Scientometrics (2019) 120:167–185
https://doi.org/10.1007/s11192-019-03126-8

A novel method to identify emerging technologies using a semi-supervised topic clustering model: a case of 3D printing industry

Yuan Zhou[1] · Heng Lin[2] · Yufei Liu[1,3] · Wei Ding[2]

Received: 12 October 2018 / Published online: 24 May 2019

Abstract

There have been recent attempts to identify emerging technologies by using topic-based analysis, but many of them have methodological deficiencies. First, analyses are unsupervised, and unsupervised methods cannot incorporate supervised knowledge that is needed to better identify technological domains. Second, those methods lack semantic interpretation, as many of them still remain at word-level analyses, we developed a novel technology-identification method that uses a semi-supervised topic clustering model (Labeled Dirichlet Multi Mixture model) to integrate technological domain knowledge. The model also generates a sentence-level semantic technological topic description through the topic description method (Various-aspects Sentence-level Description) on information extraction. We used this novel method to analyze the technology of the 3D printing industry, and successfully identified emerging technologies by differentiating new topics from the traditional topics, the results effectively demonstrated the semantic technological topic description by showing sentences. This method could be of great interest to technology forecasters and relevant policy-makers.

Keywords Emerging technologies · Semi-supervised · Topic model · Sentence-level · Technological description · 3D printing

Introduction

There is existing research that attempts to identify emerging technological topics such as technological topic classification (Wang et al. 2014), major research themes identification (Lu and Liu 2016), or subject classification (Zhang et al. 2016). These works aim to study the structure of medium- large-sized document sets or monitor the evolution of research

✉ Yufei Liu
liuyufei0418@qq.com

[1] School of Public Policy and Management, Tsinghua University, Beijing 100084, China

[2] School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China

[3] Center for Strategic Studies, Chinese Academy of Engineering, Beijing 100088, China

fields and topics at local to global levels (Heeffer and Thijs 2017; Boyack 2017; Wang et al. 2018a). We refer to these methods, collectively, as technological topic segmentation in this paper. These literatures usually use word/citation-based methods that include topic trends (Watts and Porter 2003; Guo et al. 2011; De Rassenfosse et al. 2013), co-word (Furukawa et al. 2015) or term-clumping (Zhang et al. 2014), citation or co-citations (Leydesdorff and Rafols 2009; Zhou et al. 2019; Kajikawa and Takeda 2008; Shibata et al. 2011; Cho and Shih 2011; Kong et al. 2017; Zhou et al. 2016, 2018), and combinations of these methods (Zhang et al. 2010; Wang et al. 2018b; Small et al. 2014; Breitzman and Thomas 2015). However, these methods have some limitations, depicted as follows: (1) The methods are word- or citation-based, and they have limited capability to utilize the semantic information that is embedded in the publications and patents. This requires integration of natural language processing (NLP) technologies and machine learning algorithms for scalable data processing (Kong et al. 2017; Yang et al. 2016). (2) The terminologies of technology fields are dynamic and keep changing. These methods mainly deal with existing topics, and the elicitation of new technological topics from existing ones remains difficult (Rotolo et al. 2015).

Some recent studies have used text-mining methods to identify new technological terms. These methods do not rely on the traditional word/citation data, and can be used for a variety of semantic analyses not limited to existing technological topics. For example, some methods have used unsupervised text-clustering methods to identify the characteristics of diffused technological topics (Wang and Koopman 2017; Yau et al. 2014; Roche et al. 2010). In addition, some methods have used topic models to achieve topic segmentation (Wang et al. 2014; Jeong and Song 2014; Ding 2011). However, existing topic model and text-clustering methods identify topics without combining and contrasting them with existing topics, and they are unsupervised. It cannot determine whether the identified topics are newly emerged (Waltman et al. 2010), and cannot integrated technological domain knowledge for a better clustering.

In order to address these problems, supervised classification methods that integrate existing domain knowledge have also been tentatively used in recent years, such as Support Vector Machines (SVMs) and neural networks (Kong et al. 2017; Venugopalan and Rai 2015; Kim and Choi 2014; Kim et al. 2018; Liu et al. 2019). These methods are suitable for analyzing existing topics but are unable to identify new topics that are beyond the scope of existing classification categories. Therefore, a novel method that combines supervised machine learning with unsupervised methods may be very useful to identify new technological topics, in order to fully utilize the advantages of both methods—integrating domain knowledge with supervised learning and discovering new/uncertain topics with unsupervised ones. In addition, this novel method needs to distinguish between the new topics and old ones in order to better identify the newly emerged technologies—this needs the advanced semantic description method in sentence-level, rather than the traditional keyword-based methods that cannot differentiate topics in the same technological field that often contain similar vocabulary.

Therefore, this paper proposes a novel method that combines a semi-supervised clustering model for topic segmentation and a sentence-level semantic description method for topic description. In the process of topic segmentation, a semi-supervised text-clustering model, the Labeled Dirichlet Multi Mixture (Labeled-DMM or LDMM), is used to integrate domain knowledge into technological segmentation processes, and the topics generated by topic segmentation are compared with the old topic list to identify the new ones. In the process of topic description, the Various-aspects Sentence-level Description information extraction method (VSD) is used to extract topics' semantic description at the sentence-level. This description is more explicit and specific compared to the word-level description generated by traditional keyword-based method, and it can elicit different topics by comparing to old ones—this is crucial to find emerging topics. This paper selects 3D printing technology as the case study to use the novel method for identifying emerging topics, and the results show this method is valid. This study contributes to literature by proposing a novel semi-supervised topic clustering model; in addition, it also integrates a process of topic extraction at the sentence-level, which extracts the semantic content of topics that help to better identify new topics.

The rest of this paper is organized as follows. “Literature review” section briefly presents the literature review. “Methodology” section proposes the methodology. “Case study” section analyzes the case study of 3D printing technology. “Conclusions” section concludes the paper.

