Journal of Informetrics 13 (2019) 100983

Contents lists available at ScienceDirect

Journal of Informetrics

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

Regular article

Emerging research topics detection with multiple machine learning models

Shuo Xu[a], Liyuan Hao[a], Xin An[b,*], Guancan Yang[c], Feifei Wang[a]

[a] Research Base of Beijing Modern Manufacturing Development, College of Economics and Management, Beijing University of Technology, No. 100 PingLeYuan, Chaoyang District, Beijing 100124, PR China
[b] School of Economics and Management, Beijing Forestry University, No. 35 Qinghua East Road, Haidian District, Beijing 100083, PR China
[c] School of Information Resource Management, Renmin University of China, No. 59 Zhongguancun Street, Haidian District, Beijing 100872, PR China

ARTICLE INFO

Article history:
Received 13 February 2019
Received in revised form 15 October 2019
Accepted 16 October 2019

Keywords:
Emerging research topics
Topic modeling
Dynamic Influence Model
Citation Influence Model
Machine learning

ABSTRACT

Emerging research topic detection can benefit the research foundations and policy-makers. With the long-term and recent interest in detecting emerging research topics, various approaches are proposed in the literature. Though, there is still a lack of well-established linkages between the clear conceptual definition of emerging research topics and the proposed indicators for operationalization. This work follows the definition by Wang (2018), and several machine learning models are together used to detect and foresight the emerging research topics. Finally, experimental results on *gene editing* dataset discover three emerging research topics, which make clear that it is feasible to identify emerging research topics with our framework.

1. Introduction

More and more attentions are paid to detect automatically emerging research topics from the scientific literature. The research foundations and policy-makers, which aim to promote and enhance the development of potentially promising research topics, can benefit from these studies in terms of R&D policy and portfolio management, technology opportunities analysis, and management of innovation. Several recent research projects, such as Emerging Research Areas and their Coverage (Reiss, Vignola-Gagné, Kukk, Glänzel, & Thijs, 2013) supported by the European Research Council (ERC) in 2009 and Foresight and Understanding from Scientific Exposition (FUSE) (Small, Boyack, & Klavans, 2014) funded by the Intelligence Advanced Research Projects Activity (IARPA) in 2011, promote further the development of emerging research topic identification.

With the long-term and recent interest in detecting emerging research topics, various approaches (Xu, Hao, An, Pang, & Li, 2019), such as citation-based methods (Boyack, Klavans, Small, & Ungar, 2014; Chen, 2006; Glänzel & Thijs, 2012; Jarić, Knežević-Jarić, & Lenhardt, 2014; Roche, Besagni, François, Hörlesberger, & Schiebel, 2010; Soriano, Álvarez, & Valdés, 2018; Wang, 2018) and lexical-based methods (Guo, Weingart, & Börner, 2011; Hansmann & Niemeyer, 2014; Porter, Garner, Carley, & Newman, 2019; Robinson, Ruivenkamp, & Rip, 2007), are proposed. However, as Cozzens et al. (2010) pointed

* Corresponding author.
E-mail addresses: xushuo@bjut.edu.cn (S. Xu), Leanne.H@qq.com (L. Hao), anxin@bjfu.edu.cn (X. An), yanggc@ruc.edu.cn (G. Yang), feifeiwang@bjut.edu.cn (F. Wang).

https://doi.org/10.1016/j.joi.2019.100983
1751-1577/

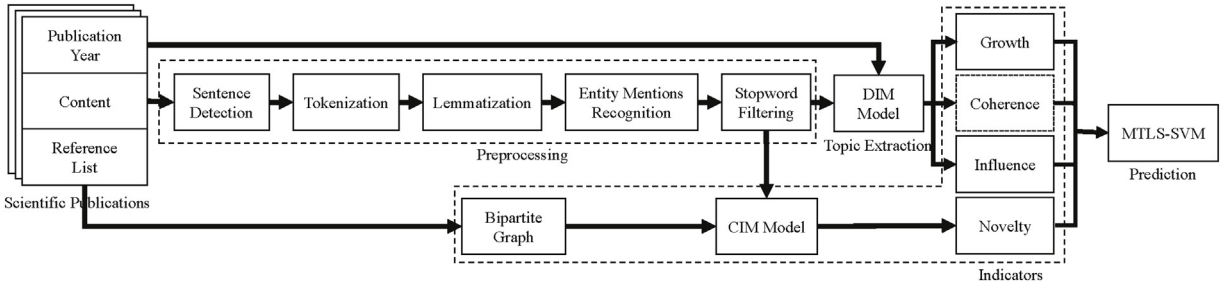

Fig. 1. Research framework of emerging topic detection.

out, most studies concentrate in fact on how to measure emerging research topics, rather than how to identify them. In our opinion, this may be related to the lack of a clear conceptual definition of emerging research topics and well-established linkages between the clear concept and the proposed indicators for operationalization, though the term *emerging research topic* or *emerging topic* is widely utilized in the literature.

Recently, Rotolo, Hicks, and Martin (2015) developed a comprehensive definition for an emerging technology and summarized five attributes to characterize the emergence of a technology: (a) radical novelty, (b) relatively fast growth, (c) coherence, (d) prominent impact, and (e) uncertainty and ambiguity. Taking into account the commonalities and differences between technologies and research topics, Wang (2018) proposed a similar definition for an emerging research topic: *A radically novel and relatively fast growing research topic characterized by a certain degree of coherence, and a considerable scientific impact*. Thus, the four attributes are attached to an emerging research topic: (a) radical novelty, (b) relatively fast growth, (c) coherence, and (d) scientific impact. This work follows this definition.

The concept of *growth* implies increase over time, and is relatively easy to measure (Ohniwa, Hibino, & Takeyasu, 2010; Porter et al., 2019; Wang, 2018). The *novelty* trait involves qualities of being original or new. When discussing emergence, this trait was mentioned by many scholars (An & Wu, 2011; Jarić et al., 2014; Porter et al., 2019; Tu & Seng, 2012; Wang, 2018). In fact, only newness but not innovation is emphasized in the literature. This is against our intuition, since the *growth* per se also implies the newness to some extent. As for *coherence*, it depends on whether the topic extraction method can ensure that the extracted topics are sufficiently coherent. It is well known that citation counts are very powerful, but it is very hard to measure the scientific impact of an emerging topic with them. In our view, main reasons are two-fold: (a) not all citations from an article are created equally, and hence they should not be counted with equal weight (Zhu, Turney, Lemire, & Vellino, 2014); (b) recent articles usually do not receive enough citations or no citation at all. Therefore, in order to measure the scientific impact of a research topic, an alternative approach should be taken.

In this article, several machine learning models (cf. Fig. 1) are together used to detect and foresight the emerging research topics. More specifically, thematic structures are extracted firstly with Dynamic Influence Model (DIM) (Gerrish & Blei, 2010) from the scientific publications. Then, the growth, coherence, influence and novelty indicators are calculated, in which the first three indicators are based on the DIM model and the last one is based on the Citation Influence Model (CIM) (Dietz, Bickel, & Scheffer, 2007). After that, the resulting indicators for the following two years are predicted with Multi-Task Least-Squares Support Vector Machine (MTLS-SVM) (Xu, An, Qiao, & Zhu, 2014) due to underling unknown relatedness amongst different indicators (Rotolo et al., 2015) and different topics. Finally, experimental results on *gene editing* dataset make clear that it is feasible to detect emerging research topics with multiple machine learning models.

2. Related work

Before delving into more specifies, discussion of the literature pertinent to emerging research topic detection is in order. For more elaborate and detailed surveys we refer the readers to (Xu, Hao, An, Pang, et al., 2019).

2.1. Bibliometric indicators for identifying emerging research topics

Many bibliometric indicators have been put foreword in the literature to identify emerging research topics, but there is no consensus among indicators. Please refer to Table 1 and the references thereafter for more details.

This study concentrates on four key attributes: (a) radical novelty, (b) relatively fast growth, (c) coherence, and (d) scientific impact. Among these indicators, radical novelty and relatively fast growth are related to time slices. The novelty is often measured by relative age of references (Jarić et al., 2014) or average publication time of scholarly articles (Tu & Seng, 2012). The growth concept can be operationalized by a variety of ways, such as the growth ratio of publications (Wang, 2018), the amount of terms (Porter et al., 2019) or the number of new authors that are attracted to a specific research topic (Guo et al., 2011). It is not difficult to see that only newness but not innovation is emphasized by Jarić et al. (2014) and Tu and Seng (2012). In the meanwhile, the growth per se also implies the newness to some extent. For instance, high growth ratio of publications usually signifies more new scientific publications. Hence, to reduce information redundancy from the growth and novelty indicators, the novelty indicator should put more weight on the innovation dimension.

Table 1
The indicators used for detecting emerging research topics.

Authors	Indicators
Jarić et al. (2014)	Novelty
Tu and Seng (2012)	Novelty, Topic's hotness
An and Wu (2011)	Novelty
Schultz and Joutz (2010)	The potential for economic impact
Ohniwa et al. (2010)	Growth, The abundance of emerging keywords
González-Alcaide, Llorente, and Ramos (2016)	Scientific activity, Size and stability of the research community, Age of the bibliographic references cited
Guo et al. (2011)	Topical direction, Growth, Diversity of base knowledge
Porter et al. (2019)	Novelty, Growth, Community, Persistence
Wang (2018)	Radical novelty, Relatively fast growth, Coherence, Scientific impact

For a research topic to be considered as emerging, its meaning should not drift from a time slice to another slice. In other words, an emerging research topic should be significant coherent. Wang (2018) measured the coherence by citations received from the publications within a given topic cluster. Boyack et al. (2011) quantified the textual coherence of each document cluster with the Jensen-Shannon divergence (JSD) (Lin, 1991). Actually, if the topic extraction approach can ensure that the extracted themes are sufficiently coherent, like the DIM model used in this study, the coherence indicator can be omitted.

An emerging technology has the potential to exert a considerable impact on the socio-economic domains, which is usually measured on the basis of patent data, such as generality indicator (Schultz & Joutz, 2010). Accordingly, a research topic should have prominent impact on future scientific development (Wang, 2018). Though citation counts are very powerful, as noted in Section 1, it is very hard to measure the scientific impact of an emerging topic based on them. Therefore, an alternative approach based on the CIM model is taken in this work.

2.2. Method of detecting emerging research topics

Last fifteen years witnessed significant progress in the field of the emerging research topics detection ever since Morris, Yen, Wu, and Asnake (2003). Several major technologies have been developed such as citation-link approaches and lexical approaches.

The citation-link methods can be further divided into three sub-groups: (a) direct citation network analysis (Shibata, Kajikawa, Takeda, & Matsushima, 2008; Waltman & van Eck, 2012), (b) co-citation network analysis (Small & Griffith, 1974), and (c) bibliographic coupling network analysis (Glänzel & Thijs, 2012; Huang & Chang, 2014). Direct citation network analysis is often used as the primary approach, since it shows the best performance in detecting emerging research topics in the early stage. Boyack and Klavans (2010) compared these three approaches, and found that the accuracy of bibliographic coupling network analysis outperformed the others. Apart from the citation-link networks based on publications, the counterparts based on authors are also utilized to detect the emerging research topics (Ma, 2012; Zhao & Strotmann, 2008).

