013

(c) In OCT 2013-JUL 2014

(d) In AUG 2014-JUN 2015

(e) In JUL 2015-OCT 2016

(f) In NOV 2016-DEC 2019

Fig. 5 Networks in the field of information science by time slices

Table 5 Community detection result in each time period

Time period	Pruning threshold δ	Community number	R index
JAN 2010-APR 2012	0.45	31	0.6333
MAY 2012-SEP 2013	0.40	25	0.5224
OCT 2013-JUL 2014	0.30	23	0.4576
AUG 2014-JUN 2015	0.35	23	0.4253
JUL 2015-OCT 2016	0.35	24	0.5347
NOV 2016-DEC 2019	0.30	28	0.4746
Total	–	154	–
Mean Modularity		–	0.5080

Word vector acquisition

In this section, we acquired a vector representation for each keyword by a Word2Vec model.

There were plenty of phrases in the keyword sets, such as “Information Retrieval” and “Citation Analysis”, which could not generate word vectors by Word2Vec thus those keywords need to be converted from phrase form into Camel-Case form for model training. For instance, “Network analysis” was converted into “NetworkAnalysis”. According to the study of Wang et al. (2015), we did some parameters setting, such as, “vector size = 300”, “window size = 7” and “minimum word frequency = 3”.

The keywords were divided based on the time setting in Table 3 and those keywords with word frequency less than 3 were removed. The number of keywords in each period is shown in Table 4.

Network construction

Further, the cosine values between the word vectors corresponding to keywords in six time periods were calculated and the keyword similarity matrixes were constructed. Using pajek,[4] we then imported the six similarity matrixes and saved them as a network file compatible with VOSviewer,[5] as shown in Fig. 5.

Topic identification

Community detection

In this section, we implemented our community detection process with Fast Unfolding algorithm. Following the studies conducted by Newman (2012) and Zeng et al. (2019), R index, namely Modularity index (Eq. (7)), was used to measure the performance of community detection. And the result of community detection in each time period is shown in Table 5.

[4] https://mrvar.fdv.uni-lj.si/pajek/.

[5] https://www.vosviewer.com/.

Table 6 Z-Score distribution of the community in MAY 2012-SEP 2013

ID	Keyword	Z-Score
1	Citation Analysis	7.6157
2	Citations	7.5343
3	H-Index	6.3953
4	Journal Impact Factor	5.0123
5	Research Evaluation	3.7920
6	Scientometrics	3.6293
7	Impact Factor	3.3852
8	Evaluation	3.2717
9	Webometrics	2.8276
10	Peer Review	2.6462
11	Indicators	2.5649
12	Web of Science	2.5022

Table 6 only presents keywords with Z-Scores greater than 2.5 (Guimerà et al., 2007)

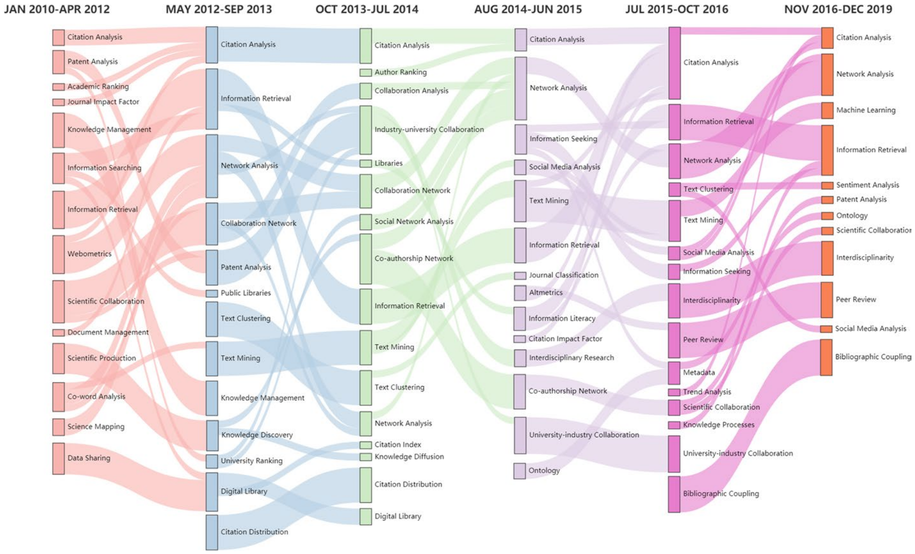


Fig. 6 Topic evolutionary pathways (Piecewise linear representation)

Pruning threshold δ denotes a minimum value of community similarity, and the edges with weight lower than δ in the network were removed. Following the study of Zeng et al. (2019), we obtained a pruning threshold for each network, and finally detected

154 communities in total. As shown in Table 5, the average modularity of our method is 0.5080, and the high peak modularity is 0.6333, which indicates that community detection algorithm has achieved better performance (Newman, 2004).

Label of community

This section aims to add a label to each community. We calculated the corresponding Z-Score value for each node in 154 communities. Referring to the study of Guimerà et al. (2007), nodes with Z-Score greater than or equal to 2.5 can be served as the core node of the community. Here, we took one of communities, which has the largest keyword number with 455in MAY 2012-SEP 2013, as an example to show the result of community label (Table 6).

From Table 6, the Z-Score of "Citation Analysis" was 7.6157, which was the highest value in the community. Consequently, we regarded "Citation Analysis" as the label of this community.

Topic evolutionary pathway analysis

This section aims to analyze the evolution of topics by furtherer identifying and visualizing the relationships among topics.

Pathway identification and visualization

According to the study of Schwartz et al. (2015), we measured the topic similarity between adjacent time periods and assumed that there is an evolutionary relationship between two topics when their similarity is greater than 0.7. The topic evolutionary pathways in the field of information science are visualized by ECharts.js,[6] as shown in Fig. 6.

Brief insights of topic evolutionary pathways

Figure 6 addressed concerns on scientific topics in IS and their evolutionary relationships in a landscape.

(1) There are three clusters on IS, indicating its key research topics and developmental trends:

- **Measurement** This cluster includes quantitative researches such as bibliometrics, informetrics, scientometrics and webometrics. For example, "Citation Analysis" has been running through the whole period in the field of information science, and "Academic Ranking" and "Author Ranking" merged with "Citation Analysis" in the evolution process. They are all common methods in bibliometrics, which could help to quantitatively assess the impact of journals, scholars, and scientific research.