Literature review

Technological topic segmentation has variable approaches. Keyword occurrence and co-word occurrence are common (Lee 2008; Hofer et al. 2010; An and Wu 2011; Schiebel et al. 2010). This method extracts keywords provided by the publication author or obtained from the title as a technological topic. These methods are simple and effective, but they are vulnerable to the inconsistent value of the keywords. Also, the description of the topic consists of a single word or phrase, which is often inadequate for interpretation. Citation analysis is another common type of technological topic segmentation method (Boyack and Klavans 2010; Small et al. 2014; Ding 2011; Upham and Small 2010). This method is usually based on the references or citation relationships between articles. The articles are clustered in different categories, and the resulting categories are regarded as technological topics. This method is widely used but its value is limited by the citation relationships. The generated categories cannot directly provide a description of this category. Technological topic segmentation based on text mining differs from these two methods because it analyzes potential technological topics through the semantic relationship of texts. Wang et al. (2014) uses LDA model for identifying underlying topic structures based on the latent relationships of technological words extracted.

Their study reveals emerging hot spots of LTE technology. Jeong and Song (2014) used a topic model to estimate the optimal time gaps among three resources (papers, patents, and web news articles) in two research domains, computer science and medical science. Yau et al. (2014) clustered scientific documents with topic modeling and used the method to identify energy technology topics. These methods do not rely on keywords or citation relationships. However, they have the problems previously described. Machine learning has provided many semi-supervised topic models or text clustering methods. These methods have great potential for solving segmentation process problems. Andrzejewski et al. (2009) proposed the DF-LDA. This is a supervised model and the result can be influenced by co-occurrence of word pairs. LTM was proposed by Chen and Liu (2014). LTM was the first lifelong-learning topic model that could learn from historical data to obtain field information for a new round of topic modeling. Nigam et al. (2000) used the Dirichlet multi mixture (DMM) model with labeled and unlabeled documents for documents division. Chen and Liu (2014) combined the representation learning process and the K-means clustering process and used labeled and unlabeled data to implement short text clustering. By using domain knowledge as supervisory knowledge, these methods can be used to implement

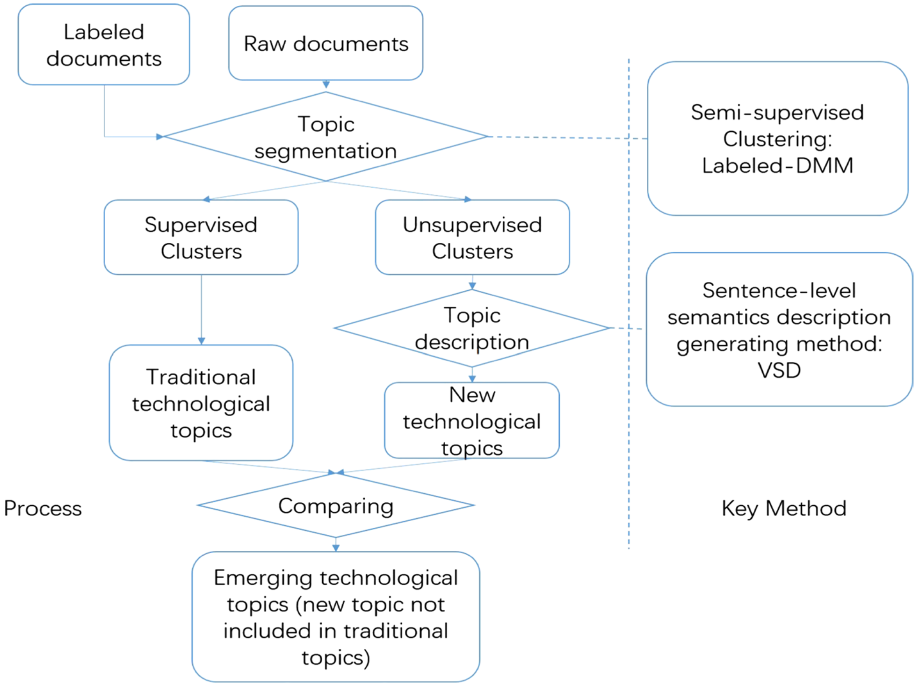

Labeled documents
Raw documents
Topic segmentation
Semi-supervised Clustering: Labeled-DMM
Supervised Clusters
Unsupervised Clusters
Topic description
Sentence-level semantics description generating method: VSD
Traditional technological topics
New technological topics
Process
Comparing
Key Method
Emerging technological topics (new topic not included in traditional topics)

Fig. 1 Novel method for emerging technology identification

the segmentation process better with domain knowledge. The development of information extraction also has great reference value for solving problems in the description process. Angeli et al. (2015) used leveraging linguistic structure mining subject-action-object like semantic relations. SAO (subject-action-object) based method is also used in patent analysis for technology planning, identifying patent infringement and identifying technological trends (Choi et al. 2012; Park et al. 2012; Yoon and Kim 2011).

Wu and Weld (2010) used heuristic matches between Wikipedia info box attribute values and corresponding sentences to construct training data. They noted good performance for open information extraction. By mining the various description aspects of a technology in the text, a more comprehensive technological topic description can be generated.

Methodology

This section presents the overall research process, particularly focusing on semi-supervised topic clustering model and sentence-level semantics description method to identify the emerging technology topics from traditional technology topics.

Overall research process

In order to carry out the analysis, the overall research process is shown in Fig. 1. The method uses traditional technological topics and the corresponding labeled document as supervision data. Topic segmentation is implemented using the semi-supervised text clustering Labeled-DMM. Among the segmentation results, the supervised clusters matter of labeled data will automatically correspond to the traditional technological topics. Clusters without labeled data will form unsupervised clusters. The sentence-level semantics descriptions of these unsupervised clusters are generated by the information extraction based Various-aspects Sentence-level Description method (VSD). Together, these clusters and descriptions constitute new technological topics. By comparing the traditional technological topics with the newly generated technological topics, emerging technological topics can be discovered. There are three main steps in the process:

- Selection of Traditional Technological topics and acquisition of corresponding labeled data.
- Topic segmentation based on semi-supervised text clustering model Labeled-DMM.
- Information extraction-based sentence-level semantics topic description generating method, VSD.

Data and traditional technological topics

The first step of the method is to obtain the raw documents and labeled data for the supervised clustering part. These topics can be obtained from the results of a technological topic segmentation, expert knowledge, a literature review, or existing popular classification criteria such as WOS classification or IPC. In selecting technical topics, the degree of coincidence between technological topics should be ensured. For example, rapid prototyping technology includes stereolithography technology, so these two technologies should not appear in the same segmentation process. Topic-related documents can be obtained by searching the database or labeling by technical experts.