Another study stream is lexical approaches. An and Wu (2011) and Hu, Hu, Deng, and Liu (2013) analyzed the co-word network to reveal the status and trends of steam cell field and library information & system, respectively. This article argues that the shortcomings for co-word network analysis are reflected in two aspects: (1) the term's meaning may drift with time, especially for emerging interdisciplinary fields; (2) high-frequent terms are often used in the co-word network analysis, ignoring low-frequent potential terms. To overcome these shortcomings, burst term detection method was developed by Kleinberg (2002). Then, Guo et al. (2011) armed this method with three indicators: sudden increases in the frequency of specific words, the number and speed by which new authors are attracted to an emerging research area, and changes in the inter-disciplinarity of cited references. In addition, the natural language processing technology was employed by Hansmann and Niemeyer (2014), Yan (2014) and Liu, Yin, Liu, and Dunford (2015) to excavate and analyze the topics, keywords or controlled terms in scientific and patent literature, so as to identify and explore emerging research topics.

In summary, to the best of our knowledge, hybrid approaches that combine citation-link and lexical techniques are under-explored for emerging research topics detection in the literature. The DIM model is utilized in this article to extract research topics from textual contents and the CIM model is used for measuring the novelty indicator. Thus, our framework can benefit simultaneously from these two-type approaches.

3. Research framework and methodology

As shown in Fig. 1, our research framework consists of four phases. The first is to detect sentence boundaries, tokenize and lemmatize each detected sentence, recognize entity mentions, and then filter stopwords as preprocessing. At the second phase, research topics are extracted with the Dynamic Influence Model (DIM) (Gerrish & Blei, 2010) from the scientific publications. The third phase is to calculate the growth, coherence, influence and novelty indicators, in which the first three indicators are based on the DIM model and the last one is based on the Citation Influence Model (CIM) (Dietz et al., 2007).

Table 2
Notations used in the DIM model.

Symbol	Description
K	Number of word topics
$M^{(t)}$	Number of publications with time t
$N_m^{(t)}$	Number of word tokens in the publication m with time t
$\vec{\vartheta}_m^{(t)}$	Multinomial distribution of topics specific to the publication m with time t
$\vec{\varphi}_k^{(t)}$	Natural parameters of words specific to the topic k and time t
$\ell_{m,k}^{(t)}$	Influence of the publication m specific to the topic k and time t
$z_{m,n}^{(t)}$	Topic associated with the n-th word token in the publication m with time t
$w_{m,n}^{(t)}$	nth word token in the publication m with time t
$\vec{\alpha}$	Hyper-parameter

Finally, the resulting indicators for the following two years are predicted with Multi-Task Least-Squares Support Vector Machine (MTLS-SVM) (Xu, An, et al., 2014). The last three phases are described in more detail in the following subsections.

3.1. Topic extraction

In order to detect emerging research topics, the dynamics of the underlying topics should be modelled explicitly. Several topic models are proposed in the literature, such as Dynamic Topic Model (DTM) (Blei & Lafferty, 2006), continuous time DTM (cDTM) (Wang, Blei, & Heckerman, 2008), Multiple timescales DTM (MDTM) (Iwata, Yamada, Sakurai, & Ueda, 2012), Dynamic Influential Model (DIM) (Gerrish & Blei, 2010), Topic over Time (ToT) (Wang & McCallum, 2006), Trend Detection Model (TDM) (Kawamae & Higashinaka, 2010) and so on. Here, DIM model is used, since it cannot only capture the evolution of topics in a sequentially organized corpus of documents, but it also can estimate a meaningful influence measure of each scholarly article. Under this model, scientific publications are grouped by year, and each year's articles are assumed to be a mixture of a set of research topics that have evolved from the last year's topics.

Intuitively, an influential scholarly article will affect how future articles are written. Thus, one should be able to discover this effect by examining the way corpus statistics change over time. This intuition is encoded in the DIM model (Gerrish & Blei, 2010), which allows for multiple threads of influence (one thread per research topic). Though DIM model aims to capture something different from citation counts, the estimated influence measure of each scholarly article has a positive correlation with the number of citations in term of Spearman rank correlation (Gerrish & Blei, 2010).

The notations is summarized in Table 2, and the graphical model representation of the DIM model is shown in Fig. 2. Note that the probability of a word v specific to the research topic k is given by $\phi_{k,v}^{(t)} = \Pr(v|\vec{\varphi}_k^{(t)}) \propto \exp(\varphi_{k,v}^{(t)})$, and let $\vec{\phi}_k^{(t)} = [\phi_{k,1}^{(t)}, \ldots, \phi_{k,v}^{(t)}, \ldots, \phi_{k,V}^{(t)}]$. Following our intuition, the more influential a publication is, the more its words should *nudge* the topic's natural parameters at the next time slice. In the DIM model, this point is formally expressed as follows (Gerrish & Blei, 2010), where $\mathcal{N}(\cdot, \cdot)$ denotes a normal distribution. It is this point that makes sure the discovered research topics to be well coherent.

$$\begin{aligned} & \vec{\varphi}_k^{(t+1)} | \vec{\varphi}_k^{(t)}, \vec{w}^{(t)}, \vec{z}^{(t)}, \vec{\ell}_k^{(t)} \\ \sim \; & \mathcal{N}(\vec{\varphi}_k^{(t)} + \exp(-\vec{\varphi}_k^{(t)}) \sum_{m=1}^{M^{(t)}} \ell_{m,k}^{(t)} \sum_{n:z_{m,n}^{(t)}=k} w_{m,n}^{(t)}, \sigma^2 \mathbf{I}) \end{aligned} \tag{1}$$

As for many Bayesian models (An, Xu, Wen, & Hu, 2014; Blei, Ng, & Jordan, 2003; Xu, Shi, et al., 2014; Xu, Zhai, et al., 2019), posterior inference cannot be done exactly in this model. A variety of algorithms have been developed in the literature, such as variational inference (Jordan, Grhahramani, Jaakkola, & Saul, 1999), Markov chain Monte Carlo (MCMC) (Andrieu, de Freitas, Doucet, & Jordan, 2003), and stochastic variational inference (Hoffman, Blei, Wang, & Paisley, 2013; Xu et al., 2018). The variational inference was originally utilized by Gerrish and Blei (2010) to approximate the posterior of the DIM model. Please refer to Gerrish and Blei (2010) for more details. It is worth mentioning that to void setting the number of topics by user, a supervised version of the DIM model is raised by Jiang, Liu, and Gao (2015), in which the author-provided keywords serve as the supervised labels. Obviously, the number of unique keywords as the number of topics is too huge for emerging research topics detection.

3.2. Indicator calculation

A research topic can be considered as emerging if it meets the characteristics of relatively fast growth, coherence, scientific impact, and radical novelty (Wang, 2018). The operationalization for each indicator is delineated in the following subsections.

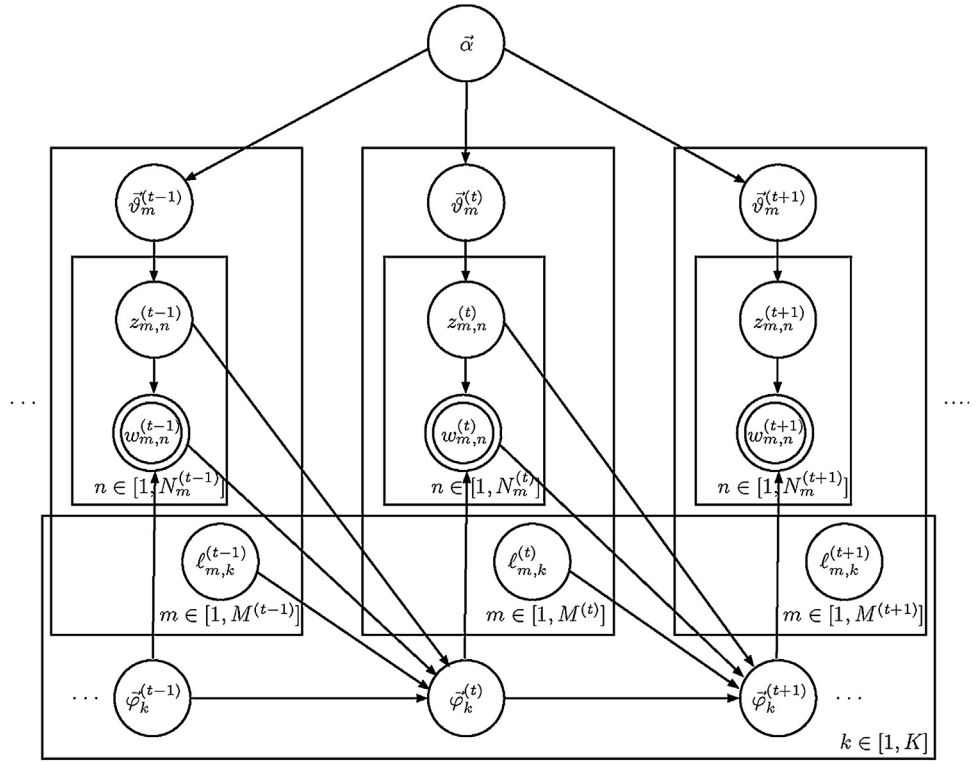

Fig. 2. The graphical model representation of the DIM model.

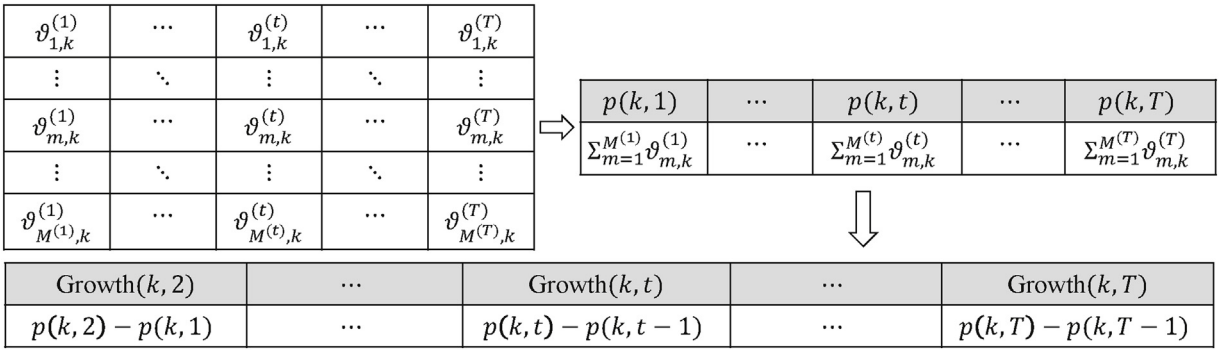

Fig. 3. Illustration to calculate the growth indicator.