[6] https://github.com/apache/echarts.

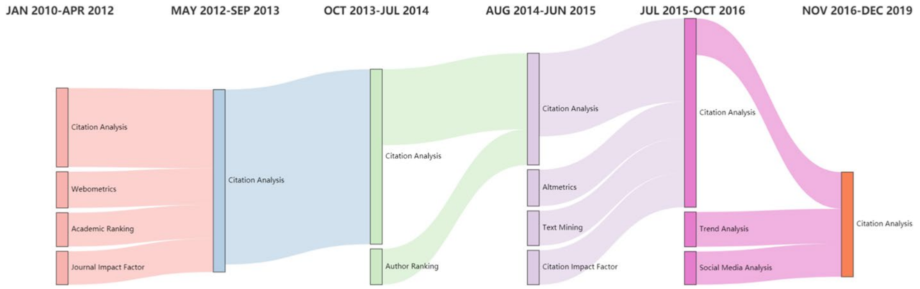


Fig. 7 Pathway of "Citation Analysis"

Therefore, it could be concluded that quantitative research is one of the vital tasks and inevitable trends in the field of information science.

- **Management** This cluster includes topics such as "Knowledge Management", "Document Management", along with further evolved topics, i.e., "Information Retrieval" and "Journal Classification", which indicates that information science and technology has been widely used to deal with real-world management problems, especially in the field of science, technology, and innovation management based on scientific data.
- **Technology** A combination of mathematical models, computer algorithms, and deep learning techniques is shown in this cluster, for instance, "Text Mining", "Text Clustering" and "Machine Learning". It is apparent that data science methods have constantly provided new solutions to IS, gradually replacing or combining with traditional analysis methods to generate new topics.

(2) Six evolutionary relationships among topics have been investigated on IS:

- **Merger and splitting** There were five research topics in the period of OCT 2013-JUL 2014, including "Collaboration Analysis", "Collaboration Network", "Social Network Analysis", "Co-authorship Network" and "Network Analysis", merging into "Network Analysis" in the period of AUG 2014-JUN 2015; While, "Text Mining" split into "Social Media Analysis" and "Text Mining" in the period of AUG 2014-JUN 2015.
- **Birth and death** "Citation Distribution" emerged as a new research topic in the period of MAY 2012-SEP 2013 and disappeared in the period of OCT 2013-JUL 2014. "Information Seeking" emerged as a new research topic in the period of AUG 2014-JUN 2015; while the topic of "University-industry Collaboration" disappeared in JUL 2015-OCT 2016.
- **Growth and contraction** Running through all time periods, "Citation analysis" constantly absorbed predecessor topics and produced successor topics, thus the sizes of its corresponding rectangles were also expanding and shrinking, which respectively reflects the growth and contraction of the topic.

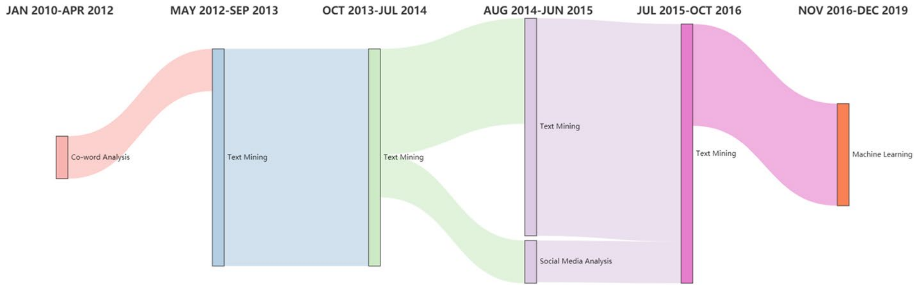


Fig. 8 Pathway of "Text Mining"

Case study on the pathways of "citation analysis" and "text mining"

In this section, we focus on two specific evolutionary pathways to conduct further investigations. The analysis in this part can also help (a) to demonstrate the rationality of time period division method, and (b) to verify the reliability of our results.

As shown in Fig. 7, pathway of "Citation Analysis" is taken as an example.

- Four topics in JAN 2010-APR 2012, "Citation Analysis", "Academic Ranking", "Journal Impact Factor" as well as "Webometrics", merged into "Citation Analysis" in MAY 2012-SEP 2013, which indicates that citation analysis was increasingly used in journal impact and webometrics. For example, Vaio and Weisdorf (2009) applied citation analysis to rank 12 international academic economic history journals; Steven and Greenberg (2011) addressed concern on citation distortion to improve the accuracy of citation analysis in academic ranking.
- "Text Mining" integrated into "Citation Analysis" in JUL 2015-OCT 2016, because text mining techniques can transform text into data and knowledge, which is complementary to citation analysis (No et al., 2015). For example, Kralj et al. (2015) presented a text mining approach in a citation network to improve the accuracy of paper categorization.
- "Altmetrics" and "Social Media Analysis" both integrated into "Citation Analysis" in NOV 2016-DEC 2019, which embodies the contribution of web text to citation analysis. For instance, Sud and Thelwall (2014) indicated that Altmetrics and social media text may help find out some papers which are important but uncited, making up the limitations of traditional citation analysis.

As shown in Fig. 8, pathway of "Text Mining" demonstrates the increasing contributions of text mining techniques to IS.