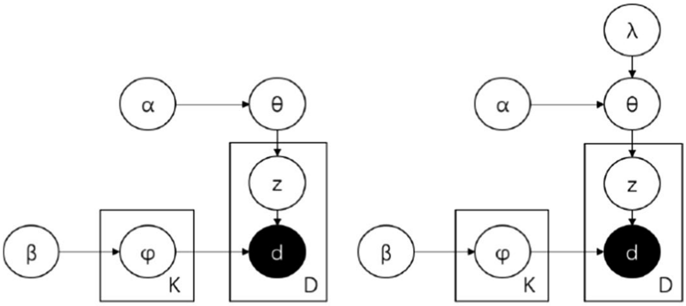
Technological topics and corresponding label data were used as the domain knowledge to guide the entire segmentation process. The selection criteria of traditional technological topics have a great influence on the generated topics. Clusters with supervised data will directly correspond to traditional technological topics. Selecting different traditional technological topics and labeled data, unsupervised clusters may also vary greatly. In practical applications, suitable technological topics should be selected according to the specific scenario to be analyzed. For example, using 3D printing technology as the research subject, if the entire production process is of interest, technological topics can be divided into materials/processes/equipment/applications. If a specific process is of interest, technological topics can be divided into Electron Beam Melting/Fused Deposition Modeling/Selective Laser Sintering/Stereolithography.

In this study, both raw documents and labeled data are from scientific articles obtained from the WOS database. The sector analyzed was the technology of 3D printing. The term “3D printing” commonly refers to additive manufacturing (AM). Traditional technological topics were selected according to an overview paper on additive manufacturing written by Wong and Hernandez (2012). In this paper, nine kinds of AM processes were noted and these processes are used as technological topics. The category of labeled data is determined by traditional technological topics. Search literature for each technological topic before 2012 (inclusive) in WOS. Select the literature type as article and search date as March 12, 2018. The abstracts are extracted after the articles are downloaded. Remove the invalid and duplicate abstracts and use the remaining abstracts as labeled data. Search formula and quantity of labeled data is shown in Table 1.

Retrieve the literature for TS= “Additive Manufacturing” in 2013–2017, select literature type as article, and search date is March 12, 2018. Extract abstracts after downloading as raw

Table 1 Data

Name	Query	Searched	Supervised
Stereo lithography apparatus (SLA)	TS = (“Stereolithography” or “stereolithography apparatus”)	1701	1653
Polyjet	TS = “Polyjet”	16	15
Fused deposition modeling (FDM)	TS = (“Fused deposition modeling” or “fused filament fabrication”)	233	227
Laminated object manufacturing (LOM)	TS = “Laminated object manufacturing”	154	150
3DP	TS = (“3*D printing” OR “three-dimensional printing” OR “3-dimensional printing”)	526	501
Prometal	TS = “Prometal”	7	7
Selective laser sintering (SLS)	TS = “Selective laser sintering”	912	888
Laser engineered net shaping (LENS)	TS = (“Laser engineered net shaping” OR “Laser powder forming”)	138	134
Electron beam melting (EBM)	TS = (“Electron-beam additive manufacturing” or “electron-beam melting” OR “electron beam additive manufacturing” or “electron beam melting”)	264	257



(a) GSDMM **(b)** Labeled-GSDMM

Fig. 2 Graphical model

Table 2 Notations

Notation	Meaning
D	Number of documents in the corpus
K	Number of mixture clusters
z_d	Cluster label of document *d*
$\vec{d}$	Documents in the corpus
$\vec{z}_{\neg d}$	Cluster labels of each document except for document *d*
$m_{z,\neg d}$	Number of documents in cluster z except for document *d*
$n_{z,\neg d}$	Number of occurrences of word w in cluster z except for document *d*
$n^{w}_{z,\neg d}$	Number of documents in cluster z except for document *d*
N_d	Number of words in document *d*
N^{w}_{d}	Number of occurrences of word w in document *d*

documents to be segmented (total of 5019 abstracts). In these documents, 20 documents for each topic, including SLA, FDM, SLS, and EBM, were manually labeled to test the effectiveness of semi-supervised clustering.

Semi-supervised text clustering Labeled-DMM

To identify the emerging technology topics from traditional technology topics, we chose the unsupervised text clustering model Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture (GSDMM) and improved it to form a semi-supervised text clustering model Labeled-DMM. The GSDMM model is a probabilistic generative model proposed by Yin and Wang (2014). The model is more efficient to solve by Gibbs Sampling, and has good performance in short text clustering (it can cope with the sparse and high-dimensional problem of short texts).

The GSDMM model is an unsupervised model shown in Fig. 2a. Assume that the documents to be segmented have k implicit clusters. In the Gibbs sampling process, the implicit topic of each document is sampled under the k topics during each iteration. After all iterations are completed, the sample expectation is the document's implicit cluster. The original GSDMM sampling formula is shown in formula (1). The meanings of variables in the formula are shown in Table 2.

$$p(z_d = z|\vec{z}_{\neg d}, \vec{d}) \propto \frac{m_{z,\neg d} + \alpha}{D - 1 + K\alpha} \cdot \frac{\prod_{w\in d}\prod_{j=1}^{N_d^w}(n_{z,\neg d}^w + \beta + j - 1)}{\prod_{i=1}^{N_d}(n_{z,\neg d} + \beta + i - 1)} \tag{1}$$

Since the model is unsupervised and cannot directly apply domain knowledge, improvements are made to the model. Input data is divided into two categories, supervised documents and documents to be segmented. Supervised documents contain known topic labels. Each document can have multiple topics, but must have at least one topic. The documents to be segmented have no known topic labels. Enable supervised document to be sampled only under known topic labels it contains. By limiting the sampling topic range, supervision documents will be divided into expected topics. For unlabeled documents whose topic distribution is similar to the supervised documents, the probability of being assigned to the corresponding topic also increases. The sampling process is limited by adding the parameter λ. The improved semi-supervised model is named Labeled-DMM. The graphical model of Labeled-DMM is shown in Fig. 2b. The sampling formula for Labeled-DMM is

$$\begin{aligned} p(z_d = z|\vec{z}_{\neg d}, \vec{d}) &\propto \lambda \cdot \frac{m_{z,\neg d} + \alpha}{D - 1 + K\alpha} \cdot \frac{\prod_{w\in d}\prod_{j=1}^{N_d^w}(n_{z,\neg d}^w + \beta + j - 1)}{\prod_{i=1}^{N_d}(n_{z,\neg d} + \beta + i - 1)} \\ \text{where, } \lambda &= \begin{cases} 0 & \text{document dis supervised document} \\ & \text{and topic } z \text{ is not in label list} \\ 1 & \text{else} \end{cases} \end{aligned} \tag{2}$$

Using Gibbs Sampling to solve the model, the entire solution process is as Table 3 shows:

Sentence-level semantics description generating method VSD

To describe technological topics more comprehensively and reflect the difference between topics, we propose a sentence-level semantics technological topic description generating method VSD. The description generated by this method is divided into three parts (Fig. 3):

- Technology names and common terms
- Methods or equipment used by the technology.
- Phrases describing the differences between technological topics.

