Technological Forecasting & Social Change 170 (2021) 120944

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

journal homepage: www.elsevier.com/locate/techfore

Tracking the dynamics of co-word networks for emerging topic identification

Lu Huang [a], Xiang Chen [a,*], Xingxing Ni [a], Jiarun Liu [a], Xiaoli Cao [a], Changtian Wang [a]

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

ARTICLE INFO

Keywords:
Emerging topics
Dynamic networks
Link prediction
Machine learning

ABSTRACT

Identifying emerging topics has been an essential study for nations to develop strategic priorities, for enterprises to create business strategies, and for institutions to define research areas. However, how to characterize emerging topics effectively and comprehensively is still very challenging. This study proposes a framework for identifying emerging topics based on a dynamic co-word network analysis, which integrates a link prediction model with machine learning techniques. Time-sliced co-word networks are weighted according to the frequency of terms' co-occurrence. A back-propagation neural network is used to forecast a future network by predicting linkages among unconnected nodes based on existing links. Four indicators are then used to sort out potential candidates of emerging topics in the predicted network. A case study on information science demonstrates the reliability of the proposed methodology, followed by subsequent empirical and expert validations.

1. Introduction

The emerging topics, especially in science and technology (S&T), have been of interest to governments, companies, and individual scientists (Small et al., 2014). The earlier they identify an emerging topic, the sooner they can get ahead of the curve, gain a competitive advantage, and benefit from the innovation. Many studies have been conducted on this challenging topic. Breitzman, Thomas, and Chang strived to continuously improve the Emerging Clusters Model for real-time processing (Thomas and Breitzman, 2006; Chang and Breitzman, 2009; Breitzman and Thomas, 2015). Small et al. (2014) combined co-citation and direct citation clustering methods in large-scale data to detect emerging topics. However, the uncertainty, ambiguity, and complexity of emerging topics increase the difficulty of identification (Wang, 2018), which has been one of the most vexing issues in the field of technology management (Huang and Yuan, 2010).

Recently, network analytics has been introduced into bibliometric studies as an effective tool for technology management research (Yang et al., 2010). It relies on its advantages in revealing the hidden relationship among various elements in the S&T system, based on valuable information in the network structure (Choudhury and Uddin, 2016). The network-based methods for identifying emerging topics are classified into citation networks, co-word networks, and hybrid networks. Those methods play vital roles in both retrospective and contemporaneous studies (Lee, 2008; Cho and Shih, 2011; Boyack et al., 2014). However, the existing methods are still falling short of truly "characterizing the potential of what is detected to be emerging" (Rotolo et al., 2015).

From accumulated data, the development of S&T creates new knowledge (Lee et al., 2009), with the network topology concurrently emerges (Li et al., 2014). The networks become highly dynamic (Choudhury and Uddin, 2016). However, the previous studies were mostly focused on static networks without considering the dynamic evolutionary information of networks. Regardless of active exploration to predict the changes in topological features (Huang and Lin, 2009; Günes et al., 2016), the accuracy of the prediction model still has room to be improved. In addition, the development and quantification of indicators with strong connections to the definition and attributes of an emerging topic are still challenging research topics.

To address these concerns, we propose a methodology based on dynamic co-word network analysis to identify emerging topics. The proposed method consists of three main steps: 1) Time-sliced co-word networks are generated, integrating with link prediction techniques to predict the dynamic evolution of networks in the future; 2) Machine learning algorithms are applied to get mass topological information, improving the prediction accuracy. A trained Back-propagation Neural Network (BNN) model is used to fit three-link prediction indexes,

* Corresponding author.
E-mail addresses: huanglu628@163.com (L. Huang), bjchenxiang@hotmail.com (X. Chen), nixingxing11@163.com (X. Ni), bitliujiarun@126.com (J. Liu), cxl163990307@163.com (X. Cao), wangchangtian98@163.com (C. Wang).

https://doi.org/10.1016/j.techfore.2021.120944
Received 29 June 2020; Received in revised form 1 June 2021; Accepted 3 June 2021
Available online 16 June 2021

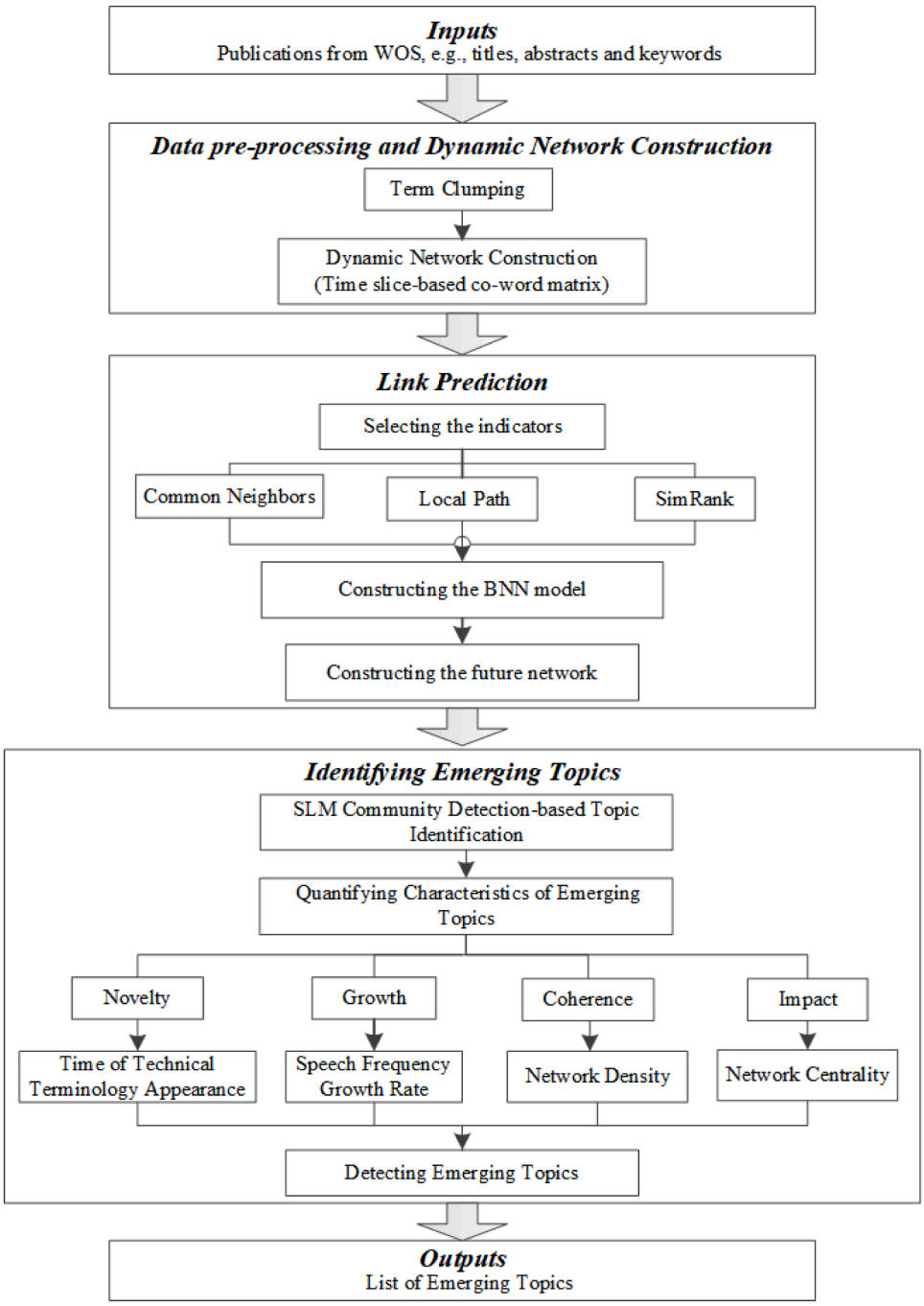


Fig. 1. The framework of the proposed emerging topic identification.

completely assessing local structure, path, and random walk information; 3) The network topology-based models are adopted to quantify the indicators in a predicted network.

The rest of this paper is organized as follows: Section 2 reviews the emerging topic identification and complex network analysis techniques. Section 3 describes the research design and methods in detail. Section 4 presents a case study on the information science field, identifying several emerging topics. Validation tests are also conducted to demonstrate the feasibility of the proposed method. Then, the insights gained from the case study and limitations are discussed with future studies in Section 5.

2. Related works

2.1. Emerging topic identification

According to Ohniwa and Hibino (2019), *emerging topic* is a relatively broad theme, which incorporates emerging technologies, methods, and scientific concepts, being composed of a cluster of terms (Deng et al., 2020). Although the emerging topic was considered important, its definition was not in consensus (Xu et al., 2020). The emerging topics are expressed in many different terms: emerging research topics, research fronts, emerging trends, emerging technologies, and emerging research fields (Xu et al., 2020). Rotolo et al. (2015) systematically reviewed various relevant definitions and described emerging technologies based on five attributes. Wang (2018) proposed a definition of an