- There was a topic of "Co-word analysis" in JAN 2010-APR 2012 and it evolved into the topic of "Text Mining" in MAY 2012-SEP 2013. Co-word analysis is a content analysis technique based on keyword co-occurrence, and during this period, many scholars combined text mining technique with co-word analysis. For example, integration of co-word analysis and other text mining techniques is helpful for researchers to get the concept network and developmental tendency in a certain field (Yang et al., 2012); Sharef et al. (2013)

Table 7 Validation settings

Validations	Experimental Settings		
	Network Construction	Time Division	Label of Topic
Our Framework	Keyword Network	PLR	Community detection + Z-Score index
CV1	Keyword Network	Fixed window	Community detection + Z-Score index
CV2	Co-word Network	PLR	Community detection + Z-Score index
EV	Keyword Network	PLR	Community detection + Hot keywords

CV1 and CV2 represent the two comparison study-based validations, respectively, and EV represents the expert knowledge-based validation

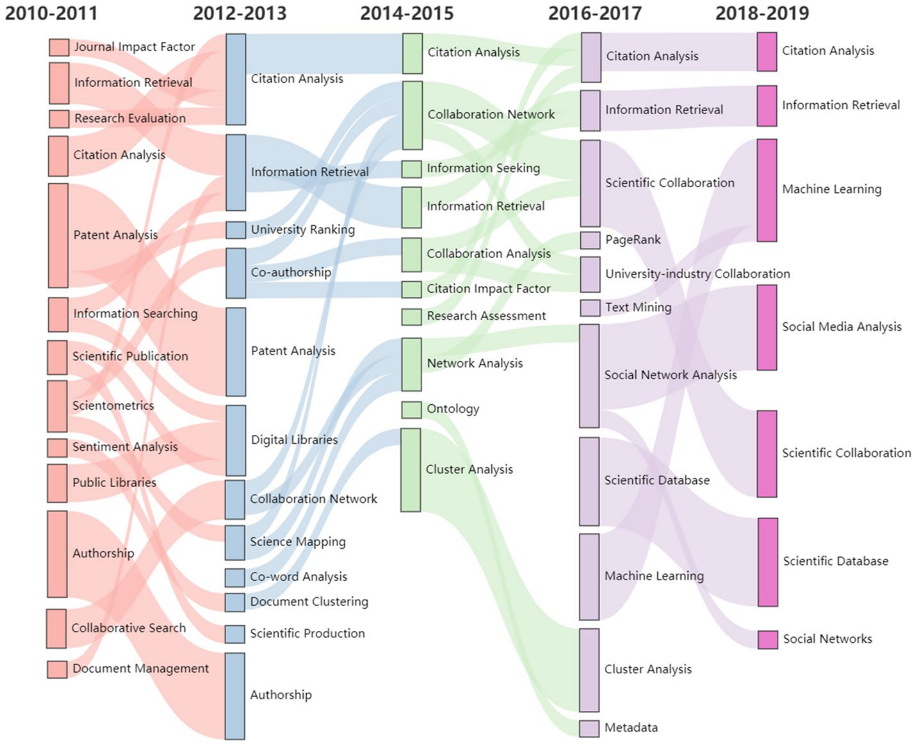


Fig. 9 Topic evolutionary pathways (fixed time window method)

applied text mining and co-word analysis to mine the context association between keywords and proposed a conceptually related lexicon clustering method.

- "Text Mining" split into "Social Media Analysis" and "Text Mining" in the period of AUG 2014-JUN 2015, which shows that "Social Media Analysis" was increasingly using "Text Mining" method and becoming a branch of it. Many studies in this period could confirm our assumption. For example, text mining techniques made great contributions to competitive intelligence, sentiment analysis and hot topic identification, which were common researches of social media analysis (Gémar & Jiménez-Quintero, 2015); Text mining

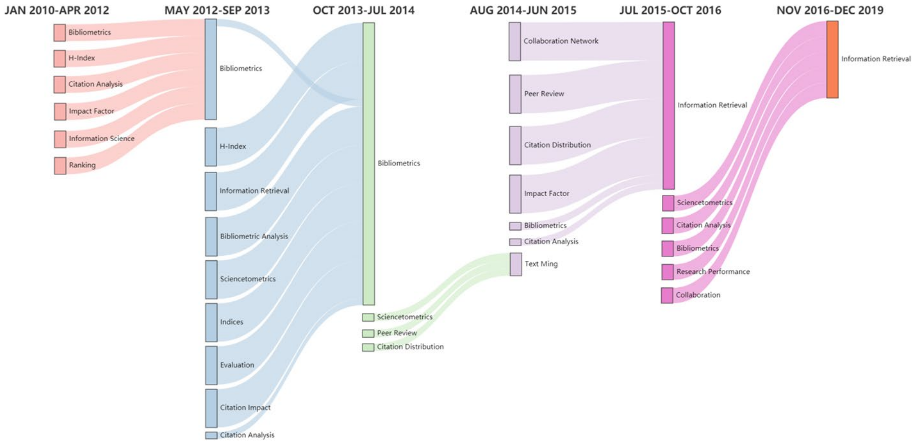


Fig. 10 Topic evolutionary pathways (Co-word Networks)

techniques could help extract vast data, capture unstructured information and encode sentiment of comments from social media and online platforms (Moreno & Terwiesch, 2014); Hu (2014) applied text mining techniques to dig out students' personal personalities and opinion tendencies in social media, which reflects that social media analysis has been one of important applications of text mining.

- It is worth noting that with the maturity of machine learning techniques in recent years, scholars increasingly apply text mining to their researches, thus it is finally included in the category of machine learning. Text mining has developed from simple mining methods such as word frequency statics and word segmentation to the technology combined with neural network and word embedding, which could mine the deep meaning of the text (Verma, 2017); Zhang et al. (2018) proposed a kernel k-means clustering method incorporated with a word embedding model and confirmed that machine learning can help text mining. These evidences explain the evolution of "Text Mining" to "Machine Learning" in NOV 2016-DEC 2019.

Validation

In this part, we conducted two comparison study-based validations and one expert knowledge-based validation to evaluate the reliability of our framework. The design of our validation is listed in Table 7.

(1) In the first comparison study-based validation (CV1), we compared the performance of our method with fixed time window method. Referring to the study of Sun et al. (2007), we set a time period of two years as one time window and divided the keyword sets into five parts, and the topic evolutionary pathways in information science from 2010 to 2019 based on fixed time window method is depicted in Fig. 9.