Mining the noun phrases in the three parts of all documents under a topic and combining these phrases into sentences through rules. Noun phrases are mined by the *N*-gram method. *N*-gram refers to the sequence of *N* items in a given piece of text or speech. When the item is a word, a sequence formed by *N* words that are always consecutive can be found. When the last word of the *N* words is a noun, the sequence has a high probability of being a notional noun phrase.

The first part of the description is the technology name and common terms. The excavated noun phrases are sorted according to the frequency of occurrence. The Noun phrases

Table 3 Algorithm1: Gibbs Sampling for Labeled DMM

```
Sampling Process
Data: Documents in the input $D^t$
Result: Cluster labels of each document $\vec{z}$
begin
initialize parameters K,$\alpha, \beta$,I(iteration number);
initialize $m_z = n_z = n_z^w = 0$
for each document $d \in D^t$ do
    sample a cluster for d:
    $z_d \leftarrow z \sim Multinomial(1/K)$
    $m_z \leftarrow m_z + 1$ and $n_z \leftarrow n_z + N_d$
    for each word $w \in d$ do
        $n_z^w \leftarrow n_z^w + N_d^w$
    end for
end for
for i = 1 to I do
    for each document $d \in D^t$ do
        $m_z \leftarrow m_z - 1$ and $n_z \leftarrow n_z - N_d$
        for each word $w \in d$ do
            $n_z^w \leftarrow n_z^w - N_d^w$
        end for
        sample a cluster for d:
        $z_d \leftarrow z \sim p(z_d = z|z_{\neg d}, d)$
        $m_z \leftarrow m_z + 1$ and $n_z \leftarrow n_z + N_d$
        for each word $w \in d$ do
            $n_z^w \leftarrow n_z^w + N_d^w$
        end for
    end for
end for
```

mined are sorted by frequency of occurrence. The phrase with the highest number of occurrences is used as the technology name, and the other n high-frequency words are used as common terms. The second part is the method or equipment used by the technology. These phrases are mined by the OpenIE method. OpenIE is a type of information extraction method designed to extract relation triples from sentences. For instance, in the sentence “Einstein constructed the theory of relativity”, a subject-relation-object triple, Einstein/constructed/the theory of relativity, can be extracted. This method has been used for the construction of knowledge maps. In all relation triples under a topic, extract triples whose relations is use/apply/employ. The objects of these triples have a high probability of describing the method or equipment that will used in this technology. There are a total of m such phrases, called the method phrases. The third part is phrases that describe the differences between technologies. The topic descriptions mined through the first two parts may appear to be very similar, and these similar topics need to be distinguished. Assume that all Noun phrases mined under a topic constitute a large document, and K topics correspond to

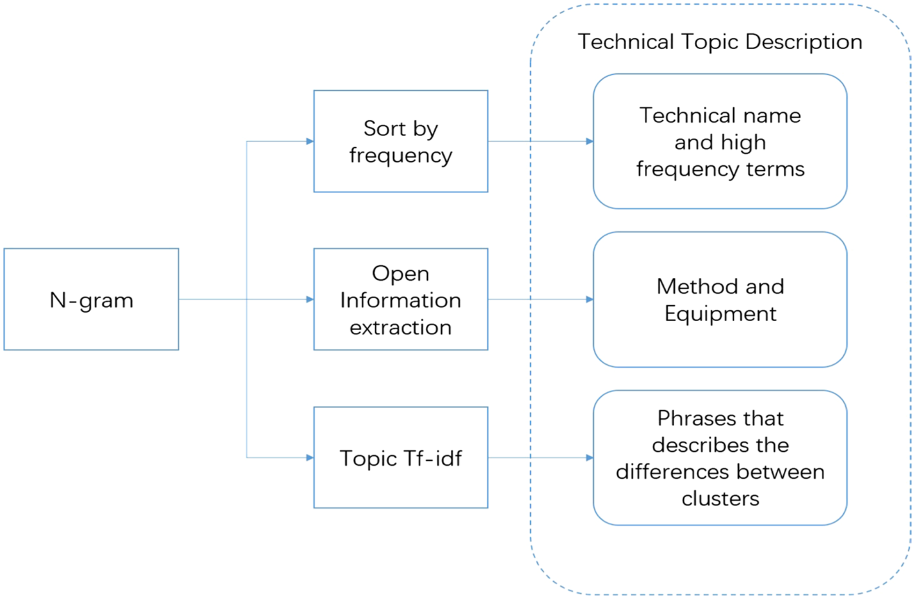

Technical Topic Description
N-gram
Sort by frequency
Open Information extraction
Topic Tf-idf
Technical name and high frequency terms
Method and Equipment
Phrases that describes the differences between clusters

Fig. 3 Technological Topic Description Generating Method VSD

K large documents. Use TF-IDF to assign weights to phrases in large documents. TF-IDF (Term Frequency-Inverse Document Frequency) is a commonly used weighting method. TF is high, indicating that the phrases appears frequently in the document. High IDF indicates that the phrases are highly discriminating between documents. After being weighted by TF-IDF, phrases that are important within the document and are discriminating between documents will get higher weights. Take the p words with the highest TF-IDF weights in the topic as high TF-IDF phrases, and the q words with the highest weight of all other topics as the low TF-IDF phrases in the topic. To increase discriminability, high TF-IDF phrases are phrases that appear only in this topic, and low TF-IDF phrases remove phrases that appear in the topic. Finally, sentence-level semantics technological topic description takes the following format:

This Cluster is technology name. This technology involves n^* common terms, etc. It usually use method or equipment like m^* method phrase. It differs from other technologies in that it is more focused on p^* high TF-IDF phrases, while less talking about q^* low TF-IDF phrases.

