Information Processing and Management 59 (2022) 102843

Contents lists available at ScienceDirect

Information Processing and Management

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

A novel emerging topic detection method: A knowledge ecology perspective

Jinqing Yang [a,b], Wei Lu [a,b], Jiming Hu [a,b,*], Shengzhi Huang [a,b]

[a] School of Information Management, Wuhan University, Wuhan 430072, PR China
[b] Information Retrieval and Knowledge Mining Laboratory, Wuhan University, Wuhan 430072, PR China

ARTICLE INFO

Keywords:
Emerging topic detection
Ecological niche
Knowledge ecosystem
Differentiated novelty
Growth index

ABSTRACT

Emerging topic detection has attracted considerable attention in recent times. While various detection approaches have been proposed in this field, designing a method for accurately detecting emerging topics remains challenging. This paper introduces the perspective of knowledge ecology to the detection of emerging topics and utilizes author-keywords to represent research topics. More precisely, we first improve the novelty metric and recalculate emergence capabilities based on the "ecostate" and "ecorole" attributes of ecological niches. Then, we take the perspective that keywords are analogous to living bodies and map them to the knowledge ecosystem to construct an **e**merging **t**opics **d**etection method based on **e**cological **n**iches (ETDEN). Finally, we conduct in-depth comparative experiments to verify the effectiveness and feasibility of ETDEN using data extracted from scientific literature in the ACM Digital Library database. The results demonstrate that the improved novelty indicator helps to differentiate the novelty values of keywords in the same interval. More importantly, ETDEN performs significantly better performance on three terms: the emergence time point and the growth rate of pre-and post-emergence.

1. Introduction

The knowledge environment is constantly evolving and developing; new knowledge constantly emerges while old knowledge becomes obsolete. The evolution of knowledge discriminates between old and new knowledge by selecting those knowledge units that fit the environmental constraints in which they live (Firestone, 2008). Like organisms, knowledge units experience a full life cycle, which generally includes creation, mobilization, diffusion, and integration. (Stary, 2014). Furthermore, this process has been also subtly summarized as a four-stage S-shaped curve: birth, growth, maturity, and senility (Braun et al., 2000; Van den Oord & Van Witteloostuijn, 2018). Therefore, the evolution of knowledge units in a knowledge ecosystem follows certain universal laws that provide strong theoretical support for detecting emerging research topics.

The study of emerging topic detection has tended to primarily focus on the design of bibliometric indicators that reveal the nature of emergence (Xu et al., 2019). More specifically, many bibliometric indicators for identifying emerging topics from different aspects have been proposed, such as **novelty** (Tu & Seng, 2012; Wang, 2018; Porter et al., 2019), **growth** (Ohniwa & Hibino, 2010; Coccia, 2012; Dang et al., 2016), **community** (Yu et al., 2016; Yoon et al., 2018; Yoo et al., 2019), and **uncertainty** (Yu et al., 2016). It has been observed that the novelty and growth indicators often appear in the scientific literature on emerging topic detection. However,

* Corresponding author at: School of Information Management, Wuhan University, Wuhan 430072, PR China.
E-mail address: hujiming@whu.edu.cn (J. Hu).

https://doi.org/10.1016/j.ipm.2021.102843
Received 19 April 2021; Received in revised form 9 November 2021; Accepted 6 December 2021
Available online 15 December 2021

the indicators proposed in each article above are different and arbitrary, which may result in the lack of well-established linkages between the concept of an emerging topic and these operationalization indicators (Xu et al., 2019). Thus, to accurately detect emerging research topics, an alternative idea should be used to re-examine extant studies. Scientific exploration was conceptualized as a search in a high-dimensional abstract landscape of problems. There exists a clear analogy with the spatial model of evolutionary biology (Börner & Scharnhorst, 2009). Several computational models and basic theories from evolutionary biology can be employed in the study of topic evolution. Researchers were inspired to explore the evolutionary laws of research topics from an ecological perspective (Van den Oord & Van Witteloostuijn, 2018). These studies are instructive in terms of exploring a novel method of emerging topic detection. Namely, the ecological niche theory can be used during the process of detecting emerging topics from a knowledge ecology perspective.

The primary objective of this study is to propose a novel method of emerging topics detection from the perspective of knowledge ecology. More specifically, we first introduce the ecological niche theory to extend the idea of emerging topic detection, which provides two dimensions for classifying bibliometric indicators based on the "ecostate" and "ecorole" attributes. Second, we improve the novelty indicator using a weighting modified method to differentiate the novelty values of research topics in the same interval; we also calculate the slope value of growth at the detection point (niche value) to represent the emergence capability in terms of the "ecorole" attribute of ecological niches. Third, keywords are analogized with living bodies and mapped to the knowledge ecosystem to construct a niche baseline for detecting emerging topics. Finally, we conduct in-depth comparative experiments to verify the effectiveness and feasibility of our approach using data extracted from scientific literature in the ACM Digital Library database.

The rest of this paper is organized as follows: Section 2 reviews ecological niche theory and knowledge ecology, previous studies on bibliometric indicators of emerging topics, and keyword-based emerging topic detection. Section 3 describes the material and methods. Section 4 primarily discusses the study's results. Finally, we conclude with an overview of our findings and implications in Section 5.

2. Theoretical frameworks

2.1. Ecological niche theory and knowledge ecology

The ecosystem has typically been thought of as an explicit analogy to academics since it involves actors and their dynamic activities. Currently, academic big data has revived the idea of an "ecology of science" due to the discovery of potential laws that have exhibited sufficient accuracy. Several computational models from evolutionary biology can be adopted in research regarding evolution. Ecological niche theory is an analytical framework for explaining and describing how individuals adapt to their habitat based on common environmental resource patterns over a period from a population ecology theory perspective (Dimmick et al., 1992). According to Elton (1927), an ecological niche is a relative temporal, spatial and functional position in an ecosystem; later, Hutchinson (1957) constructed the quantitative method of ecological niches from the perspective of multidimensional attributes. The "ecostate" and "ecorole" attributes (Zhu, 1997; Raven, 2007) can be considered two dimensions that calculate the values of ecological niches (niche values). Furthermore, the term "ecostate" refers to the state of the biological units, which is understood as their accumulated energy, biomass, individual quantity, resource possession, and science and technology development level. Additionally, "ecorole" refers to the actual dominance or influence of a biological unit over the environment, such as the rate of energy and matter consumption, rate of productivity, and ability to occupy new habitats.

An ecological niche can reveal the competition and coexistence relationship among individuals with the same or similar resource requirements (Baum & Singh, 1994). Thus, it was adopted by other disciplines such as economics (Smith et al., 2010; Seol et al., 2012; Palage et al., 2019) and management (Hannan et al., 2003; Peli, 2017; Rong et al., 2018). For example, Peng et al. (2020) argued that the actual long-term development of green technology innovation in manufacturing is reflected in the ecological "ecostate" level, while the short-term fluctuations in green technology innovation in manufacturing are at the ecological "ecorole" level and reflect the changes and trends in the industry. In addition, keywords can be analogous to living bodies; Bao et al. (2015) developed an "energy growth index" indicator to identify emerging keywords. Researchers had universally recognized that the evolution of research topics in the knowledge ecosystem is analogous to that of biological units in an ecosystem (Sice et al., 2018). To become emerging units, research topics need to not only adapt to their living environment but also achieve certain survival conditions. Therefore, this study was inspired by the above ideas. We re-examine emerging topic detection by taking the knowledge ecology perspective that a keyword is analogous to a living body.