3.2.1. *Relatively fast growth*

The popularity (Chen, Tsutsui, Ding, & Ma, 2017) for the research topic k at the time t is $p_{k,t} = \sum_{m=1}^{M^{(t)}} \vartheta_{m,k}^{(t)}$. Here, the growth is defined as the slope of the popularity, i.e., Growth$(k, t) = p_{k,t} - p_{k,t-1}$. In more details, according to the DIM model (Gerrish & Blei, 2010), each publication m with the time t is presented with a multinomial probabilistic distribution over the research topics $\vec{\vartheta}_m^{(t)}$. From this distribution, the popularity and then the growth can be easily calculated, as illustrated in Fig. 3 for the topic k. For a research topic k to be considered as emerging, the value for Growth(k, t) in recent years should be higher than the average $\frac{1}{K}\sum_k$Growth(k, t).

3.2.2. *Coherence*

The coherence can be measured with the symmetrized Kullback-Leibler divergence (symKLD) (Kullback & Leibler, 1951) or Jensen-Shannon divergence (JSD) (Lin, 1991), both of which compute the distance between two probability distributions. The JSD had been used by Boyack et al. (2011) to measure the textual coherence of each document cluster with the satisfactory

Table 3
Notation used in the CIM model.

Symbol	Description
K	Number of topics
$M, \tilde{M}$	Number of the citing and cited publications, respectively
V	Number of unique words in the citing and cited publications
$N_m, \tilde{N}_{\tilde{m}}$	Number of word tokens in the citing publication m and cited publication $\tilde{m}$, respectively
L_m	Number of documents cited by the citing publication m
$\vec{\vartheta}_m, \vec{\tilde{\vartheta}}_{\tilde{m}}$	Multinomial distribution of topics specific to the citing publication m and cited publication $\tilde{m}$, respectively
$\vec{\psi}_m$	Multinomial distribution of references specific to the citing publication m
$\vec{\lambda}_m$	Bernoulli distribution of status specific to the citing publication m
$\vec{\varphi}_k$	Multinomial distribution of words specific to the topic k
$z_{m,n}, \tilde{z}_{\tilde{m},\tilde{n}}$	Topic associated with the nth token in the citing publication m, and with the $\tilde{n}$th token in the cited publication $\tilde{m}$, respectively
$c_{m,n}$	Cited publication associated with the n-th token in the citing publication m
$s_{m,n}$	Status associated with the nth token in the citing publication m
$w_{m,n}, \tilde{w}_{\tilde{m},\tilde{n}}$	nth token in the citing publication m, and $\tilde{n}$th token in the cited publication $\tilde{m}$, respectively
$\vec{\alpha}, \vec{\beta}, \vec{\delta}_m, \vec{\mu}$	hyperparameters

performance. More specifically, the coherence for a research topic k at the time t can be defined formally on the basis of the symKLD and JSD as follows.

$$\text{Coherence}_{\text{synKLD}}(k,t) = \frac{1}{2}[\text{KLD}(\vec{\phi}_k^{(t-1)}, \vec{\phi}_k^{(t)}) + \text{KLD}(\vec{\phi}_k^{(t)}, \vec{\phi}_k^{(t-1)})] \tag{2}$$

$$\text{Coherence}_{\text{JSD}}(k,t) = \frac{1}{2}[\text{KLD}(\vec{\phi}_k^{(t-1)}, \vec{\rho}_k^{(t)}) + \text{KLD}(\vec{\phi}_k^{(t)}, \vec{\rho}_k^{(t)})] \tag{3}$$

Here $\text{KLD}(\vec{x}, \vec{y}) = \sum_v x_v \log \frac{x_v}{y_v}$ and $\vec{\rho}_k^{(t)} = \frac{1}{2}(\vec{\phi}_k^{(t-1)} + \vec{\phi}_k^{(t)})$.

Whether the symKLD or JDS is a divergence measure, meaning that if the words with high probability value in a research topic at the time t are very different from those in the same topic at the time $t-1$, the resulting symKLD and JSD values will be very high. It indicates that the meaning of this topic drifted from the time $t-1$ to t. The research topics with similar sets of words across two consecutive time slices will have a lower divergence. Therefore, for a research topic k to be considered as emerging, the value for $\text{Coherence}_{\text{symKLD}}(k,t)$ and $\text{Coherence}_{\text{JSD}}(k,t)$ in recent several consecutive years should be very close to zero.

3.2.3. Scientific influence

In the DIM model (Gerrish & Blei, 2010), the influence of the publication m over the research topic k at the time t is denoted by $\ell_{m,k}^{(t)}$, and solved by a linear regression. That is to say, this model can estimate effectively the scientific influence of each publication over each research topic at each time slice. Thus, similar to the popularity (Chen et al., 2017), the scientific influence for the research topic k at the time t is defined in this work as $\text{Influence}(k,t) = \sum_{m=1}^{M^{(t)}} \ell_{m,k}^{(t)}$. For a research topic k to be considered as emerging, the value for $\text{Influence}(k,t)$ in recent years should be higher than the average $\frac{1}{K}\sum_k \text{Influence}(k,t)$.

3.2.4. Radical novelty

Among these four indicators, the radical novelty is the most difficult to quantify. According to our understanding, if a research topic with important scientific influence is considered to be emerging, it should be highly innovative. For purpose of quantifying the radical novelty from this viewpoint, another topic model, CIM model (Dietz et al., 2007), is resorted in this article. In this model, the topical innovation and topical inheritance is incorporated via citations.

The notation is summarized in Table 3, and the graphical model representation of the CIM model is shown in Fig. 4. One can also describe the CIM model from the viewpoint of generative process as follows.

1. For each topic $k \in [1, K]$, draw $\vec{\varphi}_k \sim \text{Dir}(\vec{\beta})$;
2. For each cited document $\tilde{m} \in [1, \tilde{M}]$, draw $\vec{\tilde{\vartheta}}_{\tilde{m}} \sim \text{Dir}(\vec{\alpha})$;
3. For each cited publication $\tilde{m} \in [1, \tilde{M}]$ and each token $\tilde{n} \in [1, \tilde{N}_{\tilde{m}}]$, draw $\tilde{z}_{m,n} \sim \text{Mult}(\vec{\tilde{\vartheta}}_{\tilde{m}})$ and then $\tilde{w}_{\tilde{m},\tilde{n}} \sim \text{Mult}(\vec{\varphi}_{\tilde{z}_{\tilde{m},\tilde{n}}})$, respectively;
4. For each citing publication $m \in [1, M]$, draw $\vec{\psi}_m \sim \text{Dir}(\vec{\delta}_m)$, $\vec{\theta}_m \sim \text{Dir}(\vec{\alpha})$, and $\vec{\lambda}_m \sim \text{Beta}(\vec{\mu})$, respectively;
5. For each citing publication $m \in [1, M]$ and each token $n \in [1, N_m]$, draw $s_{m,n} \sim \text{Bern}(\vec{\lambda})$. If $s_{m,n} = 0$, draw $c_{m,n} \sim \text{Mult}(\vec{\psi}_m)$ and then $z_{m,n} \sim \text{Mult}(\vec{\tilde{\vartheta}}_{c_{m,n}})$. Otherwise, draw $z_{m,n} \sim \text{Mult}(\vec{\vartheta}_m)$. Finally, draw $w_{m,n} \sim \text{Mult}(\vec{\varphi}_{z_{m,n}})$.

From this generative process, one can see that the topic $z_{m,n}$ for a word token $w_{m,n}$ in the citing publication m can be chosen to dram from the topic mixture $\vec{\tilde{\vartheta}}_{c_{m,n}}$ of a cited publication $c_{m,n}$ or from its own topic mixture $\vec{\vartheta}_m$ that models innovative aspects. The parameters $\lambda_{m,0}$ and $\lambda_{m,1}$ control the extent from respective topic mixture with the constraint $\lambda_{m,0} + \lambda_{m,1} = 1$. Hence, by combining $\vec{\vartheta}_\cdot^{(t)}$ in the DIM model, the radical novelty for the research topic k at the time t is defined

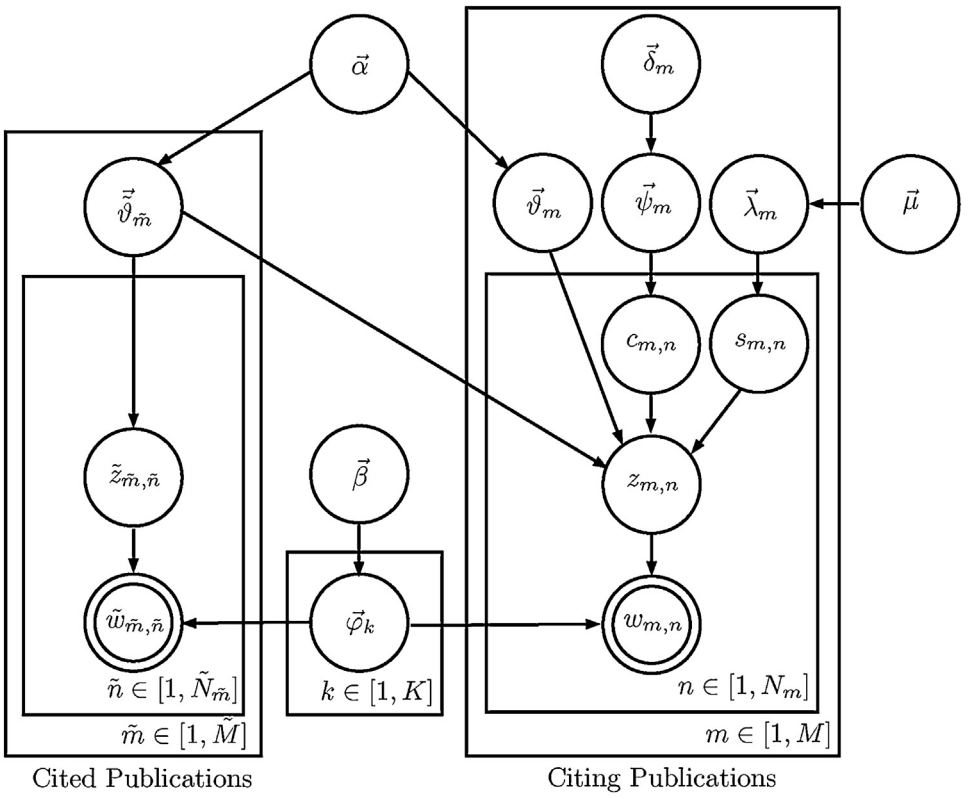

Fig. 4. The graphical model representation of the CIM model.