Case study

The 3D printing technology is an important component of industrial development and manufacturing. This paper uses the 3D printing sector as an example to verify that the proposed method can identify emergency topics in scientometric analysis.

Table 4 Metrics

Labeled				Unlabeled			
K	*H*	*C*	*V*-m	*K*	*H*	*C*	*V*-m
100	0.74	0.47	0.58	100	0.56	0.34	0.42
200	0.76	0.48	0.59	200	0.52	0.32	0.40
300	0.74	0.47	0.58	300	0.50	0.31	0.39
400	0.75	0.47	0.58	400	0.47	0.31	0.37
500	0.71	0.46	0.56	500	0.42	0.29	0.34

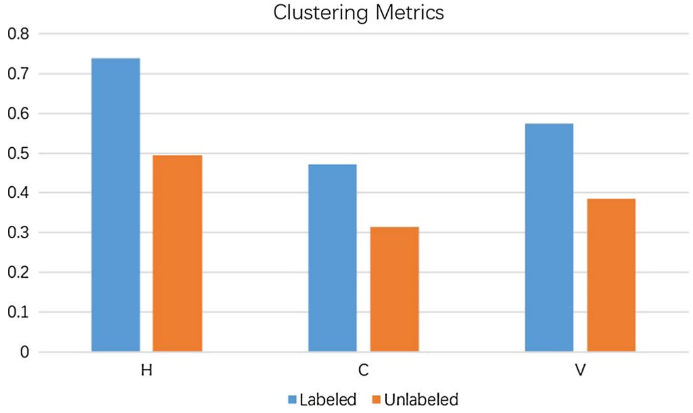


Fig. 4 Clustering method

Segmentation result

The semi-supervised text clustering method Labeled-DMM was used to segment the AM technology related literature. The clustering metrics *H–C–V* were used to evaluate the results of the segmentation. The data used in the evaluation are 80 manually-labeled articles. In *H–C–V* metrics,

- Homogeneity (*H*), describes whether the data in a result cluster belongs to the same topic.
- Completeness (*C*) describes the situation where data originally belonging to the same topic still belong to the same topic in the result cluster.
- *V*-measure (*V*-m), is a balance metric of the *H* and *C* metrics.

Here, we directly used the *H–C–V* metrics analysis tool provided by sk-learn to calculate the final metrics. The number of initial *K* clusters needs to be set during the segmentation process According to GSDMM, the number of final clusters can be automatically determined. To verify whether this characteristic is retained after improving to semi-supervised, the cases of $K=100$, 200, 300, 400, and 500 are discussed. We set the super-parameter of the GSDMM as $\alpha = 0.1$ and $\beta = 0.05$, then repeated the clustering 10 times in each case and determined the average of the evaluation metrics. The topics were segmented using our

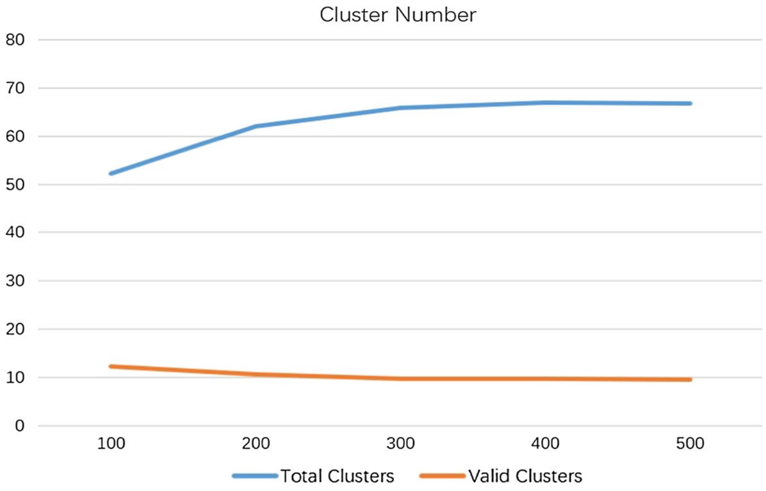


Fig. 5 Cluster amount

method and the traditional unsupervised GSDMM model respectively. The result metrics are shown in Table 4.

Table 4 shows that the final clustering metrics are similar in the five levels of *K*. This result shows that when *K* is much larger than the expected clusters, the *K* value has little effect on the segmentation result. The average of the metrics in our semi-supervised text clustering method and the traditional unsupervised method is shown in Fig. 4. The table and figure show that the segmentation metrics are improved after addition of the supervised data. The addition of domain knowledge improved the segmentation results. This supports the effectiveness of the semi-supervised text clustering based technological topic segmentation. To implement topic segmentation more accurately, it was necessary to add appropriate supervised data.

The number of clusters obtained after segmentation is shown in Fig. 5. Total clusters is the total number of clusters obtained while valid clusters is the number of clusters whose containing document is larger than 0.5% of the total number of documents (In this study the number of valid clusters equals 25). When the number of documents contained in a cluster is too small, it is difficult to analyze, and these clusters are considered as noise clusters or invalid clusters. With the increase of *K*, total clusters increased, but the number of valid clusters was almost unchanged at around 10. This result is consistent with the feature that GSDMM can automatically determine the number of clusters. It shows that in the semi-supervised clustering, the Labeled-DMM model retains the feature of automatically determining the number of clusters.

Description result

After segmentation, the clusters required description to discover any potential emerging topics that typically distinguish them from traditional technologies. Taking a group of result topics with better metrics, description of first part and second are generated for each cluster, namely the description of common terms and method phrase. Since there are many important noun abbreviations in 3D printing technology, the first part of the description includes noun abbreviations in addition to noun phrases. For the clusters whose first two parts are highly consistent, a third part description was generated to make a detailed distinction. We used the natural language parser and OpenIE tools provided by the Stanford NLP team to extract the N-grams and relational triples, where $2 \leq N \leq 4$. In order to make the description more readable, stop words were removed before the phrase was extracted. Not only commonly used stop words are removed, such as a/the/as, but also some other common words in this field are removed. These words have specific meanings, but it is difficult to identify the distinction between some topics such as addictive/manufacturing/printing. The results of the first two parts of the description are shown in Table 5. The number of documents in each cluster usually did not exceed 1000. There were few relation triples whose relation is use/employ/apply. This leads to the results that the frequency of the most high-frequency method phrase is 1. In this case, we manually selected 3 of the extracted method phrases for display. The parameters in the description take $a = b = 3$ and $c = d = 5$.