2.2. Bibliometric analysis of emerging topics

The detection of emerging topics should be theoretically grounded and practically useful. Specifically, it is essential to conceptualize a unifying mental framework that investigates the basic mechanisms of emerging research topics such as growth and change (Börner & Scharnhorst, 2009). The intuitive theory of emerging research topics is crucially derived from the evolution of research topics and the higher-level perspective of science. Additionally, the evaluation of whether a research topic is novel or emerging significantly depends on already-existing knowledge. Therefore, the new knowledge map was developed to represent the structure of all science and is based on journal articles to determine the leading areas of science (Boyack et al., 2005). From a network topology perspective, several characteristics of emerging areas were identified based on topological transitions and using quantitative graph theoretical measures such as density and diameter. (Bettencourt et al., 2009). Fanelli and Glänzel (2013) found that the scientific hierarchy provided the best rational framework for understanding the diversity of disciplines and reflecting theoretical and

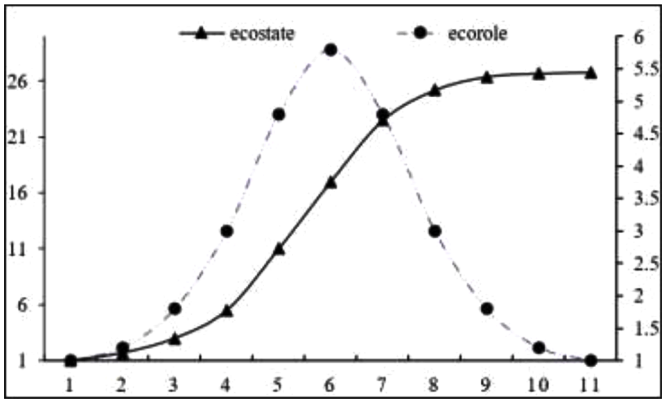

Fig. 1. The "Ecostate" and "ecorole" curves of ecological niches

methodological consensus.

Bibliometric analysis has a long tradition in scientific evaluation, and it involves bibliometric entities situated in the ecology of science such as authors (Adams, 2012; Weis & Jacobson, 2021), journals (Leydesdorff & Cozzens 1993; Weis & Jacobson, 2021), and research funding (Roshani et al., 2021). Coccia and Wang (2015) examined emerging directions in the advent of oncology technology trajectories based on breakthrough anti-cancer treatments using an approach based on a network of trends and crucial variables revealed by bibliometrics. Coccia (2018) conducted an inductive analysis based on emerging research fields to explain and characterize four features of evolving scientific areas in the context of scientific dynamics. Wang (2018, p.6) provided a new definition of emerging topics, describing it as "*a radically novel and relatively fast-growing research topic characterized by a certain degree of coherence, and a considerable scientific impact*"; this definition includes four attributes for characterizing emerging topics: (a) radical novelty, (b) relatively fast growth, (c) coherence, and (d) scientific impact (Xu et al., 2019). Additionally, an experiment showed that high rates of scientific growth are a significant signal based on the exponential model of growth (Coccia & Finardi, 2013). Specifically, higher growth rates of scientific production are observed in emerging research topics but not in established topics (Coccia, 2020), which agrees with the implication of the "ecorole" concept. It has been observed that research regarding emerging topic detection has tended to primarily focus on the design of bibliometric indicators.

2.3. Keyword-based emerging topic detection

In recent years, there has been an increasing amount of scientific literature regarding detecting emerging research topics. However, there is no consensus on the definition of research topics. In general, research topics consist of coherent research problems, concepts, methods, and techniques related to the researchers' discipline of interest (Braam et al.,1991). For example, the topic model was widely leveraged to generate research topics (Savoy, 2013; Vulić et al., 2015; Chen, 2017; Li et al., 2018; Wang et al., 2019; Wu et al., 2020). Keywords, as the minimum knowledge unit, were often used to represent research topics at a fine-grained level (Raamkumar et al., 2017; Yoon et al., 2018; Ohniwa et al., 2019). Specifically, keywords have been used as a medium for quantitatively and flexibly tracking the trajectory of research topic evolution (He, 1999). Likewise, Xu et al. (2018) observed the transition of state in the evolutionary process of interdisciplinary research by analyzing keyword evolution. Peset et al. (2020) provided a detailed analysis of the keyword evolution process by employing the survival analysis approach and found that measuring the appearance and disappearance of keywords would elucidate some relevant aspects of research topic evolution. Lu et al. (2021) investigated evolutionary patterns of author-keywords using paper metadata to predict research topic trends in computer science.

The emerging research topics appeared in the early stage of the entire life cycle and the adoption frequency of keywords was used to quantify the growth of research topics. However, it was unreasonable to utilize keyword frequency as the sole indicator when identifying emerging keywords (Dang et al., 2016). Emergence capabilities should be calculated to quantify the keyword emergence degree. Yu et al. (2016) found that the emergence degree of new keywords rises rapidly with the development of the research field. Keywords with a low degree of centrality and intermediary centrality could be considered emerging in the co-word network (Yoo et al., 2019). Certainly, novelty and fast growth were integrated when calculating emergence capability as a significant indication of emergence (Srinivasan, 2008; Tu & Seng, 2012; Small et al., 2014). The "energy growth index" indicator was also developed to characterize the evolution of keywords and identify emerging topics (Bao et al., 2015). Thus, keywords were widely adopted as proxies for research topics and were used to explore the evolution of research topics and identify emerging topics.

3. Material and methods

According to ecological niche theory, after emerging biological units have broken through a specific natural environment boundary and reached the minimum habitat threshold (Grubb, 1977), they have acquired life energy and can survive steadily in their ecosystem. Likewise, the knowledge units' emergence capabilities also reflect a stable state in the knowledge ecosystem in addition to unified