Obviously, compared with Fig. 6, many important topics are lost in Fig. 9, such as, "Technological Transition" in the period of MAY 2012-SEP 2013, "Citing Behavior" and "Concept Map" in the period of OCT 2013-JUL 2014, as well as "Knowledge Organization System" in the period of AUG 2014-JUN 2015. We could also find that there are wrong pathways identified in Fig. 9. For example, the topic "Authorship" in

Table 8 Comparisons of topic labels

Topic	SZ	SF	Supplement Keywords
#1	Patent Analysis, Patent Citation, Innovation, Data Mining, Collaboration	Publication Analysis, Patent Analysis,Library and Information Science, Co-Link Analysis, H-Core	H-Index, Patent Stock, Non-Journal Publications, Co-Term, Network Theory
#2	Citation Analysis, Research Evaluation, University Ranking, Research Collaboration, Publication Analysis	Bibliometrics, Citation Analysis, Scientometrics, Classification, Higher Education	Collaboration Network,Citation Rate, Journal Articles, Citation Behavior,Librarianship
#3	Information Retrieval, Computer Science, Image Retrieval, Clustering, Search Engines	Bibliometrics,Information Retrieval,Information Seeking, Altmetrics, Information Literacy	Metadata, Community Detection, Data Base, Elsevier, Sleeping Beauty
#4	Social Network Analysis, Co-word Analysis, Citation Network, Collaborative Networks, Co-author Networks	Social Network Analysis, Google Scholar, Concept Clustering, Credibility, Critical Thinking	University System, Scientific Ranking, Academic Research Group, Economics Departments, Domain Analysis
#5	Text Mining, Science Mapping, Mapping, Co-citation Analysis, Bibliographic Coupling	Impact Factor, Evaluation, Information Retrieval, Web of Science, Research Performance	Machine Learning, Data Accuracy, Factor Analysis, Data Accuracy, Model
#6	Network Analysis, Scientometrics, Open Access, Webometrics, Open Source	Network Analysis, Incubators, Longitudinal Study, Author Metric, Co-citation	Co-Citation Network, 3-D Computer Graphs, Anomaly Detection,Citation Potential, Liking Networks

SZ is the set of top five keywords with the highest Z-score value; SF is the set of top five keywords with the highest frequency; underlined keywords are the topic's labels selected based on word frequency or Z-Score

Table 9 Results of criterion #1

Topic	Expert1	Expert2	Expert3	Expert4	Expert5	Avg	Max	Min
#1	4	3	4	4	4	3.8	4	3
#2	5	4	5	4	4	4.4	5	4
#3	3	4	3	2	4	3.2	4	2
#4	4	4	4	5	4	4.2	5	4
#5	3	4	5	3	4	3.8	5	3
#6	3	2	4	4	4	3.4	4	2

Table 10 Results of criterion #2

Topic	Expert1	Expert2	Expert3	Expert4	Expert5	Avg	Max	Min
#1	4	3	4	5	5	4.2	5	3
#2	5	4	3	5	5	4.4	5	3
#3	5	4	4	4	5	4.4	5	4
#4	5	4	3	3	5	4	5	3
#5	4	4	4	4	5	4.2	5	4
#6	3	2	4	3	5	3.4	5	2

Table 11 Results of criterion #3

Topic	Expert1	Expert2	Expert3	Expert4	Expert5	Avg	Max	Min
#1	4	3	3	4	5	3.8	5	3
#2	5	4	4	4	5	4.4	5	4
#3	4	4	4	3	5	4	5	3
#4	5	4	3	4	5	4.2	5	3
#5	5	4	4	4	5	4.4	5	4
#6	4	2	4	4	5	3.8	5	2

the period of 2010–2011 evolved into the topic "Authorship" in the period of 2012–2013, however, the Jaccard Coefficient between these two corresponding keyword sets was as high as 0.92; Similarly, "Scientific Database" in the period of 2016–2017 evolved into "Scientific Database" in the period of 2018–2019, and the Jaccard Coefficient between their corresponding keyword set was 0.95. The experiment shows that both the accuracy of identified topics and the reasonability of identified topic evolutionary pathways of PLR perform better than the fixed time window method.

(2) In the second comparison study-based validation (CV2), the purpose is to compare the performance of our constructed keyword networks with co-word networks. The topic evolutionary pathways in information science from 2010 to 2019 based on co-word networks is depicted in Fig. 10.

Compared with Fig. 6, we could find (1) the number of evolutionary pathways identified in Fig. 10 is very small. Specifically, only 28% of the topics have evolutionary relationships with others; (2) the diversity of evolution pathways in Fig. 10 is insufficient, only identifying three relationship types, i.e., “birth”, “merging” and “death”; (3) the integrity of evolution pathways in Fig. 10 is poor, spanning at most two time periods.

In sum, the results showed that our method played better on the number, diversity and integrity of identified pathways.

(3) In the expert knowledge-based validation (EV), we conducted an experimental comparison of the performance of our methods (community detection algorithm combining with Z-score index) with the use of hot keywords (i.e., high-frequency keywords). Firstly, we randomly selected a topic from each time period and engaged these six topics for validation analysis. Secondly, for each topic, we collected 5 top keywords with the highest frequency as hot keywords and selected 5 top keywords with the highest Z-Score value. Thirdly, to help experts better understand the topic, we further selected other five keywords for each selected topic as supplements. The details are shown in Table 8.

Subsequently, we invited five experts in IS field to review these randomly selected topics and to assess whether our proposed method was a more suitable strategy for labeling topics. Specifically, for each selected topic in Table 8, each expert was asked to score three following criteria.

- Compared with the top 5 keywords with the highest word frequency, can the top 5 keywords with the highest Z-score value better represent the topic?
- Compared with the keyword with the highest word frequency, can the keyword with the highest Z-Score value better represent the topic? When the two keywords (with the highest word frequency/Z-Score value) are identical, compared with the keyword with the second-highest word frequency, can the keywords with the second-highest Z-score value better represent the topic?
- Is it reasonable that the label for the topic should be selected based on Z-Score index rather than word frequency?

The results for each criterion are shown in Tables 9, 10, 11. Here, 5 means excellent agreement, 4 means agreement, 3 means unsureness, 2 means disagreement, and 1 means strong disagreement.