Table 5 Topic description

Cluster	Description
1	This Cluster is FDM. This technology involves thermal expansion, binder jetting, inkjet printing, etc. It usually uses method or equipment like digital-subtraction technique, synthetic membrane, peroxide combination
2	This Cluster is EBM. This technology involves beam melting, electron beam, electron beam melting, etc. It usually uses method or equipment like electron beam, beam melting, electron beam melting
3	This Cluster is SLS. This technology involves laser sintering, selective laser, selective laser sintering, etc. It usually uses method or equipment like downer reactor, forcesensitive resistor, low-profile resistor
4	This Cluster is FDM. This technology involves fused deposition, deposition modeling, fused deposition modeling, etc. It usually uses method or equipment like fused deposition, 3d printer, electron microscope
5	This Cluster is LENS. This technology involves net shaping, laser net, laser net shaping, etc. It usually uses method or equipment like 3-d electrode atom probe, 3-d electrode, laser net
6	This Cluster is computed tomography. This technology involves SLA, rapid prototyping, surgical planning, etc. It usually uses method or equipment like 3d scanner, computed tomography, dlp-based 3d printing
7	This Cluster is inkjet printing. This technology involves FDM, additive printing, rapid prototyping, etc. It usually uses method or equipment like 3d printing, printed circuit board, silver nanoparticle
8	This Cluster is SLM. This technology involves bone tissue, porous structures, selective laser, etc. It usually uses method or equipment like selective laser, 3d printing, additively porous niti
9	This Cluster is SLM. This technology involves laser melting, selective laser, selective laser melting, etc. It usually uses method or equipment like electron microscopy, optical microscopy, laser metal deposition
10	This Cluster is SLM. This technology involves selective laser, topology optimization, laser melting, etc. It usually uses method or equipment like 3d printer, finite element, element model
11	This Cluster is supply chain. This technology involves rapid prototyping, life cycle, product development, etc. It usually uses method or equipment like 3d printing, provide field, computed tomographic imaging

Table 6 Description of cluster1/4/7

Cluster	Description
1	It differs from other technologies in that it is more focused on drug loading, drug release, soft tissue, green body, shape retention, while less talking about surface roughness, aerosol jet, flow rate, radio frequency and raster angle
4	It differs from other technologies in that it is more focused on raster angle, build orientation, extrusion temperature, reinforced plastic, tensile flexural, while less talking about aerosol jet, inkjet printing, flow rate, radio frequency and laser melting
7	It differs from other technologies in that it is more focused on aerosol jet, flow rate, radio frequency, laser melting, flexible electronics, while less talking about raster angle, build orientation, extrusion temperature, drug loading and reinforced plastic

Table 7 Description of cluster8/9/10

Cluster	Description
8	It differs from other technologies in that it is more focused on adaptive slicing, conformal cooling channels, shape deformation, tool paths, automatic control, while less talking about bone tissue, heat treated, columnar grains, heat treatments and microstructure mechanical
9	It differs from other technologies in that it is more focused on bone tissue, tissue regeneration, porous scaffolds, regenerative medicine, epsilon caprolactone, while less talking about melt pool, laser power, heat treated, columnar grains and heat treatments
10	It differs from other technologies in that it is more focused on heat treated, columnar grains, heat treatments, microstructure mechanical, energy input, while less talking about bone tissue, tissue regeneration, topology optimization, porous scaffolds and regenerative medicine

In the result clusters, clusters 1–6 are labeled clusters, and clusters 7–11 are unlabeled clusters. Table 5 shows that the descriptions of clusters 2, 3, 5, 6, and 11 are clear, and the discrimination with other clusters is relatively high. Clusters 1/4/7 all include FDM, and clusters 8/9/10 all naming SLM. These cluster descriptions are overlapping.

The third part description of clusters 1/4/7 is shown in Table 6. From Table 6, cluster 4 describes FDM, where raster angle, extrusion temperature, tensile flexure, and other descriptions are more biased toward the FDM process. Both cluster 1 and cluster 7 describe the 3DP process, but cluster 7 is more biased toward the printing of electronic materials, while cluster 1 is biased towards the broader 3DP process.

The names of clusters 8/9/10 are all SLM, and they all have similar aspects in common terms and method phrase. It can be considered that the three topics are different aspects of the same technical topic. The third part description of clusters 8/9/10 is shown in Table 7. Cluster 8 is more inclined to the macroscopic process. Cluster 9 is biased towards the application of the technology in the medical field. Cluster 10 is biased toward temperature changes and wafer changes in the melting process.

Emerging technologies identification

After understanding the meaning of technological topics using sentence-level semantics topic description, there are clear differences and relations between the segmented topics and the traditional technological topics. Emerging technologies can be obtained by identifying technical topics that have not been the subject of traditional technologies. To more clearly reflect the comparison results, the generated topics were divided into the following three cases.

- The technologies contained in the traditional technological topics. Clusters 1–7 are topics that are contained in traditional technology topics. Their labels are consistent, and they also show their relevance in the topic description. For example, 3DP's commonly used term "ink printing," EBM common method "beam melting," and FDM's high-TF-IDF word "extrusion temperature." This description shows the relevance of these topics and technologies. The existence of these technological topics reflects the fact that related technologies still have technological invention activities.
- Technologies not included in the traditional technological topics. Clusters 8/9/10 are topics not included in the traditional technological topics. Their names are all SLM, and common terms and common device methods are similar. These three topics correspond to the same technology SLM. Its unique terms and methods, such as "laser melting" and "fiber laser" indicate the differences between SLM and traditional topics. This topic is considered to be a potential emerging technology topic.
- Topics that do not describe technology. Cluster 11 describes the supply chain, which has little to do with the 3D printing process, but is related to the entire 3DP industry chain. Since the search query is "additively manufacturing", articles that are not related to technology may also be included. This topic is not an emerging technology topic.

In summary, new technological topics were obtained after segmenting using semi-supervised text clustering Labeled-DMM and being described by the sentence-level semantics description generation method VSD. Comparing these new topics with traditional topics, the potential emerging technology SLM that is not included in traditional technological topics was found. The result shows that the proposed method can identify emerging technologies that are not included in the traditional technological topics and demonstrates that this method can be applied to the identification of emerging technologies.