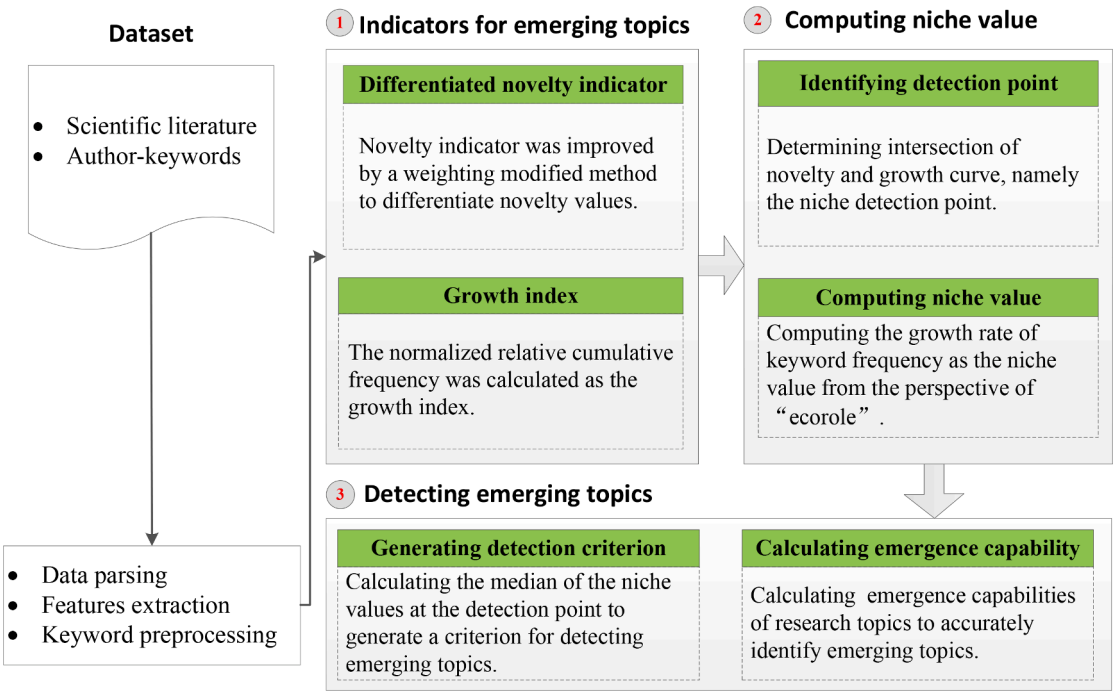

Fig. 2. Emerging topic detection method

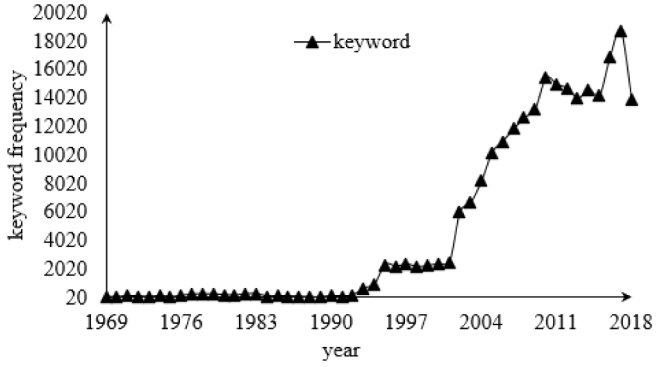

Fig. 3. Yearly distribution of keywords. The X-axis indicates the year of publication; the Y-axis is the keyword frequency.

concept connotations and competitive ability. In this regard, the crux of emergence detection is to determine detection points to measure the emergence capabilities (niche values). The "ecostate" and "ecorole" do not follow the same pattern; "ecostate" emphasizes the cumulative result of characteristics such as occupation and quantity, which generally follow an S-shaped curve. In contrast, "ecorole" primarily focuses on the ability to occupy a new habitat and characteristics such as gradient and growth rate, which exhibits a bell-shaped curve as shown in Fig.1. Therefore, the "ecorole" attribute of the ecological niche was adopted to measure the emergence capabilities of the knowledge units in the knowledge ecosystem.

The specific research method we used is as follows in Fig. 2: First, the relative cumulative frequency of keywords was calculated to express the growth index, and the relative growth rate of the cumulative keyword frequency was a weighting modified method of the novelty indicator. Second, the intersection of the novelty indicator and the growth index curve was identified as the detection point, and the ecological niche value was calculated by the slope of the growth value at the detection point. Finally, the median of the niche values was calculated to generate a criterion so that emerging topics could be accurately detected by calculating their emergence capabilities.

3.1. Sample and data

To ensure consensus with previous studies (Tu & Seng, 2012; Lu et al., 2021), the dataset was obtained from the ACM Digital

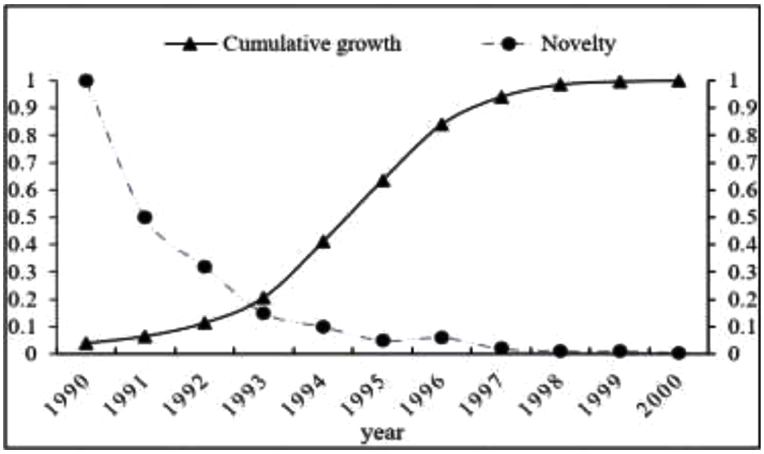

Fig. 4. Curves of novelty and growth. The X-axis indicates years elapsed from the first year in which the research topic frequency was non-zero in reverse order; the Y-axis represents the normalized number of cumulative growths.

Library database. It contains 231,384 keywords from 201,394 articles published between 1969 and 2018. The yearly distribution of keywords is shown in Fig. 3. To ensure dataset reliability, the first step is to automatically convert keywords into lowercase letters and to remove the spaces in front of and behind them. The second step is to solve some keywords' abbreviations and morphologies.

According to the research objective, detecting the emergence stage of keywords is the basic task of emerging topics detection, so the keywords' entire history records were utilized to present extant life cycle stages. Keeping in mind that 65.3% of the keywords eventually disappeared from the scientific literature (Peset et al., 2020), we constructed an experimental dataset according to the rule that keywords in the emergence stage should be selected as positive samples while keywords in the embryo stage should be selected as negative samples. Therefore, the top N keywords in descending order of frequency were selected as the positive samples, while the bottom N keywords that appeared more than 10 times were selected as the negative samples.

To verify the effectiveness and feasibility of the dataset construction, we chose the top 500 ranked keywords and employed a typical logistic function for the S-shaped curve fitting to ensure that keywords in the emergence stage were in the positive samples since the "S" shape can indicate how adoptions of new keywords grow over time (Mahajan et al., 1990; Choi et al., 2015). The formula is shown in Appendix A. We found that the S-shape curves for these high-frequency keywords fit well and their values are greater than 0.97. Therefore, this construction approach is sufficient. The "wireless sensor network" keyword is a good example and is shown in Fig. B-1 of Appendix B. In addition, to help our readers better understand the test dataset, we provide a visual presentation of several randomly selected samples in Fig. B-2 of Appendix B.

3.2. Measures of variables

3.2.1. Differentiated novelty indicator

As previously mentioned, the extant studies on emerging topic detection described the novelty bibliometric indicator according to the twofold length of time and number of publications. However, different research topics may have varying degrees of novelty in the same interval and it is redundant to utilize the number of publications since growth also implies an increase in publications over time. Therefore, this study improves the novelty indicator using a weighting modified method to differentiate novelty values for research topics in the same interval. The exponential decay function $e^{-\theta}$ was adopted as a weighting modified factor where θ represents the relative cumulative growth rate of the adoption frequency (Coccia & Finardi, 2012) and can be calculated by $\frac{f_t}{\sum_{t_0}^{t} f_t}$. The calculation for the novelty indicator is shown in equation (1):

$$Novelty = \frac{1}{t - t_0 + 1} \cdot e^{-\frac{f_t}{\sum_{t_0}^{t} f_t}} \tag{1}$$

where t_0 is the first year of non-zero frequency for the research topics in descending order, and f_t represents the frequency of research topic in year t.