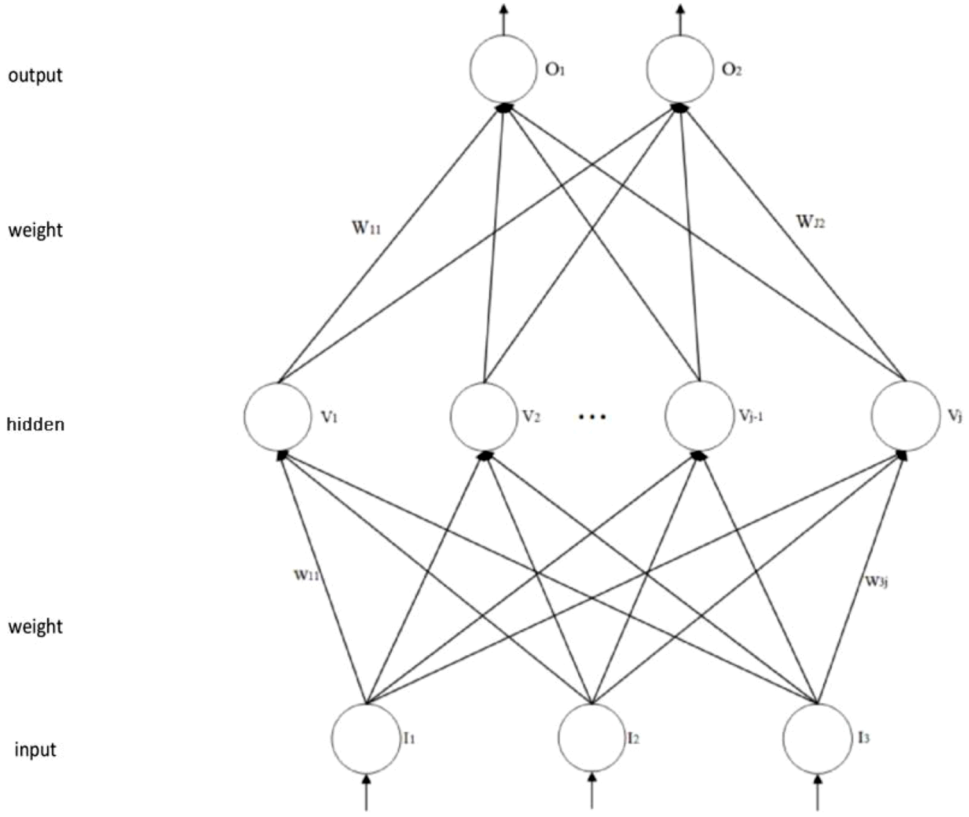


Fig. 2. The structure of the BNN (refined from Jiang et al. (2016)).

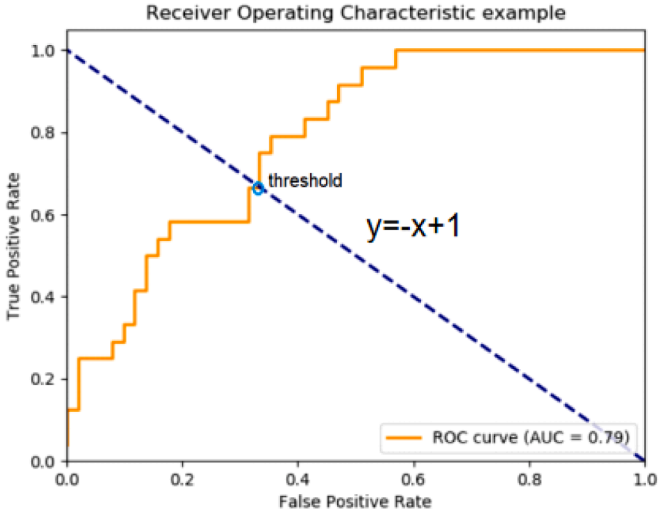


Fig. 3. Optimal threshold selection.

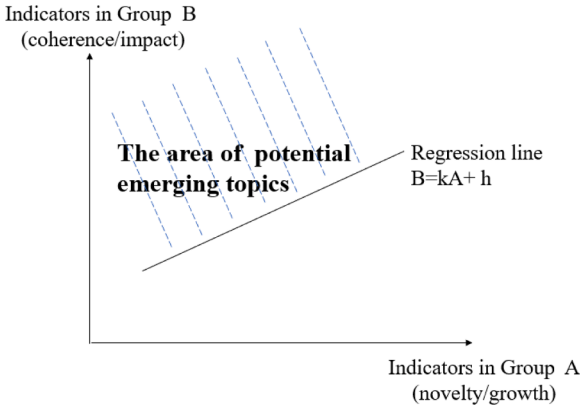


Fig. 4. Schematic diagram of emerging topic detection.

emerging topic and developed its attributes, including novelty, growth, coherence, and impact. We follow Wang (2018)'s definition for emerging topics in this study.

Various network analytics were proposed to identify emerging topics. In particular, the co-word network showed its advantages over the traditional bibliometric methods, as follows: 1) the results can be directly interpreted according to their semantics since it is a content-based (Zhao et al., 2018); 2) the relationships between terms can be mapped out (Papamitsiou and Mikalef, 2020); 3) a large corpus of information beyond term frequency can be aggregated (Gopsill et al., 2018); and 4) it supports the identification of key patterns and trends of topics (Hu et al., 2013).

2.2. Link prediction

Link prediction is a time-evolving network analysis model that can predict the likelihood of future associations among unconnected nodes based on existing links (Getoor and Diehl, 2005; Wang et al., 2017; Chen et al., 2021). The structural changes of networks in the future were predicted in this method. A cooperative network was explored to recommend potential collaborators (Guns and Rousseau, 2014; Yan and Guns, 2014; Huang et al., 2018). A network flow of technical knowledge was analyzed to discover technology opportunities (Park and Yoon, 2018). Emerging technologies were discovered based on a patent citation network (Érdi et al., 2013). The link prediction models can be categorized into three groups: similarity-based models, maximum likelihood models, and probabilistic models (Lü and Zhou, 2011; Choudhury and Uddin, 2016; Wang et al., 2017).

Table 1
Sources of journal data.

No	Journal Name	No. of Papers (2009–2016)	No. of Papers (2017–2018)
1	Scientometrics	2287	732
2	Journal of the Association for Information Science and Technology	1847	305
3	Information Research an International Electronic Journal	768	173
4	Journal of Informetrics	631	165
5	Information Processing & Management	525	145
6	Journal of Documentation	503	134
7	Journal of Information Science	442	103
8	Library & Information Science Research	358	62
9	Research Evaluation	301	59

Table 2
Stepwise results of term clumping process.

Step	Description	# Terms
1	Retrieve raw terms by NLP technique	152,468
2	Remove terms starting/ending with non-alphabetic characters, e.g., "step 1″ and "one way", "several indices"	135,157
3	Remove stop words, prepositions, and conjunctions	128,110
4	Remove meaningless terms and common terms in patents	119,641
5	Consolidate terms based on expert knowledge, e.g., "co-word analysis" and " word co-occurrence analysis"	115,601
6	Consolidate terms with the same stem, e.g., "citation count" and "citation counting"	102,502
7	Remove words of less than 3 letters	5838
8	Manually remove unrelated words	4640

Table 3
Details of the four time-slices.

Time slice	2009–2016	T1: 2009–2010	T2: 2011–2012	T3: 2013–2014	T4: 2015–2016
Nodes	4640	732	876	1107	1223
Edges	59,009	12,650	16,476	24,729	30,565

Similarity-based algorithms are a class of widely used methods of predicting the evolution of network structures (Lü and Zhou, 2011; Wang et al., 2017), and recently, deep learning techniques have been integrated to further improve prediction performance (Guns and Rousseau, 2014). In practice, the entire dataset is usually divided into two parts: 1) a training set for training and generating an initial link prediction model; 2) a test set for evaluating the performance based on metrics, such as, area under the receiver operating characteristic curve (AUC), Precision, Recall and F1 (Huang et al., 2018; Cai et al., 2020).

2.3. Community detection

There are six typical community detection methods (Newman and Girvan, 2004; Fortunato, 2010; Symeon et al., 2012): 1) Divisive algorithms, such as k-clique and Girvan-Newman algorithms (McCain, 2008); 2) Modularity-based methods: detecting and optimizing the quality of communities with graph-based measures (Ress and Gallagher, 2012); 3) Model-based methods: relying on dynamic algorithms to statistically infer the members and structure of communities (Qiu and Lin, 2011); 4) Local community detection methods: specifying a query node as a starting point and proceeding to search adjacent nodes to determine whether a node is part of a community (Branting and Mining, 2012); 5) Feature-assisted methods: leveraging additional features to explore the hidden relations between nodes that may form community structures (Wasserman and Faust, 1994); and 6) Spectral and clustering methods: including cohesive subgraph discovery and vertex clustering (Symeon et al., 2012). It has been proved in many studies that community detection methods have more advantages than traditional clustering methods (Chen, 2014; Ding et al., 2016): They are easier to find more technical details with a lower data volume demand.

The smart local moving (SML) was proposed to utilize a local move heuristic algorithm (Waltman and Eck, 2013). This modularity-based community detection has been proven to generate high-quality results even for very large networks. In our study, the SLM is adopted as the community detection algorithm.

3. Methodology

The proposed method consists of three parts: dynamic network construction, link prediction, and identifying emerging topics, as depicted in Fig. 1.

3.1. Data pre-processing and dynamic network construction

The dataset used in this paper comprises the titles, abstracts, and keywords of articles gathered from the Web of Science (WoS). The data was imported into VantagePoint[1] to retrieve key terms using a natural language processing function. The retrieved key terms are processed by a term clumping to remove noise and consolidate synonyms (Zhang et al., 2014), establishing a co-word network. The benefits of a co-word network are as follows: 1) the semantics relations among terms can be directly represented since it is a content-based; 2) it could achieve good integration performance with network analytics methods in terms of the valuable information hidden in the nodes and links of undirected networks; 3) it could also reflect technological components, such as materials, functions, manufacturing processes, applications, etc., describing the details of emerging topics with semantics.

The steps of generating a series of weighted co-word networks are as follows. First, the whole set of keywords is divided into multiple data sets by time slices, $D:\{D_t, t=1,2,\ldots,T\}$, where T denotes the sequential number of time slices, and D_t denotes the set of terms for the t^{th} time slice. Then, dynamic co-word weighted networks are generated based on the co-occurrence relationship of terms. These networks are denoted as $G:\{G_t, t=1,2,\ldots,T\}$, and the nodes in the network G_t are the terms in the t^{th} slice and the edges are weighted by the frequency of co-occurrence among terms.

3.2. Link prediction

The link prediction aims to predict the dynamic evolution of networks.