There are technologies included in the traditional technological topics that are not obtained in the segmentation topics. In the traditional technological topics, polyjet, prometal, and LOM were not found. There are two possible causes for this. First, supervised data are lacking. The number of papers that can be retrieved on the WOS is comparatively small. This results in less supervised data for related topics and makes it difficult to provide effective supervision in the segmentation process. Second, raw documents are lacking. Three keywords, polyjet, prometal, and LOM, were directly searched in the raw documents, and the number of articles obtained was less than 10. This phenomenon indicates that these technologies may have less active technical activities and while still following the traditional technological trajectories.

In the process of segmentation, the results for each segmentation process may differ. There are many groups of results in which SLM and EBM are grouped together because of similarities such as using the same melting process. In practical application, several groups of experiments should be done to get good results as the final reference. Although this method uses the 3D printing domain knowledge in the topic segmentation, it does not take advantage of unique 3D printing technology characteristics such as special processes or principles. The existing technological topics of 3D printing and related documents need to be obtained to identify emerging topics. This method therefore identifies the emerging technological topics of 3D printing technology. This method should also be able to achieve the identification of emerging technological topics in biomedicine, smart manufacturing

Table 8 Comparison of methods for topic-clustering-based methods and our method

Methodologies	Topic identification ability	Topic semantic level
Topic-clustering-based methods	Able to identify emerging topics	Word- or phrase-level semantic description
Our method	Able to identify emerging topics as well as combining and contrasting with existing topics	Sentence-level semantic description

and other fields, if technological topics and related documents in the corresponding fields can be obtained.

Conclusions

This paper proposes a novel method that integrates the semi-supervised text-clustering model and the sentence-level semantic extraction to identify emerging technological topics. The topic segmentation uses the labeled-DMM model that incorporates the industrial domain knowledge (with technological experts) throughout the segmentation process. The topic description uses the VSD method that successfully generates semantic technological topics description at sentence-level.

The key findings and contributions include the following: First, this study proposes a novel method that can be used to identify the emergence of new technological topics and differentiate new topics by contrasting to old technological topics. This method is useful for analyzing fast-changing technological or industrial domains that have newly emergent, state-of-the-art technologies. We also present a table that shows the differences between the proposed model and the topic-clustering-based methods that have been used to forecast the emerging technology (Table 8). As shown in Table 8, the main difference between our method and the existing topic-clustering approach is that the new framework explores future possibilities and understands the current topics of emerging technologies through both analysis of technological topic segmentation and description. Second, according to the analysis using the new method, the major promising technologies in new innovations of 3D printing is identified. This provides an opportunity for policy-makers and industrialists to develop 3D printing innovation strategies in the future.

There are some limitations that require further study. The method is intended to explore a future-oriented analysis of technological topics, rather than to forecast a specific event with a high uncertainty. A scenario-planning method may be useful to be integrated in further research if we wish to forecast the path-independencies produced by disruptive changes, such as technological breakthroughs.

Acknowledgements This research is supported by the National Natural Science Foundation of China (91646102, L1824039, L1724034, L1624045, L1524015, 71203117), the MOE (Ministry of Education in China) Project of Humanities and Social Sciences (Engineering and Technology Talent Cultivation) (16JDGC011), the National Science and Technology Major Project "High-end Numerical Control and Fundamental Manufacturing Equipment" (2016ZX04005002), Beijing Natural Science Foundation Project (9182013), the Chinese Academy of Engineering's China Knowledge Centre for Engineering Sciences an Technology Project (CKCEST-2019-2-13, CKCEST-2018-1-13, CKCEST-2017-1-10, CKCEST-2015-4-2), the UK-China Industry Academia Partnership Programme (UK-CIAPP\260), as well as the Volvo-supported Green Economy and Sustainable Development Tsinghua University (20153000181) and Tsinghua Initiative Research Project (2016THZW).

References

An, X. Y., & Wu, Q. Q. (2011). Co-word analysis of the trends in stem cells field based on subject heading weighting. *Scientometrics, 88*(1), 133–144.

Andrzejewski, D., Zhu, X., & Craven, M. (2009). Incorporating domain knowledge into topicmodeling via dirichlet forest priors. In *International conference on machine learning* (pp. 25–32). ACM.

Angeli, G., Premkumar, M. J. J., & Manning, C. D. (2015). Leveraging linguistic structure for opendomain information extraction. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing* (Vol. 1: Long Papers. 1, pp. 344–354).

Boyack, K. W. (2017). Investigating the effect of global data on topic detection. *Scientometrics, 111*(2), 999–1015.

Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the Association for Information Science and Technology, 61*(12), 2389–2404.

Breitzman, A., & Thomas, P. (2015). The emerging clusters model: A tool for identifying emerging technologies across multiple patent systems. *Research Policy, 44*(1), 195–205.

Chen, Z., & Liu, B. (2014). Topic modeling using topics from many domains, lifelong learning and big data. In *ICML* (pp. 703–711).

Cho, T. S., & Shih, H. Y. (2011). Patent citation network analysis of core and emerging technologies in Taiwan: 1997–2008. *Scientometrics, 89*(3), 795.

Choi, S., Park, H., Kang, D., et al. (2012). An SAO-based text mining approach to building a technology tree for technology planning. *Expert Systems with Applications, 39*(13), 1144311455.

De Rassenfosse, G., et al. (2013). The worldwide count of priority patents: A new indicatorof inventive activity. *Research Policy, 42*(3), 720–737.

Ding, Y. (2011a). Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Journal of informetrics, 5*(1), 187–203.

Ding, Y. (2011b). Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Journal of informetrics, 5*(1), 187–203.

Furukawa, T., et al. (2015). Identifying the evolutionary process of emerging technologies: A chronological network analysis of World Wide Web conference sessions. *Technological Forecasting and Social Change, 91,* 280–294.

Guo, H., Weingart, S., & Börner, K. (2011). Mixed-indicators model for identifying emerging research areas. *Scientometrics, 89*(1), 421–435.

Heeffer, S., & Thijs, B. (2017). *Lexical analysis of scientific publications for nano-level scientometrics*. New York: Springer.

Hofer, K. M., Smejkal, A. E., Bilgin, F. Z., et al. (2010). Conference proceedings as a matter of bibliometric studies: The Academy of International Business 2006–2008. *Scientometrics, 84*(3), 845–862.