3.2.2. Growth index indicator

In previous studies, researchers have generally adopted a growth index to measure research topic maturity. The growth index can be calculated through investments and achievements in scientific research such as the number of publications, the number and amount of funds, and researcher attention (Srinivasan, 2008; Cozzens et al., 2010; Small et al., 2014; Porter et al., 2019). Since a keyword appears once in a list of literature keywords, the number of publications is a proxy of keyword frequency. Considering that different research fields have different levels of research productivity, this study normalized the relative cumulative frequency for the research

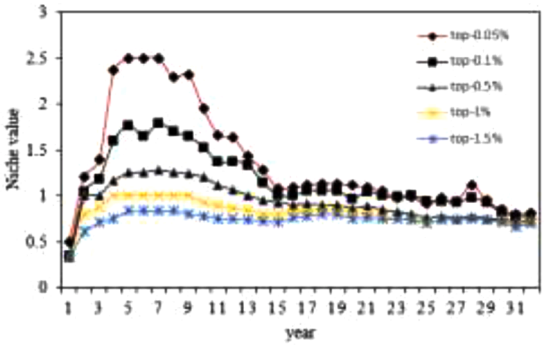

Fig. 5. ETDEN niche baseline

topics to compute the growth index as shown in equation (2):

$$Growth = norm\left(\sum_{t_0}^{t} f_t\right) \tag{2}$$

where $norm(\sum_{t_0}^{t} f_t)$ represents the normalized relative cumulative frequency of the research topics, t_0 is the first year of non-zero frequency for the research topics in descending order, and f_t is the frequency of the research topic in year t.

3.2.3. Emergence capability of research topics

To calculate the research topics' emergence capability, the first step is to identify the detection point (intersection of novelty and the growth curves as shown in Fig. 3), and the second step is to calculate the niche values based on the "ecorole" of the ecological niche.

(1) Identifying the detection point

According to niche theory, the niche value of the knowledge unit changes over time in the knowledge ecosystem. In this evolutionary process, the novelty of the keyword gradually decreases, and the cumulative growth gradually increases, as shown in Fig. 4. When placed in the same coordinate system, their normalized curves converge to a single point, which is the detection point for emerging topics. If the year at the intersection is a decimal, it is rounded to the nearest integer.

(2) Calculating the niche value

The growth rate of topic frequency was calculated to represent the emergence capability of a research topic at the detection point (Mazloumian, et al., 2011; Bao et al., 2015). To prevent a steep increase or decrease in a short period, the time interval used to calculate the niche value was set to three years, and the calculation formula is shown in equation (3).

$$niche\_value = \frac{f_{ny-1} - f_{ny+1}}{f_{ny-1}} \tag{3}$$

where ny indicates the year the niche was measured, f_{ny-1} is the research topics' frequency in the year preceding ny, and f_{ny+1} is the research topics' frequency in the year following ny.

3.3. Data analysis

3.3.1. The analysis procedure

To verify our novel ideas, the method from Tu et al. (2012) was carefully selected as the comparative exemplary (hereinafter called the benchmark). Three primary reasons were considered: First, only the novelty and growth bibliometric indicators were proposed for identifying emerging topics, which can reduce the method's complexity. Second, the chosen method also extracts keywords to represent research topics. Third, Tu's article has significant representativeness and impact; it was not only published in a high-quality journal but was also cited approximately 94 times in Google Scholar (up to 31 October 2021). The analysis procedure includes three steps. We first conducted a contrastive analysis of the niche baselines, which were calculated to generate a criterion for emerging topics. Then, we demonstrated the effectiveness of ETDEN based on our test dataset. Finally, we further verified the effectiveness and feasibility using three terms, namely the emergence time point and the pre-and post-emergence growth rate. We also provided case illustrations for interpretation.

3.3.2. Niche baseline analysis

The median of the niche values is considered the threshold value for identifying emerging topics. The threshold values for each year were connected to form the "niche baseline." In a given year, if the niche value for a research topic was higher than the niche baseline for the first time, this research topic was identified as an emerging topic in that year.

In Figs. 5 and 6, the x-axis indicates the number of years elapsed from the first year in which the research topic frequency was non-zero in reverse order. The y-axis represents the niche values or the emergence capabilities. In Fig. 6, the emergence capabilities of

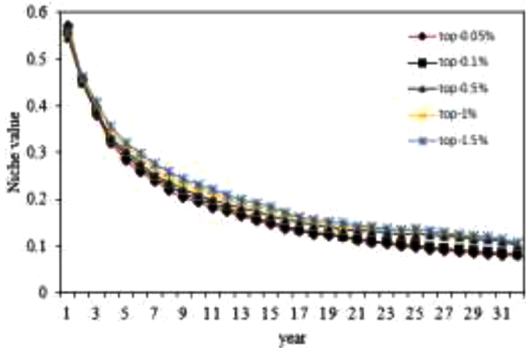

Fig. 6. Benchmark niche baseline

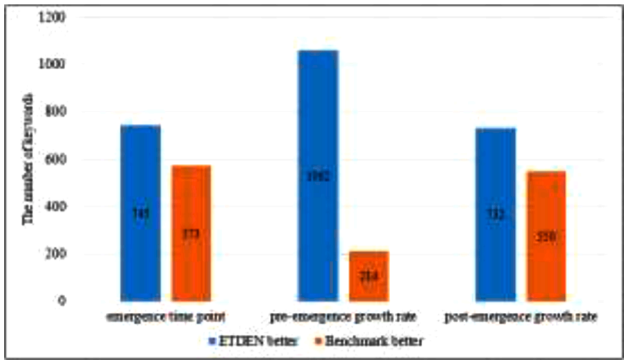

Fig. 7. Performance evaluation of ETDEN and the benchmark

Table 1
Performance of ETDEN and the benchmark

N-value	Number	Benchmark			ETDEN		
		Precision	Recall	F1	Precision	Recall	F1
0.05%	119*2	0.5022	0.9664	0.6609	0.7532	0.9747	0.8498
0.1%	238*2	0.4898	0.9580	0.6561	0.7541	0.9664	0.8471
0.5%	1190*2	0.4857	0.9244	0.6368	0.7006	0.9142	0.7933
1%	2381*2	0.4800	0.9034	0.6269	0.6784	0.8727	0.7634
1.5%	3571*2	0.4741	0.8804	0.6163	0.6517	0.8421	0.7348

Note: ETDEN is an emerging topic detection method based on ecological niches. F1 was calculated by 2 * Precision * Recall / (Precision + Recall).

research topics first rise rapidly and then gradually decline. The data exhibits a bell-shaped curve that agrees with the varying curve of the niche "ecorole". Fig. 7 shows that the research topics' emergence capabilities are the highest in the first year and then gradually decrease.

In contrast, the niche baseline of ETDEN is much more coincident with the dynamic evolution of knowledge populations in the knowledge ecosystem. In addition, comparing the top N values showed that the smaller the value of N was, the more prominent the peak of the bell-shaped curve was. This phenomenon is consistent with ecological niche theory—a smaller N indicates that the frequency of the selected keywords is higher. The emergence capabilities of keywords are stronger in the same period. However, for the niche baseline of the benchmark, different values of N are nearly indistinguishable in the early stage. Therefore, ETDEN is more in agreement with the evolutionary law for research topics in the knowledge ecosystem, making it more accurate when detecting emerging topics.