3.2.1. Selection of the indexes

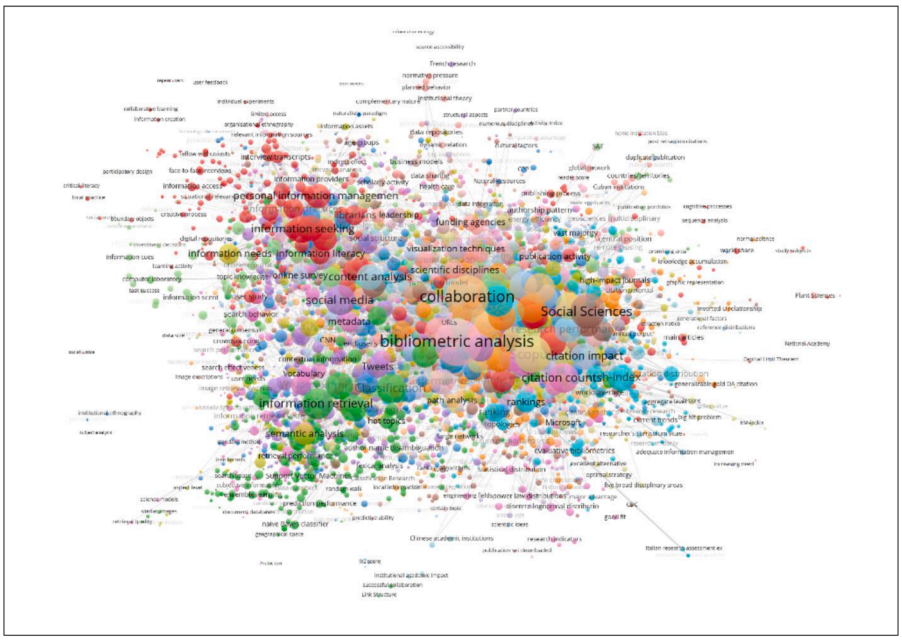
Following the study of Liben-Nowell and Kleinberg (2007), three link prediction indexes are adopted in our framework: common neighbors, local path, and SimR.

A *Common Neighbors (Newman, 2001).*

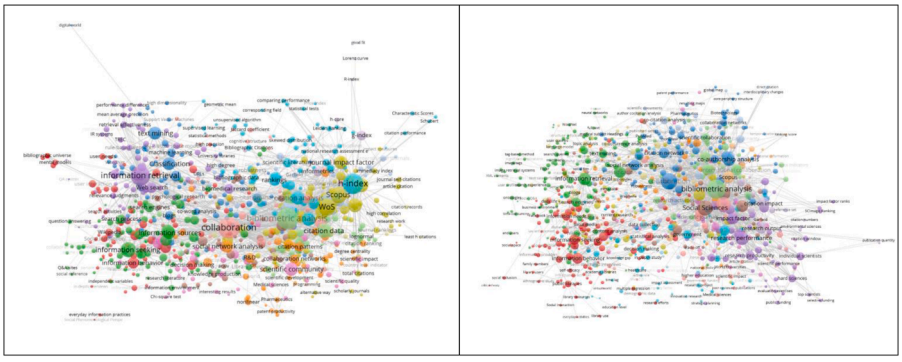
The Common Neighbors is the most straightforward index. The key idea is that the higher the number of common neighbors between two nodes, the more inclined those nodes are connected in the future. The number of common neighbors between node x and y, S_{xy}^{CN}, is defined as follows:

$$S_{xy}^{CN} = \sum_{Z\in\Gamma_{(x)}\cap\Gamma_{(y)}} w_{xz}^{\alpha_1} + w_{yz}^{\alpha_1} \quad (1)$$

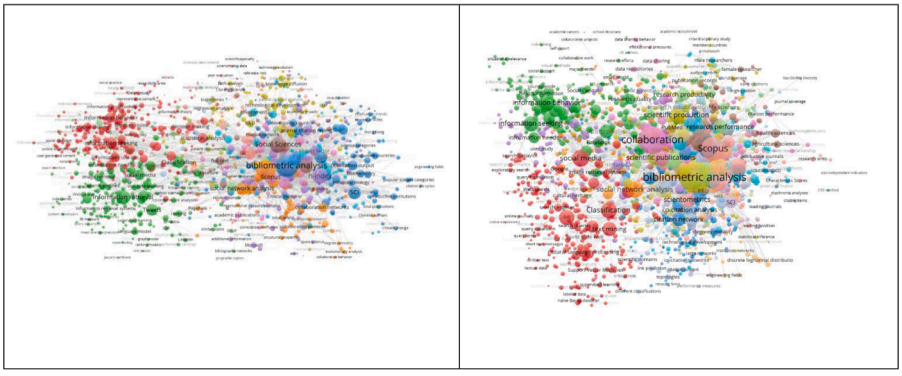
[1] https://www.thevantagepoint.com/



(a) the entire period (2009-2016)



(b) 2009-2010 (c) 2011-2012



(d) 2013-2014 (e) 2015-2016

Fig. 5. Co-word networks according to time slices. (a) the entire period (2009–2016), (b) 2009–2010, (c) 2011–2012, (d) 2013–2014, (e) 2015–2016.

where w_{xz} and w_{yz} represent the edge weights between nodes x and z, and node y and node z, respectively. $\Gamma_{(x)}$ and $\Gamma_{(y)}$ denote the neighbor sets of x and y. $\cap$ indicates the intersection of sets, and $\alpha_1(-1 \leq \alpha_1 \leq 1)$ is a parameter adjusting the influences between edge weights on prediction accuracy.

A *Local Path (Zhou et al., 2009).*

The local path index considers the contribution of third-order neighbors, i.e., the nodes linked to a neighboring node (Clauset et al., 2008). The value of the local path between node x and y, S_{xy}^{LP}, is defined as follows:

$$S_{xy}^{LP} = \left(A^2\right)_{xy} + \alpha_2\left(A^3\right)_{xy} \tag{2}$$

where $\alpha_2(0 \leq \alpha_2 \leq 1)$ is a free parameter, A is the adjacency matrix of the network, and $(A^n)_{xy}$ represents the number of paths with a length

Table 4
Details settings of the three hyperparameters.

Hyperparameter	Corresponding index	Description	Value range	Step size
α_1	Common Neighbors	The influence of edge weights.	[−1.0,1.0]	0.1
α_2	Local Path	The effect of the third-order path.	{0.1, 0.01, 0.001}	N/A
α_3	SimR	The decay factor.	[0,1]	0.1

equal to n.

A *SimR* (*Jeh and Widom, 2002*).

SimR measures the expectation that two nodes fall along one path of a random walk. The *SimR* between node x and y, S_{xy}^{SimR}, is defined as follows:

$$S_{xy}^{SimR} = \begin{cases} 1, & x = y \\ \alpha_3 \dfrac{\sum_{v\in\Gamma(x)}\sum_{v'\in\Gamma(y)} S_{vv'}^{SimR}}{n_x n_y}, & x \neq y \end{cases} \quad (3)$$

where n_x and n_y are the degrees of the nodes x and y as determined by the number of connected edges. v and v' denote a node in the set of all neighbors of x and y, respectively.

3.2.2. Constructing the bnn link prediction model

The structure of the BNN is depicted in Fig. 2. The three link prediction indexes are fed into the BNN as inputs. Then, the BNN predicts the existence of a link between two nodes in the future, where the prediction is formulated as a binary classification problem. The output of BNN is, accordingly, a yes/no result. The BNN is designed as three layers of the network, given that the inputs are three values. The input layer contains three neurons corresponding to each of the index inputs, while the output layer contains two neurons corresponding to two possible predictions. The number of neurons in the hidden layer is governed by the training process. The loss function is formulated as follows:

$$Loss = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i log f_{bp}(S_i) + (1 - y_i) log\left(1 - f_{bp}(S_i)\right)\right] \quad (4)$$

where y_i is the classification label, where 1 indicates a future edge between a pair of nodes, and 0 otherwise, and S_i represents the input values.

Then, the network $G:\{G_t, t = 1,2,\ldots,T\}$ is converted into three networks G_s^1, G_s^2, G_s^3 based on the sequence of time. The conversion process is formulated as follows:

$$G_s^1 = \sum_{t=1}^{t_1} G_t \quad (5)$$

$$G_s^2 = \sum_{t=t_1+1}^{t_2} G_t \quad (6)$$

$$G_s^3 = \sum_{t=t_2+1}^{t_3} G_t \quad (7)$$

where G_s^1 is a training set used to train the BNN model with a stochastic gradient descent (SGD) optimization (Jerez and Kristjanpoller, 2020). G_s^2 is a test set used for verifying the performance of the trained model. The final model is selected after several repeated training processes. G_s^3 is used for later prediction analysis, which is not included in the training process.

3.2.3. Construction of the future network

A future dynamic co-word weighted network $G':\{G_t, t = 1,2,\ldots,T'\}$ is forecasted based on the trained network $G:\{G_t, t = 1,2,\ldots,T\}$. Since G_s^3 is the latest network reflecting the latest development of technology, the well-trained model is applied in G_s^3 to calculate the probability of unconnected node pairs to be connected in the future. Then, the probability score is transformed into the weight of the predicted edge, and weighted edges would be added to G_s^3 to generate a dynamic weighted network G'.

The transformation process is described below:

1) Threshold selection: Following Kim and Magee (2017), the optimal threshold is set as the AUC value corresponding to the intersection point of the ROC curve and linear function (as shown in Fig. 3).
2) Weight calculation: In our study, we assume that a new edge will appear in the future when the connection probability of a node pair is higher than the threshold. The weight is computed as follows:

$$W_{xy} = \frac{S_{xy}}{Max(s)} * Avg(w) \quad \left(S_{xy} > threshold\right) \quad (8)$$

where S_{xy} is the probability of a future connection between a node pair (x_y). $Max(s)$ is the maximum value of all the probability scores, and $Avg(w)$ is the average weight of all edges in G_s^3. W_{xy} represents the weight of the new edge of a node pair (x_y).