The results in Tables show positive feedbacks of experts on our proposed methods. For example, we could find in Table 11 that the average score of all the experts on the six topics is higher than 3.8, and the maximum score of them is 5, which indicates that experts agree that the topics selected based on community discovery and Z-Score index are more accurate than those selected based on hot keywords. Some experts also further explained the reasons for the scores given: (1) Z-score index could well reflect the connection between one keyword and all other keywords in a topic, which is an important basis for selecting the topic label. (2) Word frequency cannot fully measure the importance of keywords in a topic identified based on semantic information. (3) Hot keywords (selected by word frequency) tend to contain broader semantic concepts, misleading people’s comprehending of the topic evolution.

Discussion and conclusion

This paper proposed a keyword network analytics method to identify and visualize evolutionary relationships among scientific topics. Keyword networks combined with a Word2Vec model show advantages on better capturing semantic and contextual relationships between keywords, not only containing the public attributes inherited from the words in context, but also keeping the word its own private attributes, investigating multiple semantic information of keywords; Further, community detection algorithm, which fully considers the semantic structure and topological structure of the network, was employed together with Z-Score index in the constructed keyword network to identify and label topics, more accurately reflecting knowledge structure; Moreover, piecewise linear representation method was introduced to detect the trend turning points over time and to flexibly divide the time periods, capturing the dynamics of topics. We demonstrated our method using an empirical study of ten IS journals' literature data retrieved from WoS over the time period between 2010 and 2019. Additionally, we conducted two comparison study-based validations and one expert knowledge-based validation to demonstrate the reliability of the proposed method quantitatively and qualitatively.

Possible applications

It is conceivable that our method could be applied to a wide range of bibliometric tasks.

- Our proposed strategy is not only applicable for scientific papers, but also can be applied for topic analysis of patent texts, business news, and other S&Ts data, which is helpful for researchers to grasp scientific frontiers and predict the development trends of S&Ts. For example, the proposed method could be used as an auxiliary tool for strategic decision-making, innovative resource allocation and industrial development planning (Xu et al., 2020). Also, it is possible to help discover new unknown technological opportunities.
- Topic evolution is one of interdisciplinary activities which involve a series of the interactions and evolutions among diverse subjects (Zhang et al., 2017). Therefore, our method has potentials to be integrated with other analytical approaches, such as science maps, to conduct multidisciplinary interaction analysis (Zhang et al., 2018). Taking a discipline, that is, artificial intelligence (AI) for example, our approach can distinguish the development patterns among multiple AI subfields (Qian et al., 2020), including computer vision and natural language processing.

Limitations and future work

Several limitations of our current methods and related future directions are summarized as follows. (1) Since the number of publications may be influenced by a lot of factors in addition to the change of topics, it is essential to explore topic-based time division methods, identifying the trends and contribution levels of different topics to the publication dynamics (Chen, Zhang, et al., 2017). (2) Many network analysis methods in bibliometrics have been employed for topic identification, including co-word networks (Huang et al., 2021), citation networks (Kleminski et al., 2020), bi-layer networks (Zhang et al., 2021), etc., thus topic evolution analysis combining multiple networks may be emphasized in future's work. (3) In this paper, six forms of relationship were distinguished, while it makes sense to further quantitively describe and differentiate different types of topics in the evolutionary pathways, e.g., sleeping beauties (Van Raan, 2004) and ghost city (Hu, Qi, et al., 2018). (4) More latest models with good performance would be tried in our study, such as Bert.

Acknowledgements This work was supported by the National Nature Science Foundation of China Funds [Grant No. 71774013]. Yi Zhang acknowledges supports from the Australian Research Council under Discovery Early Career Researcher Award DE190100994.

References

Arruda, H. F., Costa, L. D. F., & Amancio, D. R. (2016). Topic segmentation via community detection in complex networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science, 26*(6), 063120.

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.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd ACM international conference on machine learning* (pp. 113–120).

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 30*(2), 155–168.

Börner, K., Chen, C., & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology, 37*(1), 179–255.

Branting, L. K. (2012). Context-sensitive detection of local community structure. *Social Network Analysis and Mining, 2*(3), 279–289.

Carmona-Poyato, Á., Fernández-Garcia, N. L., Madrid-Cuevas, F. J., & Durán-Rosal, A. M. (2021). A new approach for optimal offline time-series segmentation with error bound guarantee. *Pattern Recognition, 115*, 107917.

Chae, C., Yim, J. H., Lee, J., Jo, S. J., & Oh, J. R. (2020). The bibliometric keywords network analysis of human resource management research trends: the case of human resource management journals in South Korea. *Sustainability, 12*(14), 5700.

Chang, P. C., Fan, C. Y., & Liu, C. H. (2009). Integrating a piecewise linear representation method and a neural network model for stock trading points prediction. *IEEE Transactions on Systems, Man, and Cybernetics Part c: Applications and Reviews, 39*(1), 80–92.

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

Chen, H., Zhang, G., Zhu, D., & Lu, J. (2015). A patent time series processing component for technology intelligence by trend identification functionality. *Neural Computing and Applications, 26*(2), 345–353.

Chen, H., Zhang, G., Zhu, D., & Lu, J. (2017b). Topic-based technological forecasting based on patent data: A case study of Australian patents from 2000 to 2014. *Technological Forecasting and Social Change, 119*, 39–52.

Chen, J., Chen, J., Zhao, S., Zhang, Y., & Tang, J. (2020). Exploiting word embedding for heterogeneous topic model towards patent recommendation. *Scientometrics, 125*(3), 2091–2108.

Chen, X., Chen, J., Wu, D., Xie, Y., & Li, J. (2016). Mapping the research trends by co-word analysis based on keywords from funded project. *Procedia Computer Science, 91*, 547–555.

Cheng, Q., Wang, J., Lu, W., Huang, Y., & Bu, Y. (2020). Keyword-citation-keyword network: A new perspective of discipline knowledge structure analysis. *Scientometrics, 124*(3), 1923–1943.

Cruz, P., & Cruz, H. (2020). Piecewise linear representation of finance time series: Quantum mechanical tool. *Acta Physica Polonica A., 138*(1), 21–24.

Ding, W., & Chen, C. (2014). Dynamic topic detection and tracking: A comparison of HDP, C-word, and cocitation methods. *Journal of the Association for Information Science and Technology, 65*(10), 2084–2097.

Ding, Y. (2011). Community detection: Topological vs. topical. *Journal of Informetrics, 5*(4), 498–514.