3.3.3. Results analysis

To verify the performance of the ETDEN method, the dataset was divided into five groups according to keyword frequency. In accordance with the experimental data construction strategy, the top N keywords were selected as positive samples and the bottom N were selected as negative samples; both sets of samples were in reverse order. As shown in Table 1, the values of N were set to 0.05%, 0.1%, 0.5%, 1% and 1.5%, and the corresponding number of positive and negative samples was 119*2, 238*2, 1190*2, 2381*2, and 3571*2, respectively. If the niche value for one research topic was higher than the niche baseline for the first time, the research topic

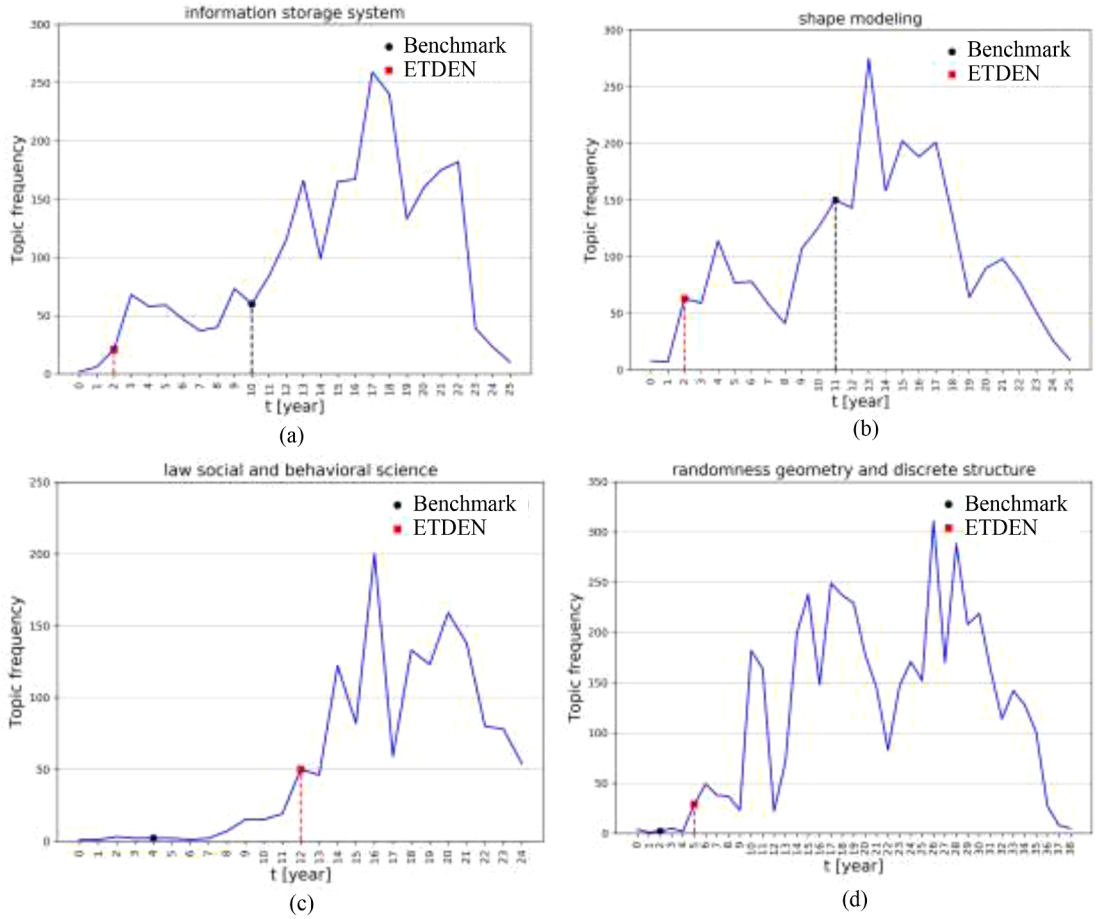

Fig. 8. Experimental results examples
Note: the red square is the emergence point identified by ETDEN and the black circle is the emergence point identified by the benchmark.

was identified as an emerging topic during that time. However, some positive samples had lower niche values than the niche baseline and the niche values of some negative samples were higher than those of the niche baseline in the extant history span. We used $P = TP/TP+FP$ and $R = TP/TP+FN$ to calculate the precision and recall, respectively. *TP* indicates the number of positive samples in the prediction results, *FP* suggests the number of negative samples that were identified as positive samples, and *FN* is the number of negative samples in the prediction results. The results are shown in Table 1.

As shown in Table 1, the F1 scores for ETDEN are higher than those of the benchmark in all five groups, which indicates that ETDEN better distinguishes emerging topics. To further demonstrate the effectiveness of ETDEN, we compared the time point of emergence and the reliability of the identified emerging topics, namely, whether ETDEN can more accurately measure the emergence capabilities of research topics and detect the time point of emergence earlier than the benchmark. An increase in research topic frequency is the most direct indication of emergence Choi et al., 2015). Therefore, the pre-and post-emergence growth rates were utilized to quantify the development potential and trend of emerging topics, as shown in equations (5) and ((6).

$$r_{be} = \frac{f_t - f_{t-1}}{f_{t-1}} \tag{5}$$

$$r_{ae} = \frac{f_{t+1} - f_t}{f_t} \tag{6}$$

where r_{be} represents the pre-emergence growth rate of the keyword frequency, r_{ae} is the post-emergence growth rate of the keyword frequency, andf_t is keyword frequency in year *t*.

We compared ETDEN with the benchmark by analyzing the effectiveness and feasibility using three terms: the emergence time point and the pre-and post-emergence growth rate. The 1,850 emerging keywords identified by the two methods were statistically analyzed, and the results are shown in Fig. 7.

As shown in Figure 7, in contrast with the benchmark, ETDEN identified more emerging topics earlier, and the identified emerging topics had a higher growth rate before emergence and a stronger growth trend. (1) For the emergence time points, 40.27% of the

emerging topics identified by ETDEN were identified earlier than those identified by the benchmark, while 30.89% of the emerging topics identified by the benchmark were identified earlier than by ETDEN. (2) Using ETDEN, 57.41% of the identified emerging topics had higher pre-emergence growth rates than those found by the benchmark, while 11.57% of the emerging topics had higher pre-emergence growth rates when using the benchmark as opposed to ETDEN. The two cases are shown in Figs. 8 (a) and (b). (3) Using ETDEN, 39.57% of the emerging topics had higher post-emergence growth rates than those found by the benchmark, while 29.72% of the emerging topics had higher post-emergence growth rates when using the benchmark as opposed to ETDEN, as shown in Figs. 8 (c) and (d). (4) Overall, 23.46% of the emerging topics identified by ETDEN exceeded the benchmark on all three of these aspects, while only 5.68% of the emerging topics identified by the benchmark are higher than those identified by ETDEN.

4. Results and Discussion

When reviewing the literature, an increasing number of bibliometric indicators have been proposed to detect emerging topics. However, there is no consensus among these bibliometric indicators, and there is a lack of specific dimensions for classifying them. Thus, it is essential for the basic theory or approach to establish solid linkages between the concept of emergence and the proposed operationalization indicators. In contrast with the past, academic big data revived the idea of an "ecology of science" by mining potential laws with sufficient accuracy. The ecological niche was introduced to detect emerging topics by viewing keywords as being analogous to a living body and mapping them to a knowledge ecosystem. More precisely, the "ecostate" and "ecorole" attributes of ecological niches roughly classify these bibliometric indicators into two groups. The primary contributions of our paper are that we re-examined emerging topic detection from the perspective of ecological niches and proposed ETDEN, which demonstrates better performance when detecting emerging topics according to three characteristics: the emergence time point and the growth rate of pre-and post-emergence.