Finally, the new weighted edges are added into G_s^3 to construct a predicted dynamic network $G':\{G_t, t = 1,2,\ldots,T'\}$, which is used to identify emerging topics. The predicted new edges are combined with the previous real network due to the following reasons.

(1) Topological structure is a vital factor for network-based analysis, and networks are dynamic. By adding new edges into the existing network, the characteristics of network topology evolution in the future can be better reflected. Accordingly, the development trend of S&T is better revealed. (2) The evaluation indicators: novelty, growth, and impact (Section 3.3.2) are time-sensitive (Tu and Seng, 2012; Small et al., 2014). Network topology-based models are adopted to quantify these indicators. Therefore, the more similar the network structure used for evaluation analysis is to the real one in the future, the more accurate the result of evaluation analysis is.

3.3. Identifying emerging topics

3.3.1. Smart local moving (SLM community detection-based topic identification

A single term cannot fully reflect a topic. Thus, we incorporate the SLM community detection algorithm (Waltman and Eck, 2013) to cluster terms into communities/topics.

3.3.2. Quantifying the characteristics of emerging topics

As described in Section 2.1, in order to distinguish emerging topics from a dynamic network G', we follow Wangˆ{\prime }s definition (2018): "A radically novel and relatively fast-growing research topic characterized by a certain degree of coherence, and a considerable scientific impact." The adopted four indicators, novelty, growth, coherence, and impact, are described as follows.

A *Novelty*

According to the study of Tu and Seng (2012), *novelty* reflects the freshness of a topic. The later a topic appears, the greater its novelty is. In our study, this indicator is modified as Eq. (9), improving the degree

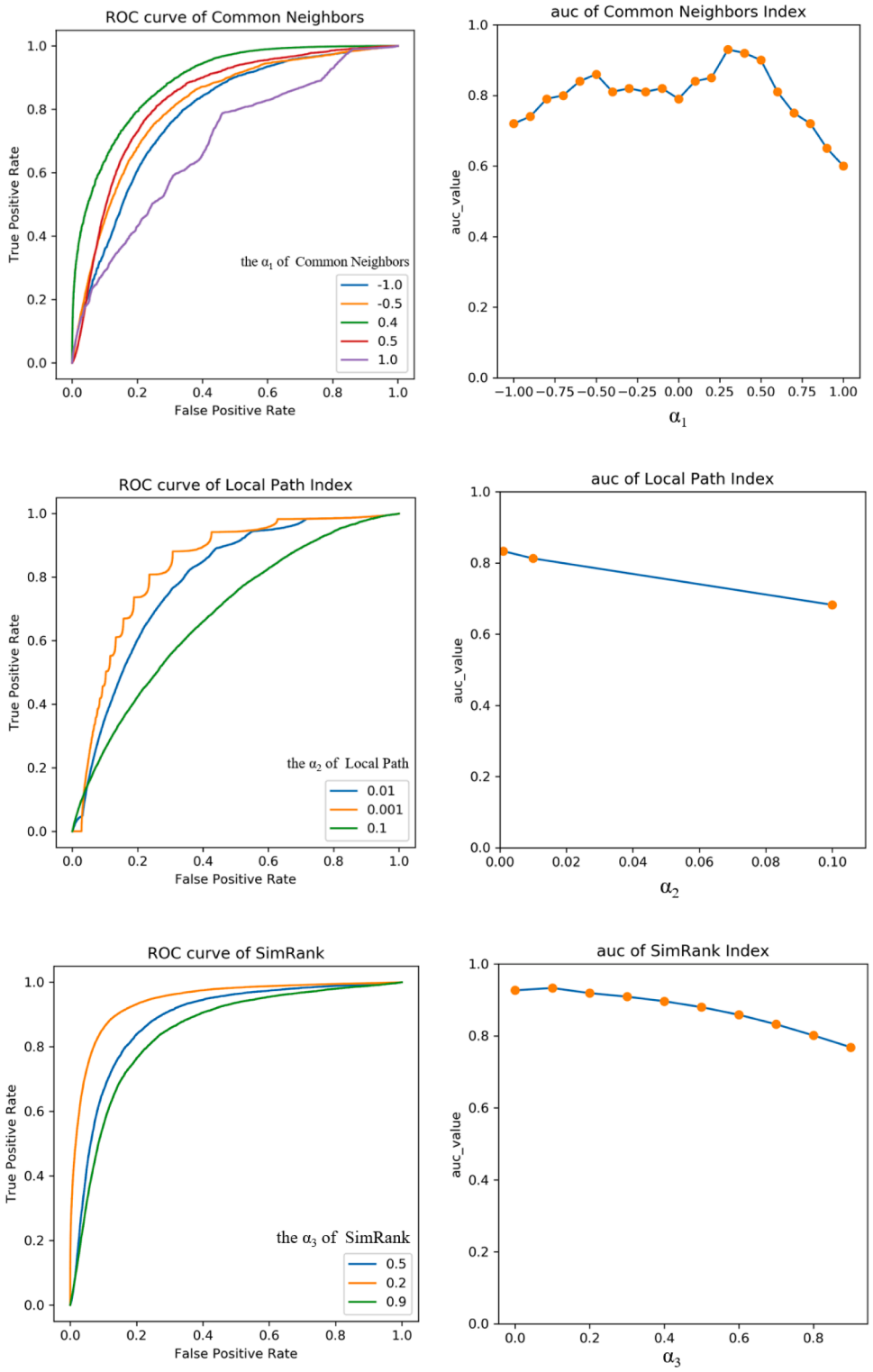


Fig. 6. The ROC and AUC values of three link prediction indexes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

of distinction:

$$Novelty = \frac{1}{n}\sum_{i=n} y_i - \frac{1}{N}\sum_{j=N} Y_j \tag{9}$$

where n is the total number of terms in the topics, and N is the total number of topics in the entire data set. y and Y represent the earliest years in which the term and the topic appeared, respectively. Accordingly, the novelty is computed by the difference between the average years in which the term and the topic appeared.

A *Growth*

The number of words belonging to emerging topics increases rapidly within a short period of time (Small et al., 2014). The indicator *growth*

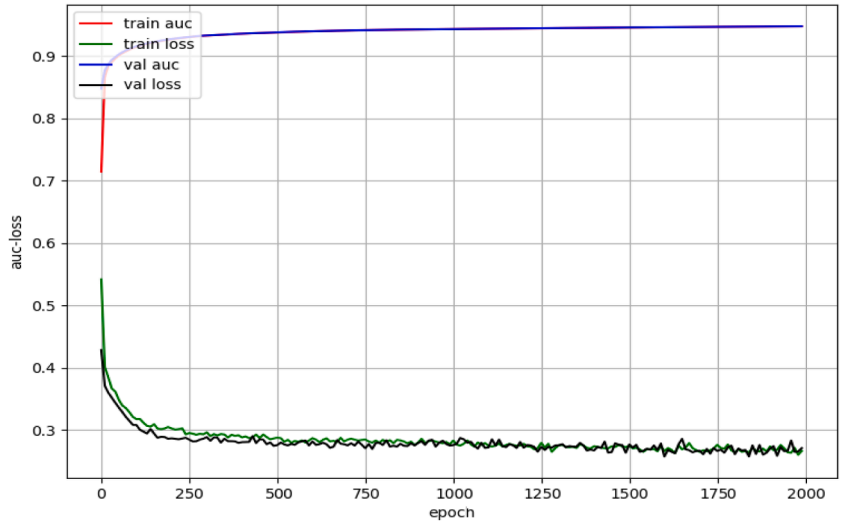


Fig. 7. BNN results of link prediction. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

reflects this phenomenon. However, random fluctuations, such as the expansion or reduction of the database, may cause a sudden increase or decrease of yearly publications in a certain research topic (Wang et al., 2018). To compensate for this fluctuation, we use smoothed word frequency instead of original word frequency to compute the *growth* of a topic, computed as follows:

$$\overline{Fre}_{i,t} = \frac{\left(Fre_{i,t-1} + Fre_{i,t}\right)}{2} \tag{10}$$

$$r_{i,t,t+1} = \frac{\overline{Fre}_{i,t,t+1}}{\overline{Fre}_{i,t}} \left(\overline{Fre}_{i,t} \neq 0\right) \tag{11}$$

$$\text{Growth} = \frac{1}{n}\sum_{i=n} r_{i,t,t+1} \tag{12}$$

where $Fre_{i,t}$ indicates the word frequency of the topic i at time t, and $r_{i,t,t+1}$ is the growth rate of the topic i during the period from t to $t+1$. In addition, Note that when the new topic with no relevant literature received $\overline{Fre}_{i,t} = 0$. In this case, since the growth rate is not available, we set $\overline{Fre}_{i,t}$ as the average word frequency of all topics at time t.

A *Coherence*

Unlike topics still in their infancy, emerging topics show coherence (Rotolo et al., 2015), which is measured as the ratio of the relative density of a topic to the other topics. As a topic matures, its connections gradually become denser. Once fully mature, the appearance of the topic in other fields become disappears. In network terms, this means that the nodes of the community corresponding to the topic have very few connections to other communities. Coherence is calculated as follows:

$$D_i = L_i / C_n^2 \tag{13}$$

$$D_j = L_j / (N-n) * n \tag{14}$$

$$Coherence = D_i / D_j \tag{15}$$

where D_i and D_j indicate the densities of community i and other communities in the network. L_i and L_j are the numbers of edges in community i and the other communities. N represents the total number of nodes in the network, and n is the number of nodes in community i. C_n^2 is the number of edges in the community in a fully connected state, $(N-n) * n$ represents the maximum number of possible edges between the nodes in the community i and the other communities in the network.