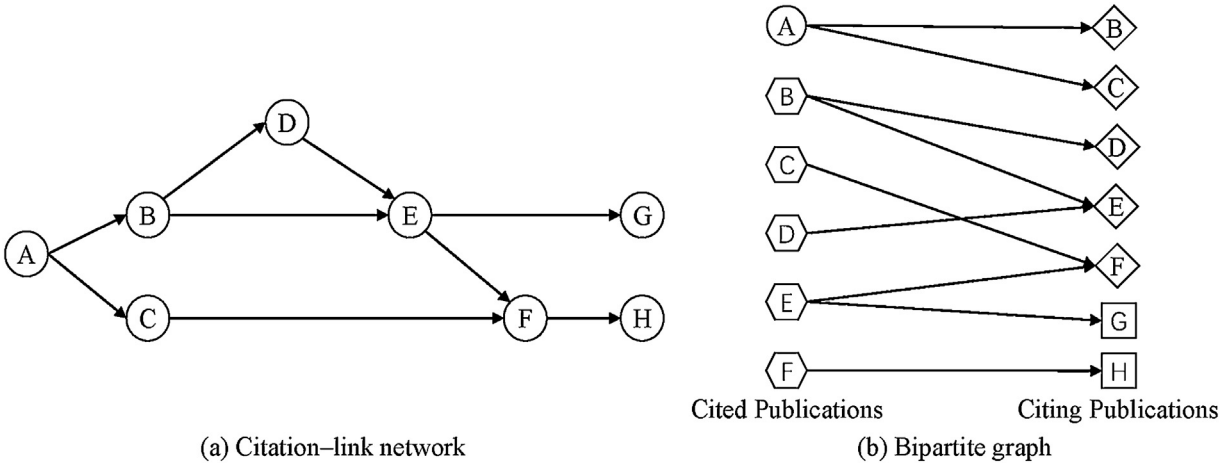

Fig. 5. An example of the citation-link network and the resulting bipartite graph.

as Novelty$(k, t) = \frac{1}{M^{(t)}} \sum_{m=1}^{M^{(t)}} [\vartheta_{m,k}^{(t)} \times \lambda_{m,1}]$. For a research topic k to be considered as emerging, the value for Novelty(k, t) in recent years should be higher than the average $\frac{1}{K} \sum_k$Novelty(k, t).

It is worth noting that the CIM model (Dietz et al., 2007) assumes that each citing publication is only influenced by the cited publications from its reference list, but not their ancestors. That is to say, the first-order Markov assumption is adopted in the model. In this way, any citation-link network can be transformed into a bipartite graph. Let's take Fig. 5(a) as an example, in which the arrow means that the publication at the arrow tail is cited by that at the arrow head. In the transformed bipartite graph, the cited and citing publication node sets consist of the publication with outgoing citation links (circle nodes in Fig. 5(b)) and those with incoming citation links (square nodes in Fig. 5(b)), respectively. The publications with incoming and outgoing links in Fig. 5(a) are two different nodes (hexagon and rhombus nodes) in Fig. 5(b).

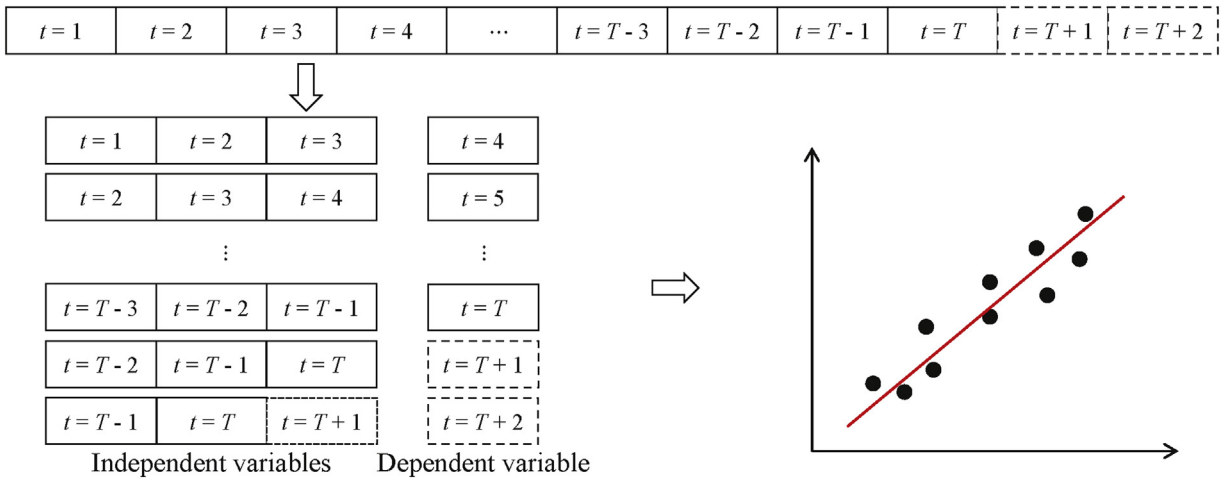

Fig. 6. Illustration of predicting the following two years' values of some indicator based on the moving-window time-series analysis ($\Delta t = 3$).

Again, posterior inference for this model cannot be done exactly. The collapse Gibbs sampling, a special case of MCMC, was originally used by Dietz et al. (2007) to approximate the posterior of the CIM model, but with incorrect posterior distributions. Here, the posterior distributions are re-derived in this article. Please refer to Appendix for more details. In this work, the number of topics for the CIM model is fixed to 100, and the symmetric Dirichlet/Beta priors α, δ, and β, δ are set at 0.5 and 0.01, respectively. The Gibbs sampling is running for 2000 iterations.

3.3. Prediction

Previous studies mainly focus on retrospectively analyzing emerging research topics of a specific field, but the nomination of emerging topics for consideration by decision makers remains largely under-studied (but see Porter et al. (2019), Small et al. (2014)). Main reason may be that the nomination of emerging topics involves a very difficult task: prediction on research topic trends for the subsequent two years.

One conventional practice for this issue is illustrated in Fig. 6. Let's suppose that the values of some indicators at $t = 1, 2, \ldots, T$ for a research topic have been calculated from the history data, and one wants to know the values for the following two years (i.e., $t = T+1, T+2$). For convenience, let the moving-window size $\Delta t = 3$. If the indicator's values covered by the moving-window are viewed as independent variables and the next value as dependent variable, one can fit a curve between independent variables and dependent variable in low-dimensional input space or high-dimensional feature space with the regression technique. Thus, it is very easy to estimate the values for the following two years with the fitted curve.

However, according to our preliminary experimental results, the prediction performance is unsatisfactory. To our point of view, the main reason is that for a research topic, the number of training instances ($T - \Delta t$) is not enough to learn a stable regression model. On closer examination, two phenomena can be found as follows: (a) there are underlying (potentially non-linear) cross relatedness amongst the interested indicators (cf. Fig. 2 in Rotolo et al. (2015)); (b) the discovered research topics from the DIM model are not completely independent with each other. This is in line with the characteristics of the problems that the multi-tasking learning technique (Caruana, 1997) is able to handle.

Therefore, the prediction problem for four indicators of all research topics can be regarded as a multi-task regression learning problem, in which the prediction for each indicator of a research topic is viewed as a task. As the state-of-the-art multi-task regression learning method, the Multi-Task Least-Squares Support Vector Machine (MTLS-SVM) (Xu, An, et al., 2014) considers the commonality of all tasks, but also the specialty of each task. Furthermore, compared to single-task (traditional) regression method, the performance indeed improves when the tasks are related (An, Wen, Zhang, & Xu, 2019; Argyriou, Evgeniou, & Pontil, 2008; Ben-David & Borbely, 2008; Chapelle et al., 2010; Xu, An, Qiao, Zhu, & Li, 2013).

Here, the RBF (radial basis function) kernel is adopted. Similar to Xu, An, et al. (2014), the grid search is used to identify proper parameters with LOO (leave-one-out) procedure. Once the optimal value of some parameters lies at the border of the search space, the search space for the parameter is increased by the multiplicative step ($2^{\pm 2}$). For more elaborate and detailed description on the MTLS-SVM and how to optimize the parameters with the grid search, we refer the readers to Xu, An, et al. (2014).

4. Experimental results & discussions

4.1. Dataset

The bibliographic data in the *gene editing* field is collected from the Web of Science core collection at the library of Beijing University of Technology. The following search strategy is utilized in this work: "TS = (gene edit*) OR TS = (crispr) OR

Table 4
Distribution of number of publications over year for *gene editing* dataset.

Pub. Year	No. of pub.	Pub. Year	No. of pub.
2000	448	2009	320
2001	442	2010	378
2002	409	2011	455
2003	446	2012	588
2004	245	2013	747
2005	221	2014	1093
2006	208	2015	1569
2007	235	2016	2625
2008	265	2017	3464
Σ Pub.	14,158	Σ Ref.	346,903

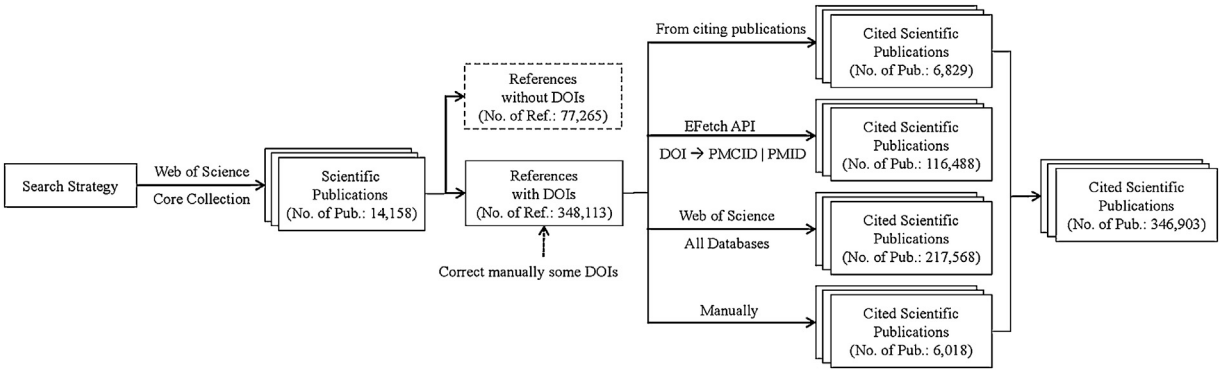

Fig. 7. Procedure for constructing *Gene Editing* dataset.

TS = (clustered regularly interspaced short palindromic repeats)". The language is limited to English, and the document type includes *article*, *proceedings paper* and *review*. The publication year spans from 2000 to 2017. The number of publications is 14,158, and Table 4 reports the distribution of number of publications over year.

In order to calculate the radical novelty indicator, the content information, such as *title* and *abstract*, for each reference is needed by the CIM model, but these information is unavailable in the reference list from Web of Science. Therefore, the cited scientific publications are separately gathered by the procedure in Fig. 7, before which the DOIs (Data Object Identifiers) of cited references are further cleaned with a method in Xu, Hao, An, Zhai, and Pang (2019). In more details, according to whether the reference attaches a DOI, the references are divided into two categories: the references with DOIs and those without DOIs. Due to the difficulty and workload of uniquely identifying each reference, the references without DOIs are directly discarded in this work.