In terms of basic mechanisms of growth and change in research topics, research topics have higher novelty, lower cumulative growth, lower influence, and a vague concept boundary in the early stage of the life cycle (Wang, 2018). With the investment of academic resources, the precise definitions of research topics gradually converge. Thus, researchers shed light on the characteristics of input and output to design quantitative indicators for emerging topics. Novelty and growth are significant indicators; the novelty trait involves newness, innovation, and an intrinsic ability to occupy new habitats, while growth reflects the rapid development of research topics such as publication volume and support through funding. Since emerging topic detection is performed in the early stage of a topic's life cycle, which is associated with higher novelty and lower cumulative growth, the novelty indicator is then more decisive. However, the benchmark's novelty metric was the inverse number of the time length, and different research topics may have undifferentiated novelty values in the same interval. We can observe that research topics with bigger and smaller cumulative frequencies have similar niche value curves in Fig. 7. Therefore, this study improves the novelty indicator using a weighting modified method to differentiate the novelty values of research topics in the same interval. The cumulative growth rate was calculated as a weighting from the "ecorole" perspective to resolve the problem of identical novelty values in the same interval.

Additionally, the extant indicators of emerging topic detection have no consensus and arbitrary selectivity, which may result in a lack of well-established linkages between the concept of an emerging topic and the operationalization indicators. It is our belief that a conceptualization of basic mechanisms can provide the intellectual framework to interlink and piece together the many indicators in existence, leading to a more comprehensive description and understanding of the nature and dynamics of emerging research topics. In a sense, academic big data revived the idea of an "ecology of science" through the discovery of potential laws that exhibited sufficient accuracy. Furthermore, according to the population ecology perspective, ecological niche theory is an analytical framework for explaining and describing how individuals adapt to their habitat based on common environmental resource patterns over a period. The "ecorole" and "ecostate" are two core properties of the ecological niche theory. In particular, the "ecorole" is more in agreement with the concept of emergence, which emphasizes intrinsic transformation; this study therefore calculated the slope value of growth at the detection point to represent the emergence capability of keywords from the "ecorole" attribute. Therefore, we improved the novelty metric and recalculated the emergence capabilities (the value of the detection point) based on the "ecostate" and "ecorole" attributes. We believe that these findings can extend the research regarding emerging research topic detection and benefit research foundations and policy-makers.

5. Conclusion

This paper proposed a novel emerging topic detection method based on ecological niches from the perspective of knowledge ecology. Specifically, the research method for emerging topic detection was extended to provide two dimensions for classifying bibliometric indicators based on the "ecostate" and "ecorole" twofold attributes of ecological niches. In addition, the novelty indicator was improved by a weighting modified method that differentiates the novelty values of research topics in the same interval; additionally, the slope value of growth at the detection point was calculated to represent the emergence capability. The experimental results demonstrated that the differentiated novelty indicator proposed in this study helps to detect the emergence time point of emerging topics as early as possible and differentiate the novelty values of keywords in the same interval. The niche baseline is more in agreement with the evolutionary law of research topics due to its calculation of niche values from the perspective of the "ecorole". ETDEN demonstrates better performance when detecting emerging topics in terms of three characteristics: the emergence time point and the growth rate of pre-and post-emergence.

This study has significant theoretical and practical implications. In theoretical terms, this study extends the idea of emerging topic detection by providing two dimensions for classifying bibliometric indicators based on the "ecostate" and "ecorole" attributes of

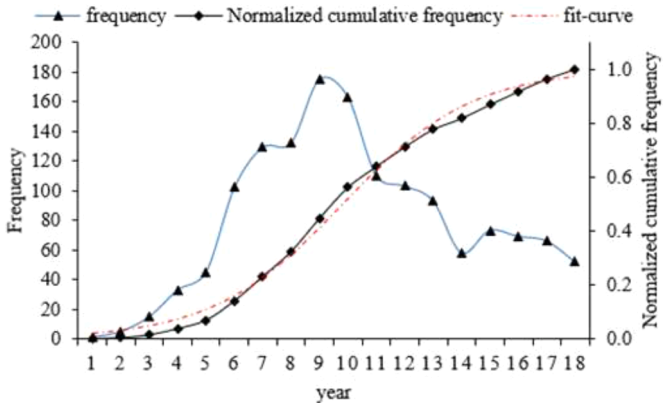

Fig. B1. S-shaped curve fitting of wireless sensor network. The X-axis indicates the number of years elapsed from the first year in which the research topic frequency was non-zero in reverse order; the right Y-axis represents the normalized number of cumulative frequency and the left is yearly frequency.