A *Impact*

The *impact* refers to the importance of a technical topic in the network (Small et al., 2014). The PageRank (Guo et al., 2011) is used to measure the impact based on a random walk. The access probability of particles is imitated via random walk in the network, where a higher probability indicates the higher impact of the node. The PageRank algorithm is formulated as follows:

$$V_t^{k+1'} = V_t^{k'} \widetilde{M} = V_t^{0'} \widetilde{M}^k \tag{16}$$

$$\widetilde{M} = s(1-\beta)M_t \tag{17}$$

$$PR_t^{'} = V_t^{k+1'} \tag{18}$$

where V_t^0 is the initial probability vector, and M_t is the state transition matrix, corresponding to the network adjacency matrix. Initially, $V_t^{0'} = \alpha PR_{t-1}^{'} + (1\text{-}\alpha)\ V_t^0$ where $\propto$ is the weight of the PageRank value from the previous period on the impact in the current period. β is a propensity parameter guiding the particle to jump to more significant nodes during the last moment, and s is a damping coefficient. The iterative process ends when $\left| V_t^{k+1'} - V_t^{k'} \right|$ reaches a pre-defined small threshold value.

The impact of a topic equals the average PageRank value of the nodes belonging to the topic, defined as follows:

$$Impact = \frac{1}{n}\sum_{i=n} PR_{T'} \tag{19}$$

where PR_i is the PageRank value of the topic node i at the time $T^{'}$.

3.3.3. Detecting emerging topics

This section aims to design a selection method based on the above four indicators, facilitating emerging topic detection from multiple possibilities. It is inappropriate to apply the mean weighting method for these four indicators due to their different characteristics. *Novelty* and *growth* are explicit indicators, well describing the development status of topics. In contrast, *coherence* and *impact* are based on the topology information of the predicted network, better depicting the great changes of topics in the development process. Thus, the four indicators are divided into two groups: Group A (*novelty* and *growth*) and Group B (*coherence* and *impact*).

In the topic evolution process, the Group A indicators constantly

Table 5
Indicator values of topics.

#Topic	Novelty Ranking	Value	Growth Ranking	Value	Coherence Ranking	Value	Impact Ranking	Value
#1	44	0.32	5	0.91	22	0.59	1	1
#2	4	0.8	49	0.62	25	0.58	2	0.96
#3	8	0.65	11	0.86	47	0.37	22	0.54
#4	22	0.56	22	0.83	34	0.51	24	0.53
#5	17	0.6	34	0.79	18	0.63	23	0.53
#6	47	0.22	1	1	43	0.43	5	0.88
#7	38	0.37	40	0.77	3	0.85	13	0.62
#8	15	0.61	33	0.79	23	0.59	6	0.8
#9	39	0.37	28	0.82	31	0.52	7	0.76
#10	9	0.65	32	0.8	32	0.52	8	0.75
#11	25	0.52	27	0.82	8	0.75	10	0.72
#12	20	0.58	19	0.85	17	0.63	11	0.68
#13	41	0.35	29	0.82	24	0.58	12	0.66
#14	30	0.46	20	0.84	19	0.61	21	0.54
#15	26	0.49	14	0.86	30	0.52	14	0.6
#16	5	0.76	25	0.82	14	0.69	15	0.58
#17	32	0.44	15	0.86	5	0.78	3	0.94
#18	11	0.63	37	0.78	12	0.7	4	0.93
#19	13	0.62	26	0.82	35	0.51	19	0.56
#20	24	0.53	13	0.86	13	0.69	17	0.56
#21	37	0.39	24	0.83	45	0.41	20	0.55
#22	19	0.59	6	0.9	15	0.65	9	0.73
#23	43	0.33	10	0.87	49	0.36	25	0.5
#24	1	1	47	0.63	33	0.52	26	0.49
#25	16	0.61	18	0.85	7	0.77	27	0.47
#26	6	0.69	45	0.72	11	0.72	29	0.44
#27	7	0.69	7	0.89	1	1	28	0.44
#28	48	0.16	2	0.99	39	0.48	30	0.43
#29	42	0.35	9	0.88	40	0.48	31	0.43
#30	27	0.49	35	0.79	20	0.6	32	0.42
#31	46	0.29	31	0.81	41	0.47	34	0.41
#32	35	0.4	44	0.74	37	0.49	33	0.41
#33	45	0.32	36	0.79	6	0.78	35	0.39
#34	33	0.42	42	0.76	42	0.46	37	0.38
#35	28	0.49	4	0.94	27	0.56	36	0.38
#36	34	0.42	8	0.89	2	0.99	38	0.37
#37	36	0.4	39	0.77	46	0.38	39	0.37
#38	14	0.62	46	0.72	38	0.49	40	0.36
#39	12	0.63	21	0.83	16	0.64	18	0.56
#40	18	0.6	12	0.86	26	0.57	41	0.35
#41	29	0.49	30	0.81	4	0.81	16	0.56
#42	3	0.81	43	0.74	36	0.49	42	0.34
#43	21	0.57	38	0.78	48	0.37	44	0.33
#44	31	0.45	23	0.83	9	0.74	43	0.33
#45	10	0.64	17	0.85	44	0.42	45	0.32
#46	23	0.56	41	0.76	28	0.56	46	0.29
#47	49	0.14	3	0.95	10	0.72	47	0.28
#48	40	0.37	16	0.86	29	0.55	48	0.27
#49	2	0.89	48	0.63	21	0.6	49	0.23

change and affect the topology of co-word networks, and then affect the Group B indicators. In order to comprehensively reflect the relative quantitative relationship between the two groups and their relative contributions on the process of topic evolution, we construct a linear regression model of Group B to Group A. The least-square method (Duan and Pan, 2017) is used to regress the linear function with coefficients k and h (as shown in Fig. 4). This line could be considered the benchmark for detecting emerging topics.

The regression function is used to expect Group B given the condition of Group A. In the topics above the regressed line, the performance of Group B is higher than the expected value (benchmark value), which are screened out as emerging topics. Specifically, we generate four linear regression models based on Group A and Group B, incorporating Novelty-Coherence, Novelty-Impact, Growth-Coherence, and Growth-Impact. We identity the topics located above the four regression lines as emerging topics.

3.4. The verification of the trained model

In order to validate the trained model, two-round validations are conducted. The first validation verifies our trained link prediction model. G_s^1 and G_s^2 are used as training and test sets, respectively. The performance of the link prediction model is evaluated in terms of five metrics: AUC, Accuracy, Precision, Recall, and F1. Furthermore, three traditional link prediction methods: Common Neighbors-based, Local Path-based, and SimR-based method are employed as compared methods.

The second validation verifies the accuracy of our prediction results (new edges). The new edges in G' are compared with the edges in the real network after G_s^3 in terms of *Precision*. Following Cai et al. (2020), the proportion of predicted new edges appearing in the future are focused on *Precision*, computed as follows:

$$Precision = \frac{L_r}{L} \tag{20}$$

where L represents the total number of predicted new edges in G', and L_r

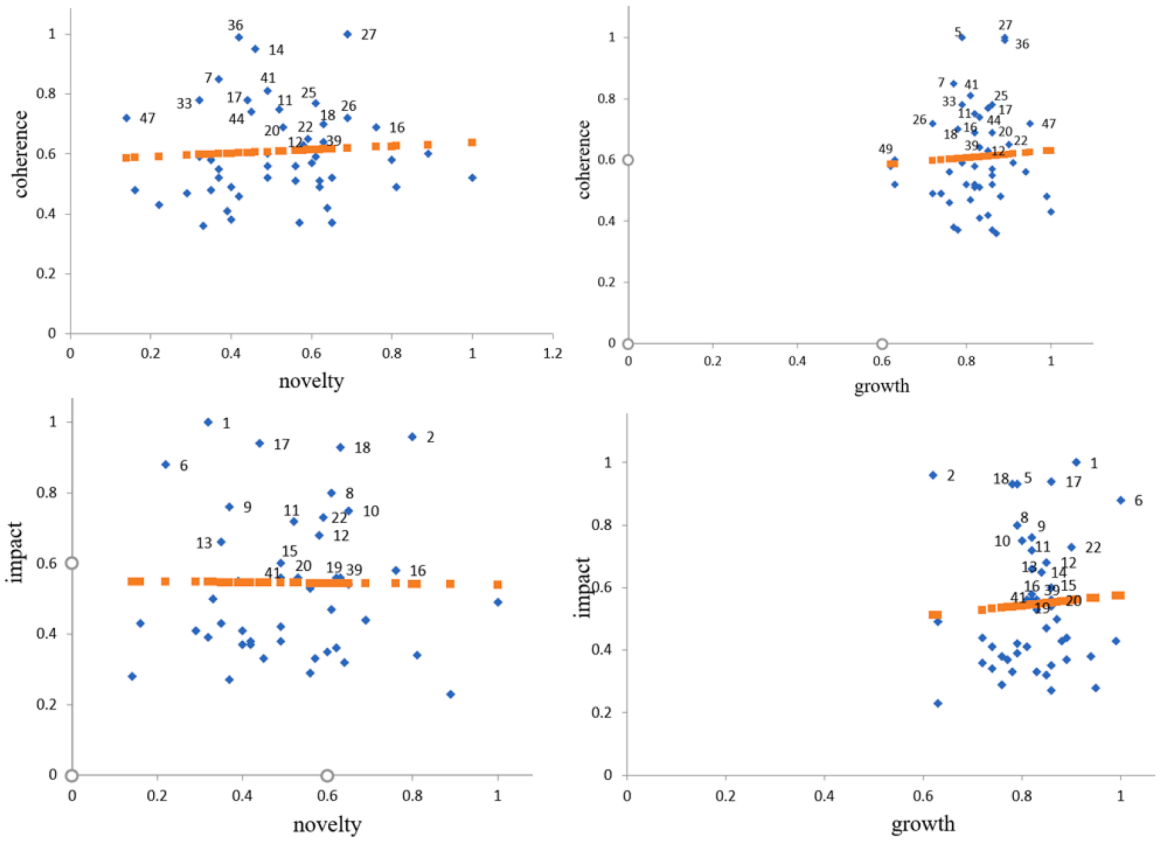


Fig. 8. Distribution maps for 49 candidate topics.