The references with DOIs are further grouped into four categories: (a) coming directly from citing publications (6829 cited references); (b) fetching from PubMed database with EFetch API[1] after mapping DOI to PMCID or PMID (116,488 cited references); (c) fetching from Web of Science all databases (217,568 cited references); (d) collecting manually (6018 cited references). In total, the number of cited scientific publications is 346,903.

This study detects the sentences in the titles and abstracts with *geniass* (Sætre et al., 2007), tokenizes and lemmatizes the splitted sentences with *geniatagger* (Tsuruoka et al., 2005). As for stopword filtering, an English stopword list from NLTK (Natural Language Toolkit)[2] is used here, but expanded with some punctuation (such as @, % and so on). All numbers in the citing and cited publications are replaced with a special word *NUMBER*. In addition, the scholarly articles often contain many entity mentions, such as protein, DNA, RNA, cell line, cell type and so on. In order to reduce the size of word vocabulary and improve the performance, these entity mentions are recognized with *geniatagger* (Tsuruoka et al., 2005), and then excluded from further analysis.

4.2. Number of research topics

The software *dtm*[3] implements the variational inference algorithm for the DIM model. Hence, it is utilized here to extract research topics. The resulting options are listed in Table 5. For the sake of identifying a proper number of topics, the perplexity, originally used in language modeling (Azzonpardi, Girolami, & van Risjbergen, 2003), is calculated for each candidate value

[1] https://www.ncbi.nlm.nih.gov/books/NBK25499/#chapter4.EFetch.
[2] http://www.nltk.org.
[3] https://github.com/blei-lab/dtm.

Table 5
The option setting for the DIM model.

Option	Value	Option	Value
initialize_lda	True	*sigma_l*	0.0001
top_chain_var	0.005	*alpha*	0.01
model	Fixed	*lda_sequence_min_iter*	50
time_resolution	1	*lda_sequence_max_iter*	100
influence_flat_years	5	*lda_max_em_iter*	100
top_obs_var	0.5	*sigma_d*	0.0001

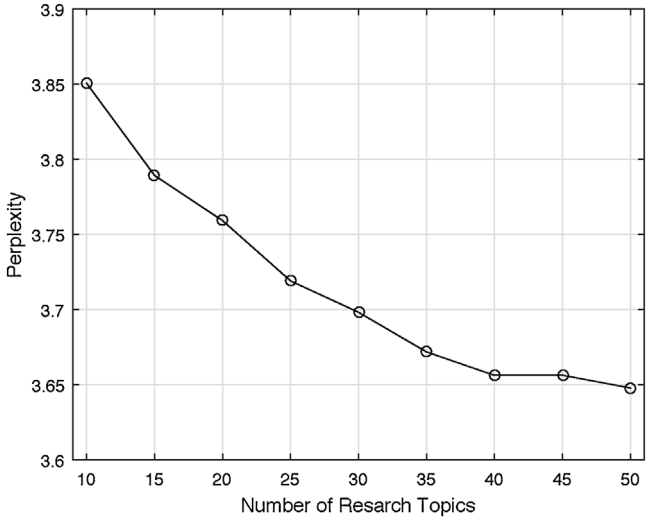

Fig. 8. The perplexity with different number of research topics.

Table 6
The prediction performance of the MTLS-SVM and LS-SVM in term of the explained variance (EV).

	LS-SVM	MTLS-SVM
Relatively fast growth	28.12	52.42
Coherence (symmetrized KL divergence)	35.75	67.89
Coherence (JS divergence)	34.35	65.64
Scientific influence	25.56	53.86
Radical novelty	27.87	57.38
Average	30.33	59.44

from 10 to 50 with a step size 5. As a standard measure for model selection, this measure is defined as the exponential of the negative normalized predictive likelihood under the model, and a lower value indicates a better modeling performance. Fig. 8 depicts the perplexity with different number of research topics. From Figure 8, one can see that the perplexity of the DIM model converges when the number of topics is 40, so the number of topics K is fixed to 40 in this work.

Since some of the research topics can represent insignificant themes, or just a collection of irrelevant words (AlSumait, Barbará, Gentle, & Domeniconi, 2009), all estimated topics from the DIM model are manually examined one by one. Four topics (id: #3, #9, #11 and #38) are directly discarded in this work. To say it in another way, in the end, 36 research topics are further analyzed to see if the emerging attributes are met.

4.3. Emerging research topics

Following Section 3.2, relatively fast growth, coherence, scientific influence and radical novelty indicators can be easily calculated from the results of the DIM and CIM model, and reported in Figs. 9–13, respectively. In Figs. 9–13, the black solid line with cross markers corresponds to the resulting average indicator. The values for the recent two years are predicted with MTLS-SVM in Section 3.3.

To illustrate the performance advantages of MTLS-SVM over traditional single task learning methods, such as LS-SVM (Suykens, Van Gestel, De Brabanter, De Moor, & Vandewalle, 2002), the explained variance (EV) (Bakker & Heskes, 2003) for the recent two years are calculated, as reported in Table 6. The EV is defined to be the total variance of the data minus the sum-squared error on the following two years as a percentage of the total data variance, which is a percentage version of the standard R^2 error measure for regression. A higher value indicates a better prediction performance. On average, LS-SVM and

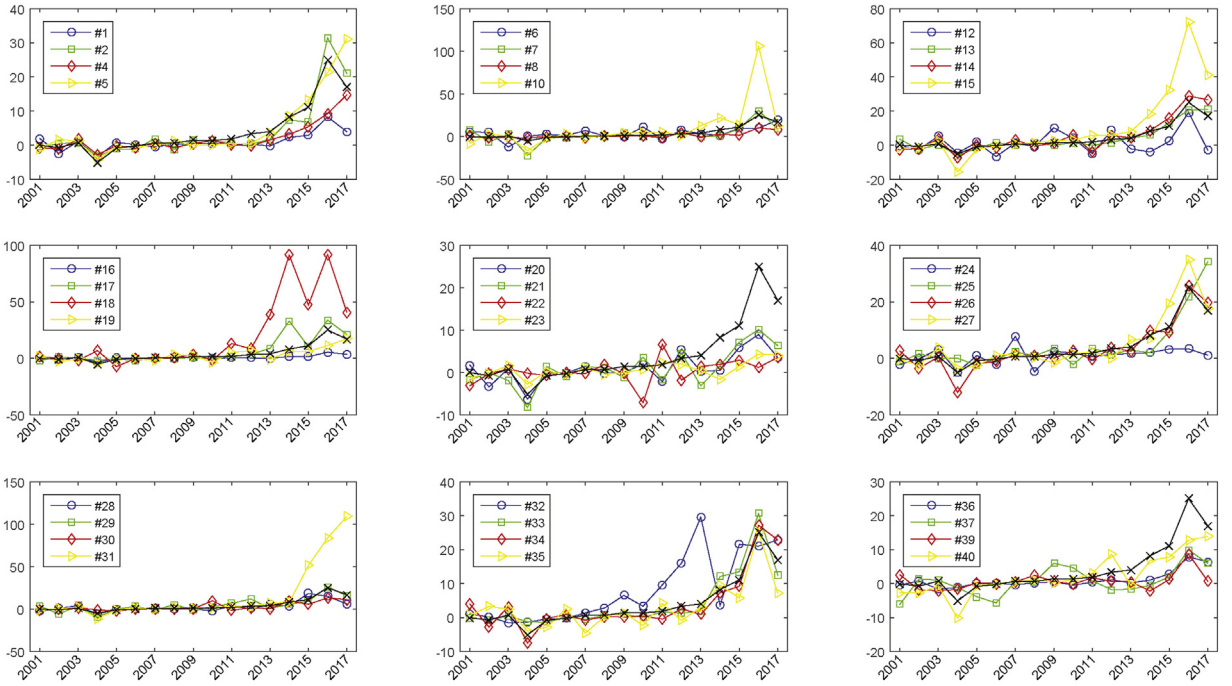

Fig. 9. The trend for the growth indicator.

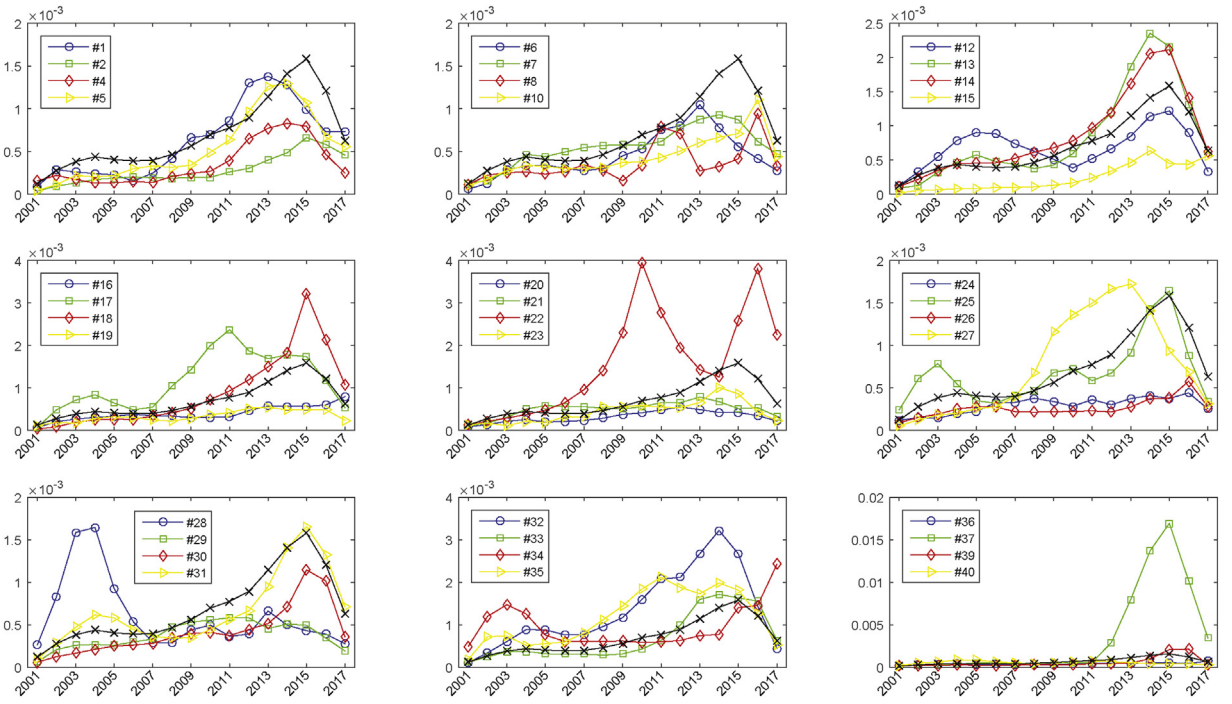

Fig. 10. The trend for the coherence indicator on the basis of symmetrized KL divergence.