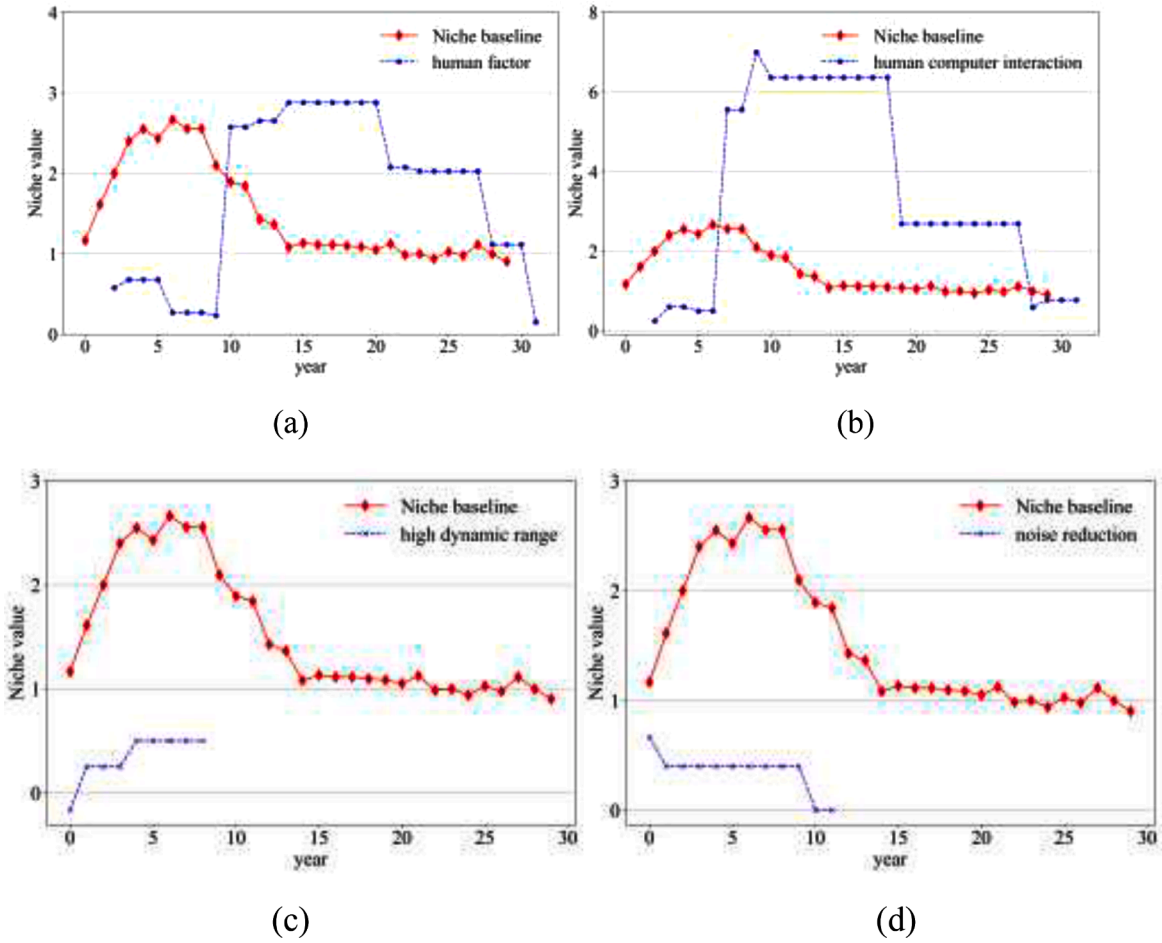

Fig. B2. Examples of positive and negative samples.

ecological niches. In a sense, we attempted to investigate the basic mechanisms of growth and change regarding emerging research topics. In practical terms, this study can help scientific policymakers to monitor areas of potential expenditure for scientific research. In addition, we created a bridge between quantitative and qualitative studies regarding emerging research topic detection. Although this paper proposed a novel method with better performance when detecting emerging topics, some shortcomings need to be improved in

future work. Future research can add to the method's complexity and incorporate various bibliometric indicators, such as community and scientific impact, to further verify the effectiveness and feasibility of these novel ideas from a knowledge ecology perspective.

CRediT authorship contribution statement

Jinqing Yang: Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Wei Lu:** Conceptualization, Methodology, Formal analysis, Supervision. **Jiming Hu:** Conceptualization, Writing – original draft, Writing – review & editing. **Shengzhi Huang:** Data curation, Formal analysis.

Acknowledgments

This work was partially supported by Major Projects of National Social Science Foundation of China (no. 17ZDA292), National Natural Science Foundation of China (no. 71874125) and The Young Top-notch Talent Cultivation Program of Hubei Province. We are also grateful to editors and anonymous reviewers for their helpful and valuable comments on our work.

Appendix A

A typical logistic function for S-shaped curve fitting was employed to ensure that keywords in the emergence stage were grouped with the positive samples since the "S" shape can indicate how the adoption of new keywords grows over time (Mahajan et al., 1990; Choi et al., 2015). The equation is as follows:

$$f(x) = \frac{k}{1 + ae^{bx}}$$

where x is the cumulative frequency of keywords in an annal year, and $a, b, k$are constants.

Appendix B

Fig. B-1. S-shaped curve fitting of "wireless sensor network". The X-axis indicates the number of years elapsed from the first year in which the research topic frequency was non-zero in reverse order; the right Y-axis represents the normalized number of cumulative frequency and the left is yearly frequency.

To help the reader better understand the test dataset, we randomly selected two positive samples, which are shown in Figs. B-2(a) & (b), and two negative samples, which are shown in Figs. B-2(c) & (d). In Fig. B-2(a), the niche value "human factor" is higher than the niche baseline in the 10th year. In Fig. B-2(b), "human computer interaction" emerges in the 7th year. In Fig. B-2(c) & (d), both negative samples are still lower than the niche values of the niche baseline.

Fig. B-2. Examples of positive and negative samples

Reference

Adams, J. (2012). The rise of research networks. *Nature, 490*(7420), 335–336.
Bao, B. K., Xu, C., Min, W., & Hossain, M. S. (2015). Cross-platform emerging topic detection and elaboration from multimedia streams. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 11*(4), 1–21.
Baum, J. A. C., & Singh, J. V. (1994). Organizational niches and the dynamics of organizational founding. *Organ. Sci., 5*(4), 483e501.
Bettencourt, L. M., Kaiser, D. I., & Kaur, J. (2009). Scientific discovery and topological transitions in collaboration networks. *Journal of Informetrics, 3*(3), 210–221.
Börner, K., & Scharnhorst, A. (2009). Visual conceptualizations and models of science. *Journal of Informetrics, 3*, 161–172.
Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics, 64*(3), 351–374.
Braam, R. R., Moed, H. F., & Van Raan, A. F. (1991). Mapping of science by combined co-citation and word analysis. I. Structural aspects. *J. Am. Soc. Inf. Sci., 42*, 233–251.
Braun, T., Schubert, A. P., & Kostoff, R. N. (2000). Growth and trends of fullerene research as reflected in its journal literature. *Chem. Rev., 100*(1), 23–38.
Chen, L. C. (2017). An effective LDA-based time topic model to improve blog search performance. *Information Processing & Management, 53*(6), 1299–1319.
Choi, J. Y., Jeong, S., & Kim, K. (2015). A study on diffusion pattern of technology convergence: Patent analysis for Korea. *Sustainability, 7*(9), 11546–11569.
Coccia, M. (2012). Evolutionary growth of knowledge in path-breaking targeted therapies for lung cancer: radical innovations and structure of the new technological paradigm. *International Journal of Behavioural and Healthcare Research, 3*(3-4), 273–290.
Coccia, M. (2018). General properties of the evolution of research fields: a scientometric study of human microbiome, evolutionary robotics, and astrobiology. *Scientometrics, 117*(2), 1265–1283.
Coccia, M. (2020). The evolution of scientific disciplines in applied sciences: dynamics and empirical properties of experimental physics. *Scientometrics, 124*(1), 451–487.
Coccia, M., & Finardi, U. (2012). Emerging nanotechnological research for future pathways of biomedicine. *International Journal of Biomedical nanoscience and nanotechnology, 2*(3-4), 299–317.
Coccia, M., & Finardi, U. (2013). New technological trajectories of non-thermal plasma technology in medicine. *Int. J. Biomed. Eng. Technol., 11*(4), 337–356.
Coccia, M., & Wang, L. (2015). Path-breaking directions of nanotechnology-based chemotherapy and molecular cancer therapy. *Technological Forecasting and Social Change, 94*, 155–169.
Cozzens, S., Gatchair, S., Kang, J., Kim, K. S., Lee, H. J., Ordóñez, G., & Porter, A. (2010). Emerging technologies: quantitative identification and measurement. *Technology Analysis & Strategic Management, 22*(3), 361–376.