Table 6
Details of nine emerging topics.

No	#Topic	Descriptive Terms
1	#11	open-access; information professionals; Linked Data; Linked Data; LIS scholars
2	#12	scientometric indicators; leading position; citation impact; co-citation; citation behavior; citation analysis
3	#16	image retrieval systems; image use; search tactics; video search system; user preferences
4	#17	research hotpots; informetrics; climate change; health care;
5	#18	information seeking; information behavior; information sharing; information requirements; information sources
6	#20	network analysis; social network; information diffusion; tree structure; centrality measures
7	#22	social media; Altmetrics; blogs; Tweets; microblog
8	#39	semantic analysis; k-means; vector space model; F1-measure; latent semantic analysis
9	#41	interdisciplinarity; subject categories; impact factor; Social Sciences; scientific activity

Note: underlined terms were manually selected to label the corresponding topic.

represents the number of predicted new edges matched to the real network after G_s^3.

4. Case study

In this Section, the proposed framework is further analyzed in-depth on Information Science (IS). As a typical interdisciplinary discipline, IS has been spearheading cross-disciplinary research, connecting fundamental studies, i.e., mathematics, physics, and computer science, with real-world demands in the social sciences (Holland and George, 2008). In this study, information scientists conducted the validation.

Following Hou et al. (2018), we collected 9540 papers in nine core IS journals published between 2009 and 2018 from Web of Science (WoS) as our analysis corpus. Among them, the 7662 papers published between 2009 and 2018 were used as training data, while the other 1878 papers published between 2017 and 2018 were used for subsequent validation. Table 1 summarizes the journals and the number of corresponding papers.[2]

4.1. Data pre-processing and dynamic network construction

From titles, abstracts, and keywords of 7662 papers, 152,468 terms were retrieved using NLP techniques. After term clumping (Zhang et al., 2014), 4640 distinct terms remained, as outlined in Table 2.

The entire dataset was then divided into four time-slices (Table 3). Co-word networks were established for both the entire dataset and each time slice subset using VOSviewer (Eck and Waltman, 2010), as shown in Figs. 5a-5e.

The following observations were found from the varied relationships between terms in time-sliced networks over time. 1) Several major research topics in IS appear in all four time-slices, including information behavior, information seeking, information retrieval systems, bibliometric analysis, and scientometrics. 2) The total number of nodes and edges in networks has increased over time, indicating that research topics of IS continue to increase.

4.2. Dynamic network-based link prediction

Following the design in Section 3.2.2, G_s^1 and G_s^2 were considered as the training set (2009–2012) and the test set (2013–2014), respectively.

[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.

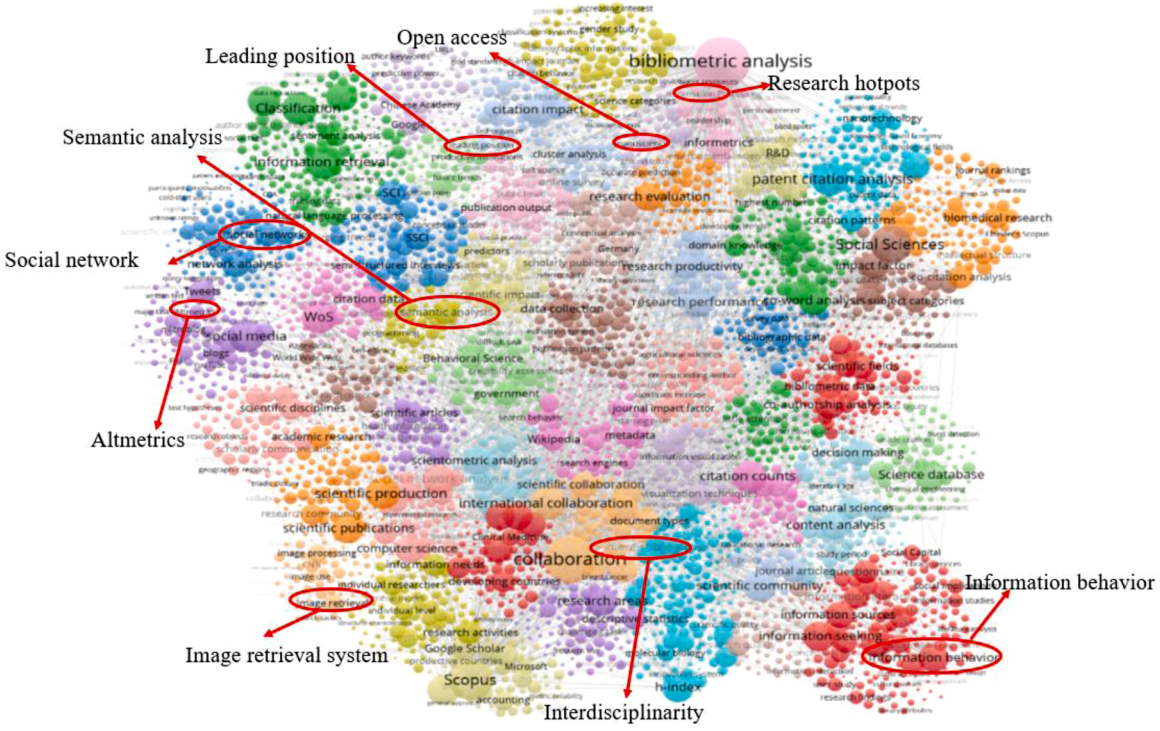


Fig. 9. Information science topics.

Table 7
The comparison of prediction performance.

Methods	Accuracy	AUC	Precision	Recall	F1
Our method	0.751	0.958	0.746	0.760	0.753
Common Neighbors-based	0.747	0.747	0.743	0.755	0.749
Local Path-based	0.747	0.801	0.741	0.758	0.750
SimR-based	0.616	0.646	0.616	0.443	0.516

Table 8
Results of expert evaluation.

No	Topics	Expert 1	Expert2	Expert3	Expert 4	Expert 5
1	open access	0.8	0.6	0.6	0.6	0.9
2	leading position	0.3	0.5	0.7	0.2	0.8
3	image retrieval systems	1	0.5	0.3	0.9	1
4	research hotpots	0.3	0.4	0.7	0.6	0.7
5	information behavior	0.5	0.6	0.5	0.5	1
6	social network	0.9	0.3	0.8	0.5	0.8
7	social media	0.9	0.6	1	0.7	1
8	semantic analysis	0.5	0.7	0.6	0.3	1
9	interdisciplinarity	0.8	0.7	0.7	0.2	0.9
	Max	1	0.7	1	0.9	1
	Min	0.3	0.3	0.3	0.2	0.7
	Avg	0.667	0.544	0.656	0.500	0.900
	Std	0.116	0.028	0.060	0.088	0.020
Mean of the group-average		0.653				

New edges were predicted based on G_s^3(2015–2016) after obtaining a well-performance link prediction model. The data between 2015 and 2016 was used instead of 2009–2016 since old data could introduce noise and negative impact on the prediction performance for streaming data analysis (Alzghoul et al., 2012; Zhang et al., 2017). Accordingly, the relatively old data (2009–2014) is used as training data, while the recent data (2015–2016) is used in the prediction analysis.

In order to achieve optimal prediction accuracy, the hyper-parameters of link prediction indexes are adjusted. The detailed parameter settings for three hyperparameters are given in Table 4, and the corresponding results are illustrated in Fig. 6. The figures on the left column represent the ROC curves for each prediction index according to the hyperparameter. The closer the curve is to the upper left indicates that link prediction indexes provide a high rate of true positives (TP). The figures on the right column are the AUC curves for each prediction index according to the hyperparameter.

Through a tuning process, the optimal values for α_1, α_2, and α_3 were set to 0.4, 0.001, and 0.1, respectively, providing the highest AUC (0.91, 0.82, and 0.92). Then, the normalized scores for three indexes were used to train the model.

Based on the results of previous steps, the three link prediction indexes of each pair of nodes (x _ y) are generated. The set of three values and classification label is used as input to the BNN model, which is described as follows:

$$\text{Sample}\left\{x\_y : \left(S_{xy}^{CN}, S_{xy}^{LP}, S_{xy}^{SimR}\right), Label : 1/0\right\}$$

where S_{xy}^{CN}, S_{xy}^{LP} and S_{xy}^{SimR} represent three link prediction indexes. The label=1 means the existence of a link between nodes x and y; otherwise, it is 0.

The training dataset consisted of 708,645 samples, of which 23,540 were positive samples. The test dataset contained 612,171 samples with 24,729 positive samples.

The SMOTE algorithm is applied to select 12,000 and 10,000 positive samples from training and test datasets, respectively, ensuring the sample ratio between training and test sets. Similarly, 35,540 and 34,729 negative examples were randomly selected to make a ratio of positive samples to negative samples of 1:1 in both datasets.

The BNN model is implemented using Keras[3] in Python. In each training case, 20% of the sample data was used to adjust the relevant

[3] https://keras.io/

Table 9
Relevant documentary proof of emerging topics.