MTLS-SVM explain 30.33% and 59.44% of the variance, respectively. This indicates obvious advantage of learning all tasks simultaneously over learning them one by one.

Compared to those for other indicators, the values for the coherence indicator based on whether symmetrized KL or JS divergence are very small. This indicates that the research topics from the DIM model are sufficiently coherent per se. Therefore, the coherence indicator is not adopted at all in this article as a criteria for emerging research topics. Furthermore,

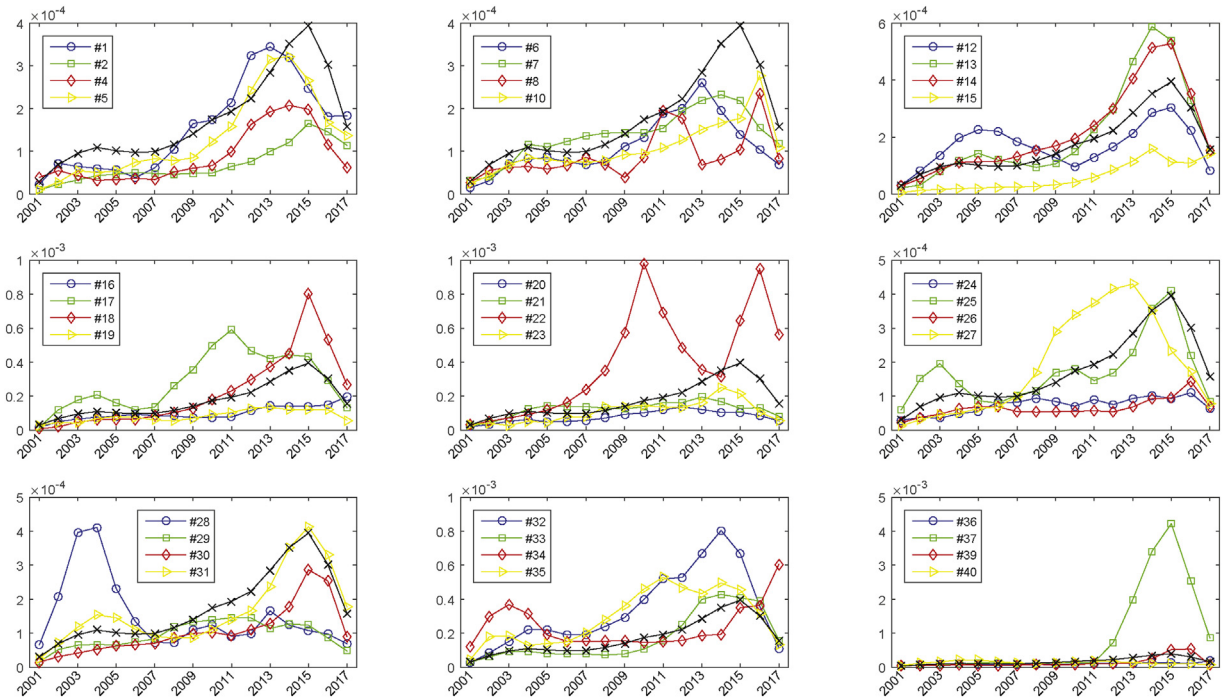

Fig. 11. The trend for the coherence indicator on the basis of JS divergence.

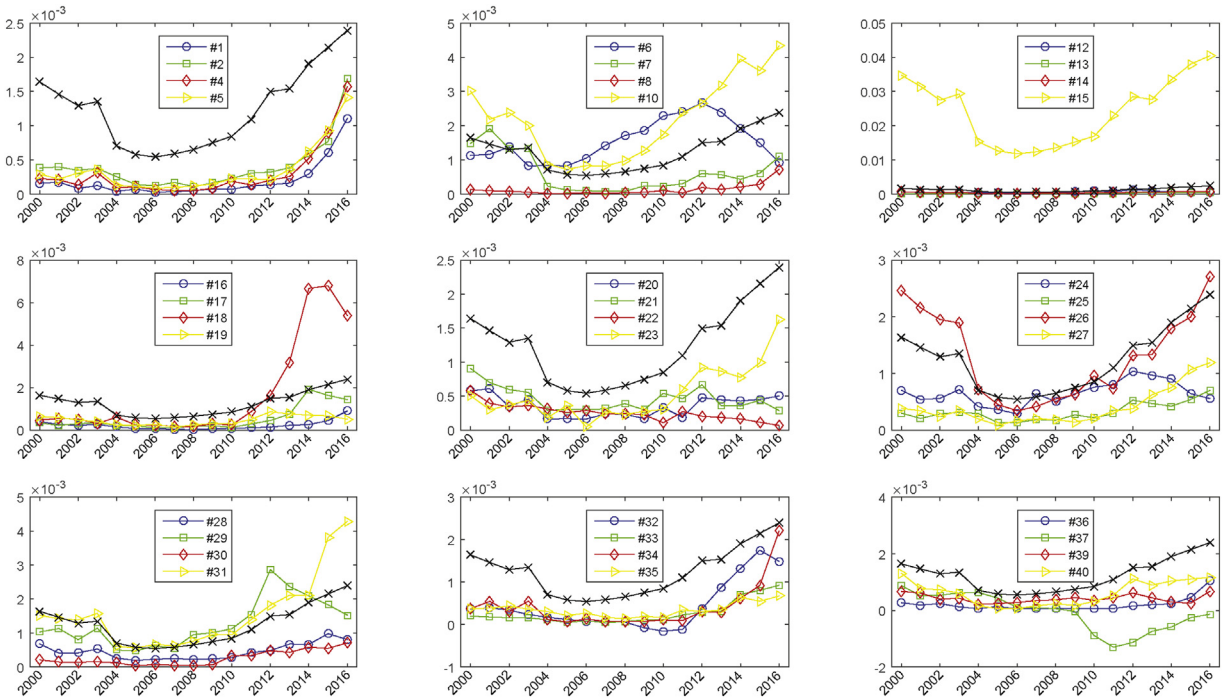

Fig. 12. The trend for the influence indicator.

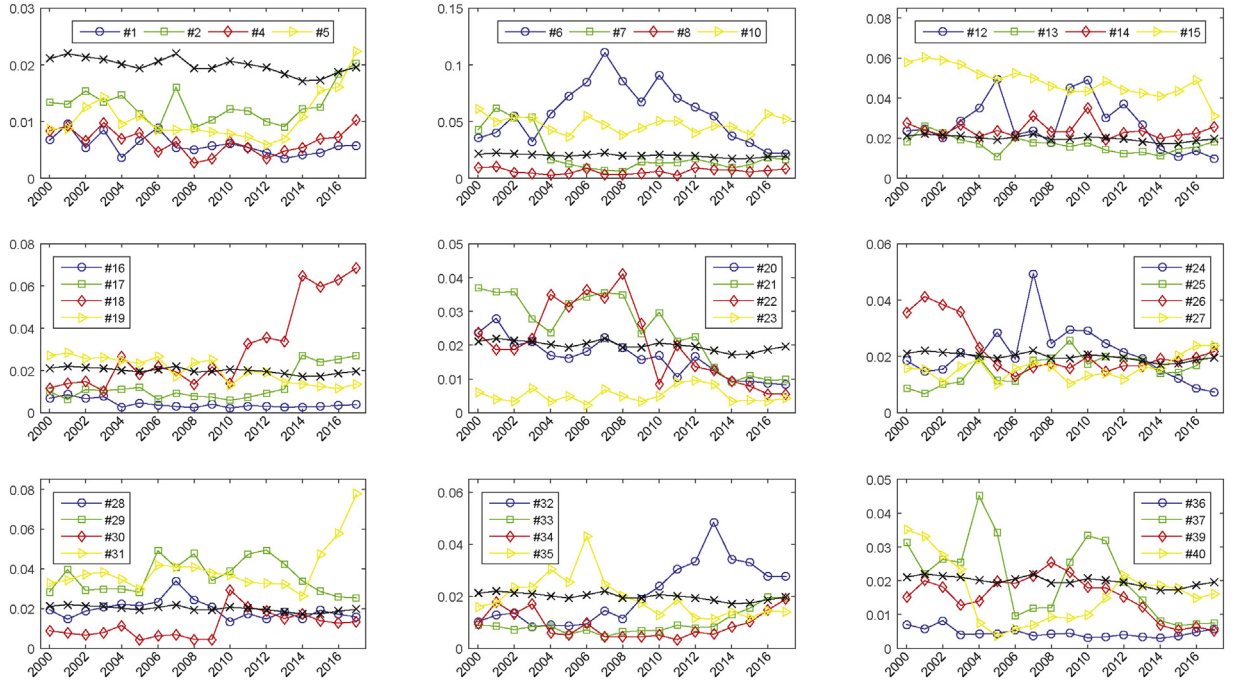

Fig. 13. The trend for the novelty indicator.

Table 7
The supports for all research topics from each indicator.

Topic	#1	#2	#4	#5	#6	#7	#8	#10	#12
Growth		•							
Influence								•	
Novelty					•			•	

Topic	#13	#14	#15	#16	#17	#18	#19	#20	#21
Growth		•	•		•	•			
Influence			•			•			
Novelty		•	•		•	•			

Topic	#22	#23	#24	#25	#26	#27	#28	#29	#30
Growth					•	•			
Influence									
Novelty					•	•		•	

Topic	#31	#32	#33	#34	#35	#36	#37	#39	#40
Growth	•			•					
Influence	•								
Novelty	•	•							

we have participated in the 2018–2019 Contest of Measuring Tech Emergence [4] pioneered by Professor Alan L. Porter with an adapted framework in Fig. 1, but only with three indicators, and won the Second Prize. This indicates that the coherence indicator may be not an important indicator for emerging topics detection with the framework based on topic models, but it should be still very important for the other bibliometric models (Wang, 2018).

For convenience, Table 7 summarizes the supports for all research topics from each indicator. From Table 7, one can see that the topic #15, #18 and #31 meet all emerging criteria defined in Section 3.2. Figs. 14–16 shows the top 10 words that have the highest probability conditioned on that topic year by year. From Figs. 14–16, sufficient coherence can be again

[4] https://vpinstitute.org/academic-portal/tech-emergence-contest/.

2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
system	system	system	system	system	system	system	system	system	system	system	system	system	synthetic	synthetic	synthetic
yeast	yeast	yeast	order	order	order	order	order	genetic	genetic	genetic	synthetic	synthetic	system	system	system
order	order	order	yeast	yeast	yeast	describe	genetic	order	order	synthetic	genetic	genetic	genetic	genetic	genetic
describe	describe	describe	describe	describe	describe	yeast	describe	describe	synthetic	order	yeast	yeast	yeast	yeast	yeast
part	part	part	part	part	part	genetic	yeast	yeast	yeast	yeast	order	order	assembly	strain	strain
product	product	product	product	product	product	part	part	part	describe	describe	describe	assembly	order	assembly	assembly
cerevisiae	produce	produce	produce	produce	genetic	product	product	synthetic	part	part	natural	natural	develop	develop	develop
produce	cerevisiae	natural	genetic	genetic	produce	produce	produce	product	natural	natural	part	describe	natural	natural	natural
construct	construct	genetic	natural	natural	natural	natural	natural	natural	complex	complex	complex	develop	strain	complex	require
present	present	construct	present	present	present	range	synthetic	produce	product	range	develop	complex	describe	order	complex

Fig. 14. The research topic #15 meeting emerging criteria in *gene editing* field.