Ding, Y., & Stirling, K. (2016). Data-driven discovery: A new era of exploiting the literature and data. *Journal of Data and Information Science, 1*(4), 1–9.

Ding, Z., Liu, R., Li, Z., & Fan, C. (2020). A thematic network-based methodology for the research trend identification in building energy management. *Energies, 13*(18), 4621.

Érdi, P., Makovi, K., Somogyvári, Z., Strandburg, K., Tobochnik, J., Volf, P., & Zalányi, L. (2013). Prediction of emerging technologies based on analysis of the US patent citation network. *Scientometrics, 95*(1), 225–242.

Firth, J. R. (1957). A synopsis of linguistic theory 1930–55. *Studies in Linguistic Analysis the Philological Society, 1957*, 1–32.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports, 486*(3–5), 75–174.

Gémar, G., & Jiménez-Quintero, J. A. (2015). Text mining social media for competitive analysis. *Tourism & Management Studies, 11*(1), 84–90.

Guimera, R., Sales-Pardo, M., & Amaral, L. A. (2007). Classes of complex networks defined by role-to-role connectivity profiles. *Nature physics, 3*(1), 63–69.

Holland, G. A. (2008). Information science: an interdisciplinary effort? *Journal of Document, 64*(1), 7–23.

Hou, J., Yang, X., & Chen, C. (2018). Emerging trends and new developments in information science: A document co-citation analysis (2009–2016). *Scientometrics, 115*(2), 869–892.

Hu, K., Wu, H., Qi, K., Yu, J., Yang, S., et al. (2018b). A domain keyword analysis approach extending Term Frequency-Keyword Active Index with Google Word2Vec model. *Scientometrics, 114*(3), 1031–1068.

Hu, X. (2014). Using social network analysis and text mining to analyze students' input on social media. *Library & Information Science Research, 32*(3), 732–741.

Huang, G., & Zhou, X. (2016). A piecewise linear representation method of hydrological time series based on curve feature. In *2016 8th international conference on intelligent human-machine systems and cybernetics (IHMSC)* (pp. 203–207). IEEE.

Huang, L., Chen, X., Ni, X., Liu, J., Cao, X., & Wang, C. (2021). Tracking the dynamics of co-word networks for emerging topic identification. *Technological Forecasting and Social Change, 170*, 120944.

Huang, L., Liu, F., & Zhang, Y. (2020). Overlapping community discovery for identifying key research themes. *IEEE transactions on engineering management*.

Isler, Y., & Kuntalp, M. (2010). Heart rate normalization in the analysis of heart rate variability in congestive heart failure. *In Proceedings of the Institution of Mechanical Engineers Part H Journal of Engineering in Medicine, 224*(3), 453.

Iwata, T., Yamada, T., Sakurai, Y., & Ueda, N. (2010). Online multiscale dynamic topic models. In *Proceedings of the 16th ACM Sigkdd international conference on knowledge discovery and data mining* (pp. 663–672).

Jeong, C., Jang, S., Park, E., & Choi, S. (2020). A context-aware citation recommendation model with BERT and graph convolutional networks. *Scientometrics, 124*(3), 1907–1922.

Jeong, D. H., & Min, S. (2014). Time gap analysis by the topic model-based temporal technique. *Journal of Informetrics, 8*(3), 776–790.

Kai, H., Qi, K., Yang, S., Shen, S., Cheng, X., Huayi, W., Zheng, J., McClure, S., & Tianxing, Y. (2018). Identifying the "Ghost City" of domain topics in a keyword semantic space combining citations. *Scientometrics, 114*(3), 1141–1157.

Katsurai, M., & Ono, S. (2019). TrendNets: Mapping research trends from dynamic co-word networks via sparse representation. *Scientometrics, 121*, 1583–1598.

Keogh, E., Chu, S., Hart, D., & Pazzani, M. (2001). An online algorithm for segmenting time series. In *Proceedings 2001 IEEE international conference on data mining* (pp. 289–296).

Keogh, E., Chu, S., Hart, D., & Pazzani, M. (2004). Segmenting time series: A survey and novel approach. *Data Min Time Ser Databases, 57*, 1–22.

Kimura, A., Kashino, K., Kurozumi, T., & Murase, H. (2008). A quick search method for audio signals based on a piecewise linear representation of feature trajectories. *IEEE Transactions on Audio, Speech and Language Processing, 16*(2), 396–407.

Kiss, A., Temesi, G., Tompa, O., Lakner, Z., & Soós, S. (2021). Structure and trends of international sport nutrition research between 2000 and 2018: Bibliometric mapping of sport nutrition science. *Journal of the International Society of Sports Nutrition, 18*(1), 12.

Klavans, R., & Boyack, K. W. (2017). Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? *Journal of the Association for Information Science and Technology, 68*(4), 984–998.

Kleminski, R., Kazienko, P., & Kajdanowicz, T. (2020). Analysis of direct citation, co-citation and bibliographic coupling in scientific topic identification. *Journal of Information Science*. https://doi.org/10.1177/0165551520962775

Kralj, J., Valmarska, A., Robnik-Šikonja, M., & Lavrač, N. (2015). Mining text enriched heterogeneous citation networks. In *Pacific-Asia conference on knowledge discovery and data mining* (pp. 672–683). Springer, Cham.

Kuhn, T. S. (1962). *The structure of scientifific revolutions*. University of Chicago Press.

Lancichinetti, A., & Fortunato, S. (2009). Community detection algorithms: A comparative analysis. *Physical review E, 80*(5), 056117.

Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems* (pp. 2177–2185).

Li, G.-C., Lai, R., D'Amour, A., Doolin, D. M., Sun, Y., Torvik, V. I., Yu, A. Z., & Fleming, L. (2014). Disambiguation and co-authorship networks of the US patent inventor database (1975–2010). *Research Policy, 43*(6), 941–955.

Liu, Z. (2005). Visualizing the intellectual structure in urban studies: A journal co-citation analysis (1992–2002). *Scientometrics, 62*(3), 385–402.

Luo, L., & Chen, X. (2013). Integrating piecewise linear representation and weighted support vector machine for stock trading signal prediction. *Applied Soft Computing Journal, 13*(2), 806–816.