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

Kajikawa, Y., & Takeda, Y. (2008). Structure of research on biomass and bio-fuels: Acitation-based approach. *Technological Forecasting and Social Change, 75*(9), 1349–1359.

Kim, S., & Choi, J. (2014). An SVM-based high-quality article classifier for systematic reviews. *Journal of Biomedical Informatics, 47*(5), 153.

Kim, K. Y., Jeong, S. Y., Park, J. H., et al. (2018). Performance comparison of Korean keyword-based document classifiers using convolutional neural networks. *International Journal of Applied Engineering Research, 13*(4), 1879–1883.

Kong, D., Zhou, Y., Liu, Y., & Xue, L. (2017a). Using the data mining method to assess the innovation gap: A case of industrial robotics in a catching-up country. *Technological Forecasting and Social Change, 119,* 80–97.

Kong, D., Zhou, Y., Liu, Y., et al. (2017b). Using the data mining method to assess the innovation gap: A case of industrial robotics in a catching-up country. *Technological Forecasting and Social Change, 119,* 80–97.

Lee, W. H. (2008). How to identify emerging research fields using scientometrics: An example in the field of Information Security. *Scientometrics, 76*(3), 503–525.

Leydesdorff, L., & Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the Association for Information Science and Technology, 60*(2), 348–362.

Liu, Y., Zhou, Y., Liu, X., et al. (2019). Wasserstein GAN-based small-sample augmentation for new-generation artificial intelligence: A case study of cancer-staging data in biology. *Engineering, 5*(1), 156–163.

Lu, L. Y. Y., & Liu, J. S. (2016). A novel approach to identify the major research themes and development-trajectory: The case of patenting research. *Technological Forecasting and Social Change, 103,* 71–82.

Nigam, K., Mccallum, A. K., Thrun, S., et al. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning, 39*(2), 103–134.

Park, H., Yoon, J., & Kim, K. (2012). Identifying patent infringement using SAO based semantic technological similarities. *Scientometrics, 90*(2), 515–529.

Roche, I., Besagni, D., François, C., Hörlesberger, M., & Schiebel, E. (2010). Identificationand characterisation of technological topics in the field of molecular biology. *Scientometrics, 82*(3), 663–676.

Rotolo, D., Hicks, D., & Martin, B. R. (2015). What is an emerging technology? *Research Policy, 44*(10), 1827–1843.

Schiebel, E., Hörlesberger, M., Roche, I., et al. (2010). An advanced diffusion model to identify emergent research issues: The case of optoelectronic devices. *Scientometrics, 83*(3), 765781.

Shibata, N., et al. (2011). Detecting emerging research fronts in regenerative medicine by the citation network analysis of scientific publications. *Technological Forecasting and Social Change, 78*(2), 274–282.

Small, H., Boyack, K. W., & Klavans, R. (2014a). Identifying emerging topics in science andtechnology. *Research Policy, 43*(8), 1450–1467.

Small, H., Boyack, K. W., & Klavans, R. (2014b). Identifying emerging topics in science and technology. *Research Policy, 43*(8), 1450–1467.

Upham, S. P., & Small, H. (2010). Emerging research fronts in science and technology: Patterns of new knowledge development. *Scientometrics, 83*(1), 15–38.

Venugopalan, S., & Rai, V. (2015). Topic based classification and pattern identification in patents. *Technological Forecasting and Social Change, 94,* 236–250.

Waltman, L., van Eck, N. J., & Noyons, E. C. M. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics, 4*(4), 629–635.

Wang, S., & Koopman, R. (2017). Clustering articles based on semantic similarity. *Scientometrics, 111*(2), 1017–1031.

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

Wang, B., Liu, Y., Zhou, Y., et al. (2018a). Emerging nanogenerator technology in China: A reviewand forecast using integrating bibliometrics, patent analysis and technology roadmapping methods. *Nano Energy, 46,* 322–330.

Wang, Y., Urban, F., Zhou, Y., & Chen, L. (2018b). Comparing the technology trajectories of Solar PV and Solar water heaters in China: Using a Patent Lens. *Sustainability, 10,* 4166.

Watts, R. J., & Porter, A. L. (2003). R&D cluster quality measures and technology maturity. *Technological Forecasting and Social Change, 70*(8), 735–758.

Wong, K. V., & Hernandez, A. (2012). A review of additive manufacturing. *ISRN Mechanical Engineering, 2012*(2), 30–38.

Wu, F., & Weld, D. S. (2010). Open information extraction using Wikipedia. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 118–127). Association for Computational Linguistics.

Yang, Z., Tang, J., & Cohen, W. (2016). Multi-modal Bayesian embeddings for learning social knowledgegraphs. In *International joint conference on artificial intelligence* (pp. 22872293). AAAI Press.

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

Yin, J., & Wang, J. (2014). A dirichlet multinomial mixture model-based approach for short textclustering. In *ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 233–242). ACM.

Yoon, J., & Kim, K. (2011). Identifying rapidly evolving technological trends for R&D planning using SAO-based semantic patent networks. *Scientometrics, 88*(1), 213–228.

Zhang, J., Liu, X., & Wu, L. (2016). The study of subject-classification based on journal coupling andexpert subject-classification system. *Scientometrics, 107*(3), 1149–1170.

Zhang, L., Liu, X., Janssens, F., et al. (2010). Subject clustering analysis based on ISI categoryclassification. *Journal of Informetrics, 4*(2), 185–193.

Zhang, Y., et al. (2014). "Term clumping" for technical intelligence: A case study on dye-sensitized solar cells. *Technological Forecasting and Social Change, 85,* 26–39.

Zhou, Y., Li, X., Lema, R., & Urban, F. (2016). Comparing the knowledge bases of wind turbine firms in Asia and Europe: Patent trajectories, networks, and globalisation. *Science and Public Policy, 43*(4), 476–491.

Zhou, Y., Li, X., Lema, R., & Urban, F. (2019). How do low-carbon policies promote green diffusionamong alliance-based firms in China? An evolutionary-game model of complex networks. *Journal of Cleaner Production, 210,* 518–529.

Zhou, Y., Pan, M., & Urban, F. (2018). Comparing the international knowledge flow of china's wind and solar photovoltaic (pv) industries: Patent analysis and implications for sustainable development. *Sustainability, 10*(6), 1883.