2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
gene	gene	gene	gene	gene	gene	gene	gene	gene	target	target	target	target	target	target	target
target	target	target	target	target	target	target	target	target	gene	gene	gene	gene	genome	genome	genome
specific	specific	specific	specific	specific	specific	specific	editing	editing	editing	genome	genome	genome	gene	gene	editing
system	system	system	system	system	system	editing	specific	genome	genome	editing	editing	editing	editing	editing	gene
high	high	high	high	editing	editing	system	system	specific	system	system	system	system	system	system	crispr
method	method	method	editing	high	genome	genome	genome	system	specific	specific	specific	specific	talen	crispr	system
assay	assay	assay	method	method	high	method	method	method	method	efficient	efficient	efficient	specific	efficient	efficient
efficiency	efficiency	editing	assay	genome	method	high	efficiency	efficient	efficient	method	method	talen	efficient	talen	efficiency
editing	editing	efficiency	efficiency	efficiency	efficiency	efficiency	high	efficiency	efficiency	efficiency	efficiency	efficiency	efficiency	efficiency	method
site	site	site	genome	assay	assay	efficient	efficient	high	high	site	site	method	crispr	specific	specific

Fig. 15. The research topic #18 meeting emerging criteria in *gene editing* field.

2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
cell	cell	cell	cell	cell	cell	cell	cell	cell	expression	expression	expression	expression	expression	cell	cell
effect	effect	effect	expression	expression	expression	expression	expression	expression	cell	cell	cell	cell	cell	expression	expression
activity	expression	expression	effect	effect	effect	effect	effect	effect	effect	effect	effect	effect	induce	result	result
expression	activity	activity	activity	activity	activity	activity	activity	activity	activity	activity	induce	induce	effect	increase	increase
induce	induce	induce	induce	induce	induce	induce	induce	induce	induce	induce	activity	activity	result	induce	activity
result	result	result	result	result	result	result	result	result	activation	activation	result	result	activity	activity	show
activation	activation	activation	activation	activation	activation	activation	activation	activation	result	result	activation	activation	increase	effect	activation
apoptosis	apoptosis	role	role	role	role	role	pathway	pathway	increase	increase	increase	increase	activation	activation	induce
role	role	apoptosis	increase	pathway	pathway	pathway	role	role	pathway	pathway	pathway	pathway	pathway	pathway	pathway
growth	growth	increase	pathway	increase	increase	increase	increase	increase	role	role	role	role	show	show	signaling

Fig. 16. The research topic #31 meeting emerging criteria in *gene editing* field.

observed, though the values for the coherence indicator on the base of symmetrized KL or JS divergence are above the average, especially for topic #18.

The research topic #15 is about the genome editing in yeast (*Saccharomyces cerevisiae*). Saccharomyces cerevisiae has long been the most tractable organism for eukaryotic cell biology, owing to its genetic malleability, greatly facilitated by a preference for homologous recombination over non-homologous end joining for double stranded break repair (DiCarlo et al., 2013). Over the yeasts, biologists have taken advantage of this preference, allowing for the site-specific installation of genetic material and genomic edits with base-pair precision (Moqtaderi & Geisberg, 2013).

The research topic #18 means the assays to evaluate the efficiency of genome editing methods. Once one has chosen a CRISPR strategy to introduce a gene edit into the target cells, verification and characterization of the edit are usually required so as to monitor the result. In the literature, a variety of assays (such as T7E1, TIDE, IDAA, NGS and so on) may be used, depending on your experimental purpose and the nature of the gene edit (Sentmanat, Peters, Florian, Connelly, & Pruett-Miller, 2018).

The research topic #31 expresses the process of programmed cell death, or apoptosis. This is a very important problem for successful genome editing in practice, since electroporation with editing plasmids often leads to massive cell death (Li et al., 2018). In the meanwhile, gene therapy targeting the apoptotic machinery has great potential to benefit patients with threatening malignancies provided the availability of efficient and specific gene delivery and administration systems (Jia, Chen, & Yang, 2012).

It is not trivial to evaluate quantitatively the performance of our approach in the presence of inaccessible ground truth. But since the DIM and CIM models used in this work are both generative, one can run the inference algorithm on the scientific publications for the following two years exclusively. That is to say, global hidden variables (such as $\vec{\varphi}_k$ in the CIM model) and hyper-parameters are held fixed, and local ones (such as $\vec{\vartheta}_m^{(t)}$ and $\vec{\varphi}_k^{(t)}$ in the DIM model) are estimated for the unseen publications. Thus, the indicators in Section 3.2 can be calculated from the inferred results. The indicators' values from the inferred results are then compared to those predicted with MTLS-SVM. The MTLS-SVM can explain 52.42%, 53.86% and 57.38% of variance for the growth, influence and novelty indicators, respectively (cf. Table 6).

5. Conclusions

Emerging research topic detection can benefit the research foundations and policy-makers in terms of R&D policy and portfolio management, technology opportunities analysis, and management of innovation. With the long-term and recent interest in detecting emerging research topics, various approaches, such as citation-link and lexical methods, are proposed in the literature. Though, there is still a lack of well-established linkages between the clear conceptual definition of emerging research topics and the proposed indicators for operationalization.

In this study, thematic structures are extracted firstly with Dynamic Influence Model (DIM) from the scientific publications. Then, the growth, coherence, influence and novelty indicators are calculated, in which the first three indicators are based on the DIM model and the last one is based on the Citation Influence Model (CIM). As for the CIM model, the collapsed Gibbs sampling is done separately for the cited and citing publication parts. After that, the resulting indicators for the following two years are predicted with Multi-Task Least-Squares Support Vector Machine (MTLS-SVM) due to underling unknown relatedness amongst different indicators and different topics.

Experimental results on *gene editing* dataset discover three emerging research topics as follows: (a) the genome editing in yeast (*Saccharomyces cerevisiae*), (b) the assays to evaluate the efficiency of genome editing methods, and (c) the process of programmed cell death, or apoptosis. But, due to no benchmark datasets public available for emerging research topics, it is not trivial to evaluate quantitatively the performance of our framework in the presence of inaccessible ground truth. Nevertheless, the Second Prize for the 2018–2019 Contest of Measuring Tech Emergence with a similar framework makes clear that it is feasible to identify emerging research topics with our methodology. However, a scientific verification of our methodology still needs to be further investigated in the near future.

Last but not least, though four indicators are defined in this work according to the definitions of emerging technology and topic (Rotolo et al., 2015; Wang, 2018), the coherence indicator may be not an important indicator for emerging topics detection with the framework based on topic models, since the models can make sure the discovered topics to be sufficiently coherent. But it should be still very important for the other bibliometric models (Wang, 2018). In addition, in real-world application on nominating emerging topics, to reduce false positives, one should combine these indicators so that each nominated topic should receive multiple supports from the designed indicators. A feasible solution is to nominate the research topics and related evidences from half of the indicators at least, and then let domain experienced experts make decision.

Author contributions

Shuo Xu: Conceived and designed the analysis; Wrote the paper.
Liyuan Hao: Collected the data; Wrote the paper.
Xin An: Conceived and designed the analysis; Performed the analysis.
Guancan Yang: Performed the analysis.

Feifei Wang: Contributed data or analysis tools;

Acknowledgements

This research received the financial support from Social Science Foundation of Beijing Municipality under grant number 17GLB074, and Natural Science Foundation of Guangdong Province under grant number 2018A030313695. Our gratitude also goes to the anonymous reviewers for their valuable comments.

Appendix A.

As a matter of fact, due to conditional independence (Bishop, 2006), the collapsed Gibbs sampling for the CIM model can be done separately for the cited and citing publication parts in Fig. 4, but Dietz et al. (2007) mistakenly mixed them. The posterior distributions are re-derived in this article as follows.

(1) Cited Publication Part

$$\begin{aligned} & P(\vec{z}_{\tilde{m},\tilde{n}}|\vec{\tilde{w}},\vec{\tilde{z}}_{\neg(\tilde{m},\tilde{n})},\vec{\alpha},\vec{\beta}) \\ \propto\ & \frac{n^{(\tilde{w}_{\tilde{m},\tilde{n}})}_{\tilde{z}_{\tilde{m},\tilde{n}},\leftarrow}+\beta_{\tilde{w}_{\tilde{m},\tilde{n}}}-1}{\sum_{v=1}^{V}(n^{(v)}_{\tilde{z}_{\tilde{m},\tilde{n}},\leftarrow}+\beta_v)-1}\times(n^{(\tilde{z}_{\tilde{m},\tilde{n}})}_{\tilde{m},\leftarrow}+\alpha_{\tilde{z}_{\tilde{m},\tilde{n}}}-1) \end{aligned} \tag{4}$$

(2) Citing Publication Part

$$\begin{aligned} & P(z_{m,n}|\vec{w},\vec{z}_{\neg(m,n)},\vec{c},\vec{s},\vec{\alpha},\vec{\beta}) \\ =\ & \frac{n^{(w_{m,n})}_{z_{m,n},\rightarrow}+\beta_{w_{m,n}}-1}{\sum_{v=1}^{V}(n^{(v)}_{z_{m,n},\rightarrow}+\beta_v)-1}\times\begin{cases}\dfrac{n^{(z_{m,n})}_{c_{m,n},\rightarrow}+\alpha_{z_{m,n}}-1}{\sum_{k=1}^{K}(n^{(k)}_{c_{m,n},\rightarrow}+\alpha_k)-1}, & s_{m,n}=0\\ \dfrac{n^{(z_{m,n})}_{m}+\alpha_{z_{m,n}}-1}{\sum_{k=1}^{K}(n^{(k)}_{m}+\alpha_k)-1}, & s_{m,n}=1\end{cases} \end{aligned} \tag{5}$$

$$\begin{aligned} & P(s_{m,n}|\vec{z},\vec{s}_{\neg(m,n)},\vec{\alpha},\vec{\mu}) \\ =\ & \frac{n^{(s_{m,n})}_{m}+\mu_{s_{m,n}}-1}{\sum_{b=0}^{1}(n^{(b)}_{m}+\mu_b)-1}\times\begin{cases}\dfrac{n^{(z_{m,n})}_{c_{m,n},\rightarrow}+\alpha_{z_{m,n}}-1}{\sum_{k=1}^{K}(n^{(k)}_{c_{m,n},\rightarrow}+\alpha_k)-1}, & s_{m,n}=0\\ \dfrac{n^{(z_{m,n})}_{m}+\alpha_{z_{m,n}}-1}{\sum_{k=1}^{K}(n^{(k)}_{m}+\alpha_k)-1}, & s_{m,n}=1\end{cases} \end{aligned} \tag{6}$$