Mathieu, R. G., & Gibson, J. E. (1993). A methodology for large-scale R&D planning based on cluster analysis. *IEEE Transactions on Engineering Management, 40*(3), 283–292.

McCain, K. W. (2008). Assessing an author's influence using time series historiographic mapping: The oeuvre of Conrad Hal Waddington (1905–1975). *Journal of the American Society for Information Science and Technology, 59*(4), 510–525.

Mei, Q. Z., & Zhai, C. X. (2005). Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *Proceedings of the 11th ACM Sigkdd international conference on knowledge discovery and data mining* (pp. 198–207).

Miao, Z., Du, J., Dong, F., Liu, Y., & Wang, X. (2020). Identifying technology evolution pathways using topic variation detection based on patent data: A case study of 3D printing. *Futures, 118*, 102530.

Mikolov, T., Sutskever, I., Chen, K., et al. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems, 26*, 3111–3119.

Moreno, A., & Terwiesch, C. (2014). Doing business with strangers: Reputation in online service marketplaces. *Information Systems Research, 25*(4), 865–886.

Newman, M. E. J. (2004). Fast algorithm for detecting community structure in networks. *Physical review E, 69*(6), 066133.

Newman, M. E. J. (2012). Communities, modules and large-scale structure in networks. *Nature Physics, 8*(8), 25–31.

Newman, M. E. J., & GIirvan, M. (2004). Finding and evaluating community structure in networks. *Physical review, 69*(2), 108–113.

Nguyen, T. H. D., Melcer, E., Canossa, A., Isbister, K., & Seif El-Nasr, M. (2018). Seagull: A bird's-eye view of the evolution of technical games research. *Entertainment Computing, 26*, 88–104.

No, H. J., An, Y., & Park, Y. (2015). A structured approach to explore knowledge flows through technology-based business methods by integrating patent citation analysis and text mining. *Technological Forecasting & Social Change, 97*, 181–192.

Onan, A. (2019). Two-Stage Topic Extraction Model for Bibliometric Data Analysis Based on Word Embeddings and Clustering. *IEEE Access, 7*, 145614–145633.

Onan, A., & Toolu, M. A. (2020). Weighted word embeddings and clustering-based identification of question topics in mooc discussion forum posts. *Computer Applications in Engineering Education., 29*, 675–689.

Palla, G., Barabási, A.-L., et al. (2007). Quantifying social group evolution. *Nature, 446*(7136), 664.

Park, I., & Yoon, B. (2018). Technological opportunity discovery for technological convergence based on the prediction of technology knowledge flow in a citation network. *Journal of Informetrics*, *12*(4), 1199–1222.

Pépin, L., Kuntz, P., Blanchard, J., Guillet, F., & Suignard, P. (2017). Visual analytics for exploring topic long-term evolution and detecting weak signals in company targeted tweets. *Computers & Industrial Engineering, 112*, 450–458.

Qi, L., Wang, Y., Chen, J., Liao, M., & Zhang, J. (2021). Culture under complex perspective: A classification for traditional Chinese cultural elements based on NLP and complex networks. *Complexity, 2021*, 1–15.

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.

Qiu, J., & Lin, Z. (2011). A framework for exploring organizational structure in dynamic social networks. *Decision Support Systems, 51*(4), 760–771.

Rabitz, F., Olteanu, A., Jurkevičienė, J., & Budžytė, A. (2021). A topic network analysis of the system turn in the environmental sciences. *Scientometrics, 126*(3), 2107–2140.

Rees, B. S., & Gallagher, K. B. (2012). Overlapping community detection using a community optimized graph swarm. *Social Network Analysis & Mining, 2*(4), 405–417.

Ren, H., Renoust, B., Melançon, G., Viaud, M.-L. & Satoh, S. (2018). Exploring temporal communities in mass media archives.

Schwartz, R., Reichart, R., & Rappoport, A. (2015). Symmetric pattern based word embeddings for improved word similarity prediction. In *Proceedings of the nineteenth conference on computational natural language learning.*

Sharef, N. M., Martin, T., & Azmimurad, M. A. (2013). Conceptually related lexicon clustering based on word context association mining. *International Journal of Information Processing & Management, 4*(3), 40–50.

Sharma, D., Kumar, B., Chand, S., & Shah, R. R. (2021). Uncovering research trends and topics of communities in machine learning. *Multimedia Tools and Applications, 80*(6), 9281–9314.

Sheng, Z., Hailong, C., Chuan, J., & Shaojun, Z. (2015). An adaptive time window method for human activity recognition. In *2015 IEEE 28th Canadian conference on electrical and computer engineering (CCECE)* (pp. 1188–1192). IEEE.

Silvestrini, P., Amato, U., Vettoliere, A., Silvestrini, S., & Ruggiero, B. (2017). Rate equation leading to hype-type evolution curves: A mathematical approach in view of analysing technology development. *Technological Forecasting and Social Change*, *116*, 1–12.

Steven, A. G. (2011). Understanding belief using citation networks. *Journal of Evaluation in Clinical Practice, 17*(2), 389–393.

Su, L. X., Lyu, P. H., Yang, Z., & Ding, S. (2015). Scientometric cognitive and evaluation on smart city related construction and building journals data. *Scientometrics, 105*(1), 449–470.

Sud, P., & Thelwall, M. (2014). Evaluating altmetrics. *Scientometrics, 98*(2), 1131–1143.

Sun, J. M., Yu, P. S., Papadimitriou, S., & Faloutsos, C. (2007). GraphScope: Parameter-free mining of large Time-eevolving graphs. In *Proceedings of the 13th ACM Sigkdd international conference on Knowledge discovery and data mining* (pp. 687–696). New York: ACM.

Sun, X., & Ding, K. (2018). Identifying and tracking scientific and technological knowledge memes from citation networks of publications and patents. *Scientometrics, 116*(3), 1735–1748.

Symeon, P., Yiannis, K., Athena, V., & Ploutarchos, S. (2012). Community detection in social media, performance and application considerations. *Journal of Data Mining Knowledge Discovery, 24*(3), 515–554.

The, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association, 101*, 1566–1581.