No	Topics	Relevant documentary proof
1	#11 open access; information professionals; Linked Data; Linked Data; LIS scholars	Dong et al. (2018) has explored open access as a new research topic.
2	#12 scientometric indicators; leading position; citation impact; co-citation; citation behavior; citation analysis	Dong et al. (2018) and Hou et al. (2018) found that citation analysis was an emerging topic and important research front.
3	#16 image retrieval systems; image use; search tactics; video search system; user preferences	In the era of Web 2.0, retrieving image data quickly and accurately from vast databases has become a hot research topic (Piras and Giacinto, 2017). Netease Technology and The Drum reported that Google and Pinterest had announced a shift in their research focus from text-based search engines to voice and image search (SEO Sydney, 2018; The Drum, 2017). In 2018, eWeek reported that Microsoft had increased investments in deep learning to improve the retrieval accuracy and experiences of being an image search (Bayern, 2018).
4	#17 research hotpots; informetrics; climate change; health care	Feng et al. (2020) identified and analyzed several research hotpots, e.g., health care.
5	#18 information seeking; information behavior; information sharing; information requirements; information sources	Hou et al. (2018) have explored some new research topics, including information behavior, seeking behavior, information use, and interactive information retrieval.
6	#20 network analysis; social network; information diffusion; tree structure; centrality measures	Dong et al. (2018) identified network analytics as one of the most important future interdisciplinary topics.
7	#22 social media; Altmetrics; blogs; Tweets; microblog	Hou et al. (2018) concluded that Altmetrics has the most potential for frontier research of any pursuit in the field of IS. Dong et al. (2018) considered some high emergence degree topics, including social media and Altmetric.
8	#39 semantic analysis; k-means; vector space model; F1-measure; latent semantic analysis	Ruder et al. (2018) pointed out that semantic analysis has made amazing progress and become an important research topic. Semantic analysis was beginning to be a significant research topic of deep learning, as a report presented at the INDABA conference indicated (Herman, 2018).
9	#41 interdisciplinarity; subject categories; impact factor; Social Sciences; scientific activity	Chang (2018) proposed that the trend of interdisciplinarity is increasingly obvious.

hyperparameters, including the number of neurons in the hidden layer and the activation functions. The number of neurons in the hidden layer was set to 10, and the ReLu activation function was used for the input and hidden layers. Three eigenvalues were mapped non-linearly, and the output layer was selected as "Softmax". Fig. 7 depicts the training process with the finally adjusted hyperparameters, where the loss is continuously decreased. After 2000 epochs, the training converged, and the AUC value reached 0.965.

We applied the trained model in the network of 2015–2016 (G_s^3) and generated 3525 new edges in total. Then, these edges were added into the previous network (2009–2016) to construct a future network G' with 4640 terms, which was used for identifying emerging topics.

4.3. Emerging topic identification

SLM community detection algorithm was applied in network G' to cluster the 4640 terms into 49 topics, as shown in Fig. 9. Then, the four indicators: novelty, growth, coherence, and impact, were computed for each topic and standardized using min-max normalization (Isler and Kuntalp, 2010). Table 5 shows the indicator values for each topic.

Following the design in Section 3.3.3, 49 topics were distributed in four scatter plots (Fig. 8). The topics located above the regression line were selected for further process.

Finally, nine topics that repeatedly occur above the four regression lines were selected as emerging topics. The selected topics are listed in Table 6, which are highlighted in Fig. 9. For each topic, a set of descriptive terms with high frequency were selected. The terms with the highest frequency were considered as labels for each topic.

4.4. Experimental validation

4.4.1. Verification of the trained model and prediction results

The performance of the proposed link prediction model was evaluated in comparison with three traditional link prediction methods in terms of AUC, Accuracy, Precision, Recall, and F1. The compared methods include Common Neighbors-based, Local Path-based, and SimR-based methods. Table 7 summarizes the compared results, showing that our model outperforms the other methods in all five evaluation metrics. It proved that our link prediction model achieved good performance.

We further verified the prediction accuracy of our model (new edges) compared to the edges in the real network of 2017–2018. Concretely, we acquired papers published in IS journals in 2017–2018 and generated a network with 1121 nodes and 54,726 edges. Among 3525 new edges predicted by our model, 3322 edges were found in real 2017–2018. Accordingly, the *Precision* was 0.94. The results show that the proposed link prediction model can provide reliable results.

4.4.2. Validation of identified emerging topics

(1) *Expert knowledge-based validation*

In this Section, the emerging topic identification is qualitatively evaluated by leading domain experts. Specifically, we designed an expert evaluation, and five IS experts from the University of Technology Sydney, KU Leuven, Chinese Academy of Sciences, Xidian University, and Beijing Institute of Technology were engaged in the evaluation. Based on traditional questionnaire survey, experts were asked to score the emerging topics listed in Table 6. Each expert gave a continuous score between 0 and 1 to express his/her agreement on the emerging topic. A closer value to 1 indicated excellent agreement, while a close value to '0′ indicated strong disagreement. The results of five experts' questionnaires are summazired in Table 8.

Three experts gave the maximum score "1″, and the maximum score of the other two experts was 0.7 and 0.9, respectively, indicating that five experts were generally satisfied with the analysis results. Although there was still room for improvement, an average score of 0.653 (close to 0.7) is considered acceptable for most emerging topic identification methods (Zhang et al., 2018).

(1) *Empirical-based validation*

Since there are different understandings of 'emergence,' we further triangulated our results with the ones in other literature. The detailed information is described in Table 9. Note that our analysis is based on the data of 2009–2016, and the forecasted emerging topics were targeted since the year 2016. A list of empirical evidence was given in Table 9, indicating the alignment between our results and the literature. Therefore, it is reasonable to consider the emerging topics identified in

IS as reliable topics. In turn, the methodology of our study is also reliable.[4][5][6]

5. Discussion and conclusions

In this paper, we proposed a dynamic co-word network analysis to identify emerging topics. The link prediction method was introduced to reveal the dynamic changes of co-word networks. Meanwhile, a machine learning algorithm was applied to fit three link prediction indexes, fully assessing local structure, path, and random walk information, to improve the accuracy of the link prediction method. Emerging topics were measured and identified with four indicators: novelty, growth, coherence, and impact. The feasibility and reliability of the proposed method were validated on a case study in the IS application.

5.1. Potential applications

In addition to emerging topic identification, the proposed strategy can be modified to satisfy various potential applications. For example, the combination of link prediction and neural network analysis could be applied to forecast development trends of S&T. The emerging industrial and commercial topics can be analyzed based on multiple types of scientific and technical data, such as patents and business news.

5.2. Limitations

Still, our research has several limitations. 1) Link prediction can only predict appearing connections in the future, but disappearing links cannot be identified. 2) The information of emerging terms could be missing because of the data pre-processing that could remove some low frequency and twig terms. 3) The case study was only limited to journal papers from the WoS database, regardless that many emerging topics may appear in other databases, e.g., conference papers.

5.3. Future works

Future works will include improving the link prediction method. For instance, the current major direction in complex network research is graph embedding methods due to their superior efficiency and prediction accuracy. Furthermore, quantitative validation of the entire method will be considered, as well as low-frequency terms or twig terms. Future works will also include extending the case study from social science to natural science, from journal papers to entire academic papers, funding information, and social media for better universality of the model.

Acknowledgments

This work was supported by the National Nature Science Foundation of China Funds [Grant No. 71774013].