Dang, Q., Gao, F., & Zhou, Y. (2016). Early detection method for emerging topics based on dynamic bayesian networks in micro-blogging networks. *Expert Syst. Appl., 57*, 285–295.
Dimmick, J. W., Patterson, S. J., & Albarran, A. B. (1992). Competition between the cable and broadcast industries: A niche analysis. *J. Media Econ., 5*, 13–30.
Elton, C. (1927). *Animal Ecology* (pp. 63–68). London: Sidgwick and Jackson.
Fanelli, D., & Glänzel, W. (2013). Bibliometric evidence for a hierarchy of the sciences. *PLoS One, 8*(6), e66938.
Firestone, J. M. (2008). On doing knowledge management. *Knowledge Management Research & Practice, 6*(1), 13–22.
Grubb, P. J. (1977). The maintenance of species-richness in plant communities: the importance of the regeneration niche. *Biological reviews, 52*(1), 107–145.
Hannan, M. T., Carroll, G. R., & Polos, L. (2003). The organizational niche. *Sociol. Theory, 21*(4), 309e340.
He, Q. (1999). Knowledge discovery through co-word analysis. *Library Trends, 48*(1), 133–159.
Hutchinson, G. E. (1957). Concluding Remarks. Cold Spring Harbor Symp. *Quant. Biol., 22*, 415–427.
Leydesdorff, L., & Cozzens, S. (1993). The delineation of specialties in terms of journals using the dynamic journal set of the SCI. *Scientometrics, 26*(1), 135–156.
Li, X., Zhang, A., Li, C., Ouyang, J., & Cai, Y. (2018). Exploring coherent topics by topic modeling with term weighting. *Information Processing & Management, 54*(6), 1345–1358.
Lu, W., Huang, S., Yang, J., Bu, Y., Cheng, Q., & Huang, Y. (2021). Detecting research topic trends by author-defined keyword frequency. *Information Processing & Management, 58*(4), Article 102594.
Mahajan, V., Muller, E., & Bass, F. M. (1990). New product diffusion models in marketing: A review and directions for research. *Journal of marketing, 54*(1), 1–26.
Mazloumian, A., Eom, Y. H., Helbing, D., Lozano, S., & Fortunato, S. (2011). How citation boosts promote scientific paradigm shifts and nobel prizes. *PLoS One, 6*(5), e18975.
Ohniwa, R. L., & Hibino, A. (2019). Generating process of emerging topics in the life sciences. *Scientometrics, 121*(3), 1549–1561.
Palage, K., Lundmark, R., & Söderholm, P. (2019). The innovation effects of renewable energy policies and their interaction: the case of solar photovoltaics. *Environmental Economics and Policy Studies, 21*(2), 217–254.
Peli, G. (2017). Population adaptation with newcomers and incumbents: The effects of the organizational niche. *Industrial and Corporate Change, 26*(1), 103–124.
Peng, B., Zheng, C., Wei, G., & Elahi, E. (2020). The cultivation mechanism of green technology innovation in manufacturing industry: From the perspective of ecological niche. *J. Cleaner Prod., 252*, Article 119711.
Peset, F., Garzón-Farinós, F., González, L. M., García-Massó, X., Ferrer-Sapena, A., Toca-Herrera, J. L., & Sánchez-Pérez, E. A. (2020). Survival analysis of author keywords: An application to the library and information sciences area. *Journal of the Association for Information Science and Technology, 71*(4), 462–473.
Porter, A. L., Garner, J., Carley, S. F., & Newman, N. C. (2019). Emergence scoring to identify frontier R&D topics and key players. *Technological Forecasting and Social Change, 146*, 628–643.
Raamkumar, A. S., Foo, S., & Pang, N. (2017). Using author-specified keywords in building an initial reading list of research papers in scientific paper retrieval and recommender systems. *Information Processing & Management, 53*(3), 577–594.
Raven, R. (2007). Niche accumulation and hybridisation strategies in transition processes towards a sustainable energy system: An assessment of differences and pitfalls. *Energy policy, 35*(4), 2390–2400.
Rong, K., Lin, Y., Li, B., Burström, T., Butel, L., & Yu, J. (2018). Business ecosystem research agenda: more dynamic, more embedded, and more internationalized. *Asian Bus Manage, 17*, 167–182.
Roshani, S., Bagherylooieh, M. R., Mosleh, M., & Coccia, M. (2021). What is the relationship between research funding and citation-based performance? A comparative analysis between critical disciplines. *Scientometrics, 126*(9), 7859–7874.
Savoy, J. (2013). Authorship attribution based on a probabilistic topic model. *Information Processing & Management, 49*(1), 341–354.
Seol, H., Park, G., Lee, H., & Yoon, B. (2012). Demand forecasting for new media services with consideration of competitive relationships using the competitive Bass model and the theory of the niche. *Technol. Forecast. Soc. Chang., 79*(7), 1217–1228.
Sice, P. V., Thirkle, S. A., & Ogwu, S. A. (2018). MIKE: Management, Information and Knowledge Ecology. *International Journal of Systems and Society (IJSS), 5*(1), 13–27.
Small, H., Boyack, K. W., & Klavans, R. (2014). Identifying emerging topics in science and technology. *Research policy, 43*(8), 1450–1467.
Smith, A., Voß, J. P., & Grin, J. (2010). Innovation studies and sustainability transitions: the allure of the multi-level perspective and its challenges. *Res. Policy, 39*(4), 435–448.
Srinivasan, R. (2008). Sources, characteristics and effects of emerging technologies: Research opportunities in innovation. *Industrial Marketing Management, 37*(6), 633–640.
Stary, C. (2014). Non-disruptive knowledge and business processing in knowledge life cycles–aligning value network analysis to process management, 18 (4), 651-686.
Tu, Y. N., & Seng, J. L. (2012). Indices of novelty for emerging topic detection. *Information processing & management, 48*(2), 303–325.
Van den Oord, A., & Van Witteloostuijn, A. (2018). A multi-level model of emerging technology: An empirical study of the evolution of biotechnology from 1976 to 2003. *PLoS One, 13*(5), Article e0197024.
Vulić, I., De Smet, W., Tang, J., & Moens, M. F. (2015). Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. *Information Processing & Management, 51*(1), 111–147.
Wang, Q. (2018). A bibliometric model for identifying emerging research topics. *Journal of the association for information science and technology, 69*(2), 290–304.
Wang, R., Zhou, D., & He, Y. (2019). Atm: Adversarial-neural topic model. *Information Processing & Management, 56*(6), Article 102098.
Weis, J. W., & Jacobson, J. M. (2021). Learning on knowledge graph dynamics provides an early warning of impactful research. *Nat. Biotechnol.*, 1–8.
Wu, C., Kanoulas, E., & de Rijke, M. (2020). Learning entity-centric document representations using an entity facet topic model. *Information Processing & Management, 57*(3), Article 102216.
Xu, J., Bu, Y., Ding, Y., Yang, S., Zhang, H., Yu, C., & Sun, L. (2018). Understanding the formation of interdisciplinary research from the perspective of keyword evolution: a case study on joint attention. *Scientometrics, 117*(2), 973–995.
Xu, S., Hao, L., An, X., Yang, G., & Wang, F. (2019). Emerging topics detection with multiple machine learning models. *Journal of Informetrics, 13*(4), Article 100983.
Yoo, S., Jang, S., Byun, S. W., & Park, S. (2019). Exploring human resource development research themes: A keyword network analysis. *Human Resource Development Quarterly, 30*(2), 155–174.
Yoon, Y. S., Zo, H., Choi, M., Lee, D., & Lee, H. W. (2018). Exploring the dynamic knowledge structure of studies on the Internet of things: Keyword analysis. *ETRI Journal, 40*(6), 745–758.
Yu, H., Wei, Y. M., Tang, B. J., Mi, Z., & Pan, S. Y. (2016). Assessment on the research trend of low-carbon energy technology investment: A bibliometric analysis. *Appl. Energy, 184*, 960–970.
Zhu, C. (1997). The niche ecostate-ecorole theory and expansion hypothesis. *Acta Ecologica Sinica, 17*(3), 324–332.