Tseng, Y. H., Lin, C. J., & Lin, Y. I. (2007). Text mining techniques for patent analysis. *Information Processing & Management, 43*(5), 1216–1247.

Vaio, G. D., & Weisdorf, J. L. (2009). Ranking economic history journals: A citation-based impact-adjusted analysis. *Discussion Papers, 4*(1), 1–17.

Van Raan, A. F. (2004). Sleeping beauties in science. *Scientometrics, 59*(3), 467–472.

Verma, M. (2017). Cluster based ranking index for enhancing recruitment process using text mining and machine learning. *International Journal of Computer Applications, 157*(9), 23–30.

Wang, B., Liu, S., Ding, K., Liu, Z., & Xu, J. (2014a). Identifying technological topics and institution-topic distribution probability for patent competitive intelligence analysis: A case study in LTE technology. *Scientometrics, 101*(1), 685–704.

Wang, C., Blei, D., & Heckerman, D. (2008). Continuous time dynamic topic models. In *Proceedings of the international conference on uncertainty in artificial intelligence* (pp. 579–586).

Wang, Q., She, J., Song, T., Tong, Y., Chen, L., & Xu, K. (2016). Adjustable time-window-based event detection on twitter. In *international conference on web-age information management* (pp. 265–278). Springer, Cham.

Wang, X., & Mccallum, A. (2006). Topics over time: a non-Markov continuous-time model of topical trends. In *Acm Sigkdd International conference on knowledge discovery & data mining* (pp. 424–433). ACM.

Wang, X., Cheng, Q., & Lu, W. (2014b). Analyzing evolution of research topics with NEViewer: A new method based on dynamic co-word networks. *Scientometrics, 101*(2), 1253–1271.

Wang, Y., Liu, Z., & Sun, M. (2015). Incorporating linguistic knowledge for learning distributed word representations. *PloS one, 10*(4), e0118437.

Wasserman, S., & Faust, K. (1994). Social network analysis methods and applications. *Contemporary Sociology, 91*(435).

Wu, H., Yi, H., & Li, C. (2021). An integrated approach for detecting and quantifying the topic evolutions of patent technology: a case study on graphene field. *Scientometrics, 126*, 1–21.

Xie, J., Kelley, S., & Szymanski, B. K. (2013). Overlapping community detection in networks: The state-of-the-art and comparative study. *Acm Computing Surveys (csur), 45*(4), 1–35.

Xu, Y., Zhang, S., Zhang, W., Yang, S., & Shen, Y. (2019). Research front detection and topic evolution based on topological structure and the PageRank algorithm. *Symmetry, 11*(3), 310.

Xu, H., Winnink, J., Yue, Z., Liu, Z., & Yuan, G. (2020). Topic-linked innovation paths in science and technology. *Journal of Informetrics, 14*(2), 101014.

Yan, C., Yi, C., Wu, L., & Fang, J. (2015). Trend Feature Extraction in Condition Monitoring by a New Piecewise Linear Representation Method. In *First international conference on information sciences, machinery, materials and energy* (pp. 1378–1383). Atlantis Press.

Yang, B., Liu, D., & Liu, J. (2010). Discovering communities from social networks: methodologies and applications. In *Handbook of social network technologies and applications* (pp. 331–346). Springer.

Yang, Y., Wu, M., & Cui, L. (2012). Integration of three visualization methods based on co-word analysis. *Scientometrics, 90*(2), 659–673.

Yau, C. K., Porter, A., Newman, N., & Suominen, A. (2014). Clustering scientifc documents with topic modeling. *Scientometrics, 100*(3), 767–786.

You, H., Li, M., Hipel, K. W., et al. (2017). Development trend forecasting for coherent light generator technology based on patent citation network analysis. *Scientometrics, 111*(1), 297–315.

Zeng, Q., Hu, X., & Li, C. (2019). Extracting keywords with topic embedding and network structure analysis. *Data Analysis and Knowledge Discovery, 3*(7), 52–60.

Zhang, F., & Wu, S. (2021). Measuring academic entities' impact by content-based citation analysis in a heterogeneous academic network. *Scientometrics, 126*, 1–26.

Zhang, Y., Lu, J., Liu, F., Liu, Q., Porter, A., Chen, H., et al. (2018). Does deep learning help topic extraction? a kernel k-means clustering method with word embedding. *Journal of Informetrics, 12*(4), 1099–1117.

Zhang, Y., Porter, A. L., Hu, Z., Guo, Y., & Newman, N. C. (2014). "Term clumping" for technical intelligence: A case study on dye-sensitized solar cells. *Technological Forecasting and Social Change, 85*, 26–39.

Zhang, Y., Wu, M., Miao, W., Huang, L., & Lu, J. (2021). Bi-layer network analytics: A methodology for characterizing emerging general-purpose technologies. *Available at SSRN 3830937*.

Zhang, Y., Zhang, G., Zhu, D., & Lu, J. (2017). Scientific evolutionary pathways: Identifying and visualizing relationships for scientific topics. *Journal of the Association for Information Science & Technology, 68*(8), 1925–1939.

Zhou, D., Ji, X., Zha, H., & Giles, C. L. (2006). Topic evolution and social interactions: how authors effect research. In *Proceedings of the 15th ACM international conference on Information and knowledge management* (pp. 248–257).

Zhou, H. K., Yu, H., & Hu, R. (2017). Topic evolution based on the probabilistic topic model: A review. *Frontiers of Computer Science, 11*(5), 786–802.

Zhou, P., & Jiang, D. (2020). Study on the evolution of hot topics in the urban development. *Evolutionary Intelligence*. https://doi.org/10.1007/s12065-020-00391-y

Zhou, X., Huang, L., Porter, A., Vicentegomila, J. M., & Phillips, F. (2019). Tracing the system transformations and innovation pathways of an emerging technology: solid lipid nanoparticles. *Technological Forecasting and Social Change, 146*, 785–794.

Zhu, J., Li, X., Peng, M., Huang, J., Qian, T., Huang, J., Liu, J., Hong, R. & Liu, P. (2015). Coherent topic hierarchy: A strategy for topic evolutionary analysis on microblog feeds. In *International conference on web-age information management* (pp. 70–82). Springer, Cham.