References

Alzghoul, A., Löfstrand, M., Backe, B., 2012. Data stream forecasting for system fault prediction. Comput. Indus. Eng. 62 (4), 972–978.
Boyack, K.W., Klavans, R., Small, H., et al., 2014. Characterizing the emergence of two nanotechnology topics using a contemporaneous global micro-model of science. J. Eng. Technol. Manage. 32 (32), 147–159.
Branting, L.K., 2012. Context-sensitive detection of local community structure. Soc. Netw. Anal. Min. 2 (3), 279–289.
Breitzman, A., Thomas, P., 2015. The emerging clusters model: a tool for identifying emerging technologies across multiple patent systems. Res. Policy. 44 (1), 195–205.
Cai, F., Chen, J., Zhang, X., Mou, X., Zhu, R., 2020. Link prediction based on deep latent feature model by fusion of network hierarchy information. Tehnič. Vjesni. 27 (3), 912–922.
Chang, C.K.N, Breitzman, A., 2009. Using patents prospectively to identify emerging, high-impact technological clusters. Res. Eval. 18 (5), 357–364.
Chang, Y.W., 2018. Exploring the interdisciplinary characteristics of library and information science (lis) from the perspective of interdisciplinary lis authors. Libr. Inf. Sci. Res. 40 (2), 125–134.
Chen, C., 2014. CiteSpace II: detecting and visualizing emerging trends and transient patterns in scientific literature. J. Assoc. Info. Sci. Technology 57 (3), 359–377.
Chen, W., Qu, H., Chi, K., 2021. Partner selection in china interorganizational patent cooperation network based on link prediction approaches. Sustain. 13 (2), 1003.
Cho, T.S., Shih, H.Y., 2011. Patent citation network analysis of core and emerging technologies in taiwan: 1997–2008. Scientomet. 89 (3), 795.
Choudhury, N., Uddin, S., 2016. Time-aware link prediction to explore network effects on temporal knowledge evolution. Scientomet. 108 (2), 745–776.
Clauset, A., Moore, C., Newman, M.E.J, 2008. Hierarchical structure and the prediction of missing links in networks. Nat. 453 (7191), 98–101.
Deng, S., Xia, S., Hu, J., Li, H., Liu, Y., 2020. Exploring the topic structure and evolution of associations in information behavior research through co-word analysis. Journal Libr. Info. Sci., 096100062093812
Ding, Z., Zhang, X., Sun, D., Luo, B., 2016. Overlapping community detection based on network decomposition. Sci. Rep. 6 (1), 1–11.
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. Scientomet.: Int. J. Quant. Aspect. Sci. Sci. Polic. 115 (2), 849–868.
Duan, Q., Pan, X., 2017. Identification of emerging topics in science using social media. J. Chi. Socie. Scienti. Technic. Info. 12 (36), 1216–1223.
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. Scientomet. 95 (1), 225–242.
Feng, J., Mu, X., Wang, W., Xu, Y., 2020. A topic analysis method based on a three-dimensional strategic diagram. J. Info. Sci., 016555152093090
Fortunato, S., 2010. Community detection in graphs. Phys. Rep. 486 (3–5), 75–174.
Getoor, L., Diehl, C.P., 2005. Link mining. ACM. SIGKDD. Explorat. Newslett. 7 (2), 3–12.
Gopsill, J.A., Shakespeare, P., Snider, C.M., Newnes, L., & Hicks, B.J., 2018. Investigating the evolving knowledge structures in new technology development. In: IFIP International Conference On Product Lifecycle Management. pp. 523–533.
Günes, I., Gündüz-Öğüdücü, S., Cataltepe, Z., 2016. Link prediction using time series of neighborhood-based node similarity scores. Data Mini. Knowled. Discover. 30 (1), 147–180.
Guns, R., Rousseau, R., 2014. Recommending research collaborations using link prediction and random forest classifiers. Scientomet. 101 (2), 1461–1473.
Guo, H., Weingart, S., Börner, K., 2011. Mixed-indicators model for identifying emerging research areas. Scientomet. 89 (1), 421–435.
Herman, E., 2018. The deep learning indaba report. ACM SIGMultimedia Records 9 (3), 5.
Holland, G.A., 2008. Information science: an interdisciplinary effort? J. Doc. 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). Scientomet. 115, 869–892.
Hu, C., Hu, J., Deng, S., Liu, Y., 2013. A co-word analysis of library and information science in china. Scientomet. 97 (2), 369–382.
Huang, L., Yuan, Y., 2010. Evaluation on the industrialization potential of emerging technologies based on principal component and cluster analysis. 2010 12th International Conference On Computer Modelling and Simulation. IEEE, pp. 317–322.
Huang, L., Zhu, Y., Zhang, Y., Zhou, X., Jia, X., 2018. A link prediction-based method for identifying potential cooperation partners: a case study on four journals of informetrics. 2018 Portland International Conference On Management of Engineering and Technology (PICMET). IEEE, pp. 1–6.
Huang, Z., Lin, D.K., 2009. The time-series link prediction problem with applications in communication surveillance. INFORMS. J. Comput. 21 (2), 286–303.
Isler, Y., Kuntalp, M., 2010. Heart rate normalization in the analysis of heart rate variability in congestive heart failure. Proceed. Institut. Mech. Eng. Part H J. of Eng. Medi. 224 (3), 453.
Jeh, G., & Widom, J., 2002. SimRank: a measure of structural-context similarity. In: Proceedings of the Eighth ACM SIGKDD International Conference On Knowledge Discovery and Data Mining. pp. 538–543.
Jerez, T., Kristjanpoller, W., 2020. Effects of the validation set on stock returns forecasting. Expert. Syst. Appl. 150, 113271.
Jiang, L., Zhang, J., Xuan, P., Zou, Q., 2016. BP neural network could help improve pre-mirna identification in various species. Biomed. Res. Int 1–11.
Kim, J., & Magee, C.L., 2017. Dynamic patterns of knowledge flows across technological domains: empirical results and link prediction. SSRN. Electr. J.. Available at SSRN: 10.2139/ssrn.2990729.
Lee, S., Yoon, B., Park, Y., 2009. An approach to discovering new technology opportunities: keywordbased patent map approach. Technov. 29 (6), 481–497.
Lee, W.H., 2008. How to identify emerging research fields using scientometrics: an example in the field of Information Security. Scientomet. 76 (3), 503–525.
Li, X., Du, N., Li, H., Li, K., Gao, J., Zhang, A., 2014. A deep learning approach to link prediction in dynamic networks. Proceedings of the 2014 SIAM International

[4] https://seosydney.com/digital-marketing/google-lens-discovering-visual-search-engine/http://www.thedrum.com/news/2017/10/16/the-future-search-can-image-and-voice-search-cater-paid-resu

[5] https://www.lightstalking.com/microsoft-launches-artificial-intelligence-powered-image-search-for-google-rival-bing/

[6] https://www.kamperh.com/slides/ruder+kamper_indaba2018_talk.pdf

Conference On Data Mining. Society for Industrial and Applied Mathematics, pp. 289–297.
Liben-Nowell, D., Kleinberg, J., 2007. The link-prediction problem for social networks. J. Assoc. Info. Sci. Technol. 58 (7), 1019–1031.
Lü, L., Zhou, T., 2011. Link prediction in complex networks: a survey. Physica A: Statisti. Mechani. App. 390 (6), 1150–1170.
McCain, K.W., 2008. Assessing an author's influence using time series historiographic mapping: the oeuvre of conrad hal waddington (1905–1975). J. Assoc. Info. Sci. Technol. 59, 510–525.
Newman, M.E.J., Girvan, M., 2004. Finding and evaluating community structure in networks. Physic. Rev. E 69 (2), 026113.
Newman, M.E.J., 2001. Clustering and preferential attachment in growing networks. Work. Pap. 64 (2), 025102.
Ohniwa, R.L., Hibino, A., 2019. Generating process of emerging topics in the life sciences. Scientomet. 121 (3–4), 1549–1561.
Papamitsiou, Z., & Mikalef, P., 2020. Mapping the intellectual progress in e-business, e-services and e-society from 2001 to 2019. Responsible Design, Implementation and Use of Information and Communication Technology, 252–265.
Park, I., Yoon, B., 2018. Technological opportunity discovery for technological convergence based on the prediction of technology knowledge flow in a citation network. J. Informetr. 12 (4), 1199–1222.
Piras, L., Giacinto, G., 2017. Information fusion in content based image retrieval: a comprehensive overview. Info. Fusi. 37, 50–60.
Qiu, J., Lin, Z., 2011. A framework for exploring organizational structure in dynamic social networks. Decis. Suppo. Syst. 51 (4), 760–771.
Rees, B.S., Gallagher, K.B., 2012. Overlapping community detection using a community optimized graph swarm. Soc. Netw. Anal. Mini. 2 (4), 405–417.
Rotolo, D., Hicks, D., Martin, B.R., 2015. What is an emerging technology? Res. Polic. 44 (10), 1827–1843, 2015.
Small, H., Boyack, K.W., Klavans, R., 2014. Identifying emerging topics in science and technology. Res. Polic. 43 (8), 1450–1467.
Symeon, P., Yiannis, K., Athena, V., Ploutarchos, S., 2012. Community detection in social media, performance and application considerations. J. Data Mini. Knowled. Discover. 24 (3), 515–554.
Thomas, P., Breitzman, A., 2006. A method for identifying hot patents and linking them to government-funded scientific research. Res. Eval. 15 (2), 145–152.
Tu, Y.N., Seng, J.L., 2012. Indices of novelty for emerging topic detection. Inf. Process. Manage 48 (2), 303–325.
Waltman, L., Van Eck, N.J., 2013. A smart local moving algorithm for large-scale modularity-based community detection. Europ. Physic. J. B 86 (11), 471.
Wang, Q., 2018. A bibliometric model for identifying emerging research topics. J. Assoc. Info. Sci. Technol. 69 (2), 290–304.
Wang, W., Feng, Y., Jiao, P., Yu, W., 2017. Kernel framework based on non-negative matrix factorization for networks reconstruction and link prediction. Knowled. Bas. Syst. 137 (1), 104–114.
Wasserman, S., Faust, K., 1994. Social Network analysis: Methods and Applications. Cambridge University Press.
Xu, S., Hao, L., An, X., Pang, H., Li, T., 2020. Review on emerging research topics with key-route main path analysis. Scientomet. 122, 607–624.
Yan, E., Guns, R., 2014. Predicting and recommending collaborations: an author-, institution-, and country-level analysis. J. Informetr. 8 (2), 295–309.
Yang, C., Park, H., Heo, J., 2010. A network analysis of interdisciplinary research relationships: the korean government's r&d grant program. Scientomet. 83 (1), 77–92.
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. J. Informetr. 12 (4), 1099–1117.
Zhang, Y., Porter, A.L., Hu, Z., et al., 2014. "Term clumping" for technical intelligence: a case study on dye-sensitized solar cell. Technol. Forecast. Soc. Chang. 85, 26–39.
Zhang, Y., Zhang, G., Zhu, D., Lu, J., 2017. Scientific evolutionary pathways: identifying and visualizing relationships for scientific topics. J. Assoc. Info. Sci. Technol. 68 (8), 1925–1939.
Zhao, W., Mao, J., Lu, K., 2018. Ranking themes on co-word networks: exploring the relationships among different metrics. Inf. Process. Manage. 54 (2), 203–218.
Zhou, T., Lü, L., Zhang, Y.C., 2009. Predicting missing links via local information. Europ. Physic. Journal B 71 (4), 623–630.

Lu Huang, Ph.D, associate professor of School of Management and Economics, Beijing Institute of Technology. She takes charge of Chinese National Science Foundation (Award #71774013– "Complex network-based global R&D cluster identification and evaluation") and published more than 40 research articles which were indexed by SCI/SSCI/EI. Her current research focuses on technology forecasting and assessment, particularly the study of emerging science and technology topics.

Xiang Chen, received the Ph.D. degree in management science and engineering from Beihang University in 2003. He is currently a Professor with the School of Management and Economics, Beijing Institute of Technology (BIT), China. His-research interests include intelligent recommender system, data mining, and artificial intelligence algorithm.

Xingxing Ni, working toward the master's degree in management science and engineering with the School of management and economics, Beijing Institute of Technology, China. Her research interests include technology innovation management, complex network and bibliometric analysis.

Jiarun Liu, is an undergraduate student in management science and engineering with the School of management and economics, Beijing Institute of Technology, China. His-research interests include technology innovation management, complex network and bibliometric analysis.

Xiaoli Cao, working toward the master's degree in management science and engineering with the School of management and economics, Beijing Institute of Technology, China. Her research interests include technology innovation management, complex network and bibliometric analysis.

Changtian Wang, working toward the master's degree in management science and engineering with the School of management and economics, Beijing Institute of Technology, China. His-research interests include technology innovation management, complex network and bibliometric analysis.