$$\begin{aligned} & P(c_{m,n}|\vec{z},\vec{c}_{\neg(m,n)},\vec{\alpha},\{\vec{\delta}_m\}_{m=1}^{M}) \\ =\ & \frac{n^{(c_{m,n})}_{m}+\delta_{m,c_{m,n}}-1}{\sum_{\ell=1}^{L_m}(n^{(\ell)}_{m}+\delta_{m,\ell})-1}\times\frac{n^{(z_{m,n})}_{c_{m,n},\rightarrow}+\alpha_{z_{m,n}}-1}{\sum_{k=1}^{K}(n^{(k)}_{c_{m,n},\rightarrow}+\alpha_k)-1} \end{aligned} \tag{7}$$

where $\vec{\tilde{z}}_{\neg(\tilde{m},\tilde{n})}$ represents the topic assignments for all word tokens except $\tilde{w}_{\tilde{m},\tilde{n}}$, $\vec{z}_{\neg(m,n)}$, $\vec{c}_{\neg(m,n)}$ and $\vec{s}_{\neg(m,n)}$ represent the topic, cited publication and status assignments for all word tokens except $w_{m,n}$. Here, $n^{(v)}_{k,\leftarrow}$ is the number of tokens of word v in the cited publications which are assigned to topic k, $n^{(k)}_{\tilde{m},\leftarrow}$ represent the number of tokens in the cited publication $\tilde{m}$ which are assigned to the topic k, $n^{(v)}_{k,\rightarrow}$ is the number of tokens of word v in the citing publications which are assigned to topic k, $n^{(\ell)}_{m}$) is the number of tokens in the citing publication m which are assigned to the cited publication ℓ, $n^{(b)}_{m}$ is the number of tokens in the citing publication m which are assigned to status b, $n^{(k)}_{\tilde{m},\rightarrow}$ represents the number of tokens in the citing publications which are assigned to the topic k in the cited publication $\tilde{m}$, and $n^{(k)}_{m}$ represents the number of tokens in the citing publication m which are assigned to topic k.

Eq. (4) is the same as that in the standard LDA model (Griffiths & Steyvers, 2004). Using the expectation of Dirichlet and Beta distributions, one can easily obtain the resulting model parameters as follows.

$$\tilde{\vartheta}_{\tilde{m},k}=\frac{n^{(k)}_{\tilde{m},\leftarrow}+n^{(k)}_{\tilde{m},\rightarrow}+\alpha_k}{\sum_{k=1}^{K}(n^{(k)}_{\tilde{m},\leftarrow}+n^{(k)}_{\tilde{m},\rightarrow}+\alpha_k)}=\frac{n^{(k)}_{\tilde{m},\leftarrow}+n^{(k)}_{\tilde{m},\leftarrow}+\alpha_k}{\tilde{N}_{\tilde{m}}+\sum_{k=1}^{K}(n^{(k)}_{\tilde{m},\leftarrow}+\alpha_k)} \tag{8}$$

$$\vartheta_{m,k} = \frac{n_{m,\rightarrow}^{(k)} + \alpha_k}{\sum_{k=1}^{K}(n_{m,\rightarrow}^{(k)} + \alpha_k)} \tag{9}$$

$$\varphi_{k,v} = \frac{n_{k,\leftarrow}^{(v)} + n_{k,\rightarrow}^{(v)} + \beta_v}{\sum_{v=1}^{V}(n_{k,\leftarrow}^{(v)} + n_{k,\rightarrow}^{(v)} + \beta_v)} \tag{10}$$

$$\psi_{m,\ell} = \frac{n_m^{(\ell)} + \delta_{m,\ell}}{\sum_{\ell=1}^{L_m}(n_m^{(\ell)} + \delta_{m,\ell})} \tag{11}$$

$$\lambda_{m,b} = \frac{n_m^{(b)} + \mu_b}{\sum_{b=0}^{1}(n_m^{(b)} + \mu_b)} \tag{12}$$

References

AlSumait, L., Barbará, D., Gentle, J., & Domeniconi, C. (2009). Topic significance ranking of LDA generative models. In R. Goebel, J. Siekmann, & W. Wahlster (Eds.), *Proceedings of the European conference on machine learning and principles and practice of knowledge discovery in databases* (pp. 67–82). http://dx.doi.org/10.1007/978-3-642-04180-8_22

An, X. Y., & Wu, Q. Q. (2011). Co-word analysis of the trends in stem cells field based on subject heading weighting. *Scientometrics*, *88*, 133–144. http://dx.doi.org/10.1007/s11192-011-0374-1

An, X., Xu, S., Wen, Y., & Hu, M. (2014). A shared interest discovery model for co-author relationship in SNS. *International Journal of Distributed Sensor Networks*, *2014*, 1–9. http://dx.doi.org/10.1155/2014/820715

An, X., Wen, Y., Zhang, Y., & Xu, S. (2019). Evaluation of the forestry and environmental conservation policies in western China with multi-output regression method. *Computers and Electronics in Agriculture*, *157*, 239–246. http://dx.doi.org/10.1016/j.compag.2018.12.035

Andrieu, C., de Freitas, N., Doucet, A., & Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine Learning*, *50*, 5–43. http://dx.doi.org/10.1023/A:1020281327116

Argyriou, A., Evgeniou, T., & Pontil, M. (2008). Convex multi-task feature learning. *Machine Learning*, *73*, 243–272. http://dx.doi.org/10.1007/s10994-007-5040-8

Azzonpardi, L., Girolami, M., & van Risjbergen, K. (2003). Investigating the relationship between language model perplexity and IR precision-recall measures. In *Proceedings of the 26th international ACM SIGIR conference on research and development in information retrieval*. pp. 369–370. New York, NY, USA: ACM. http://dx.doi.org/10.1145/860435.860505

Bakker, B., & Heskes, T. (2003). Task clustering and gating for Bayesian multitask learning. *Journal of Machine Learning Research*, *4*, 83–99. http://dx.doi.org/10.1162/153244304322765658

Ben-David, S., & Borbely, R. S. (2008). A notion of task relatedness yielding provable multiple-task learning guarantees. *Machine Learning*, *73*, 273–287. http://dx.doi.org/10.1007/s10994-007-5043-5

Bishop, C. M. (Ed.). (2006). *Pattern recognition and machine learning*. Springer.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on machine learning*. pp. 113–120. New York, NY, USA: ACM. http://dx.doi.org/10.1145/1143844.1143859

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

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 American Society for Information Science*, *61*, 2389–2404. http://dx.doi.org/10.1002/asi.21419

Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., et al. (2011). Clustering more than two million biomedical publications: Comparing the accuraries of nine text-based similarity approaches. *PLoS ONE*, *6*, e18029.

Boyack, K. W., Klavans, R., Small, H., & Ungar, L. (2014). Characterizing the emergence of two nanotechnology topics using a contemporaneous global micro-model of science. *Journal of Engineering and Technology Management*, *32*, 147–159. http://dx.doi.org/10.1016/j.jengtecman.2013.07.001

Caruana, R. (1997). Multitask learning. *Machine Learning*, *28*, 41–75. http://dx.doi.org/10.1023/A:1007379606734

Chapelle, O., Shivaswamy, P., Vadrevu, S., Weinberger, K., Zhang, Y., & Tseng, B. (2010). Multi-task learning for boosting with application to web search ranking. In *Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining*. pp. 1189–1198. New York, NY, USA: ACM. http://dx.doi.org/10.1145/1835804.1835953

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, *57*, 359–377. http://dx.doi.org/10.1002/asi.20317

Chen, B., Tsutsui, S., Ding, Y., & Ma, F. (2017). Understanding the topic evolution in a scientific domain: An exploratory study for the field of information retrieval. *Journal of Informetrics*, *11*, 1175–1189. http://dx.doi.org/10.1016/j.joi.2017.10.003

Cozzens, S., Gatchair, S., Kang, J., Kim, K.-S., Lee, H. J., Ordóñez, G., et al. (2010). Emerging technologies: Quantitative identification and measurement. *Technological Forecasting & Social Change*, *22*, 361–376. http://dx.doi.org/10.1080/09537321003647396

DiCarlo, J. E., Norville, J. E., Mali, P., Rios, X., Aach, J., & Church, G. M. (2013). Genome engineering in Saccharomyces cerevisiae using CRISPR-C as systems. *Nucleic Acids Research*, *41*, 4336–4343. http://dx.doi.org/10.1093/nar/gkt135

Dietz, L., Bickel, S., & Scheffer, T. (2007). Unsupervised prediction of citation influences. In *Proceedings of the 24th international conference on machine learning*. pp. 233–240. New York, NY, USA: ACM. http://dx.doi.org/10.1145/1273496.1273526

Gerrish, S. M., & Blei, D. M. (2010). A language-based approach to measuring scholarly impact. *Proceedings of the 27th international conference on machine learning*, 375–382.

Glänzel, W., & Thijs, B. (2012). Using 'core documents' for detecting and labelling new emerging topics. *Scientometrics*, *91*, 399–416. http://dx.doi.org/10.1007/s11192-011-0591-7

González-Alcaide, G., Llorente, P., & Ramos, J. M. (2016). Bibliometric indicators to identify emerging research fields: Publications on mass gatherings. *Scientometrics*, *109*, 1283–1298. http://dx.doi.org/10.1007/s11192-016-2083-2

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 5228–5235. http://dx.doi.org/10.1073/pnas.0307752101

Guo, H., Weingart, S., & Börner, K. (2011). Mixed-indicators model for identifying emerging research areas. *Scientometrics*, *89*, 421–435. http://dx.doi.org/10.1007/s11192-011-0433-7

Hansmann, T., & Niemeyer, P. (2014). Big data – Characterizing an emerging research field using topic models. In *Proceedings of the IEEE/WIC/ACM international joint conferences on web intelligence and intelligent agent technologies*. pp. 43–51. Washington, DC, USA: IEEE Computer Society. http://dx.doi.org/10.1109/WI-IAT.2014.15

Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*, *14*, 1303–1347.

Hu, C.-P., Hu, J.-M., Deng, S.-L., & Liu, Y. (2013). A co-word analysis of library and information science in China. *Scientometrics*, *97*, 369–382. http://dx.doi.org/10.1007/s11192-013-1076-7
Huang, M.-H., & Chang, C.-P. (2014). Detecting research fronts in OLED field using bibliographic coupling with sliding window. *Scientometrics*, *98*, 1721–1744. http://dx.doi.org/10.1007/s11192-013-1126-1
Iwata, T., Yamada, T., Sakurai, Y., & Ueda, N. (2012). Sequential modeling of topic dynamics with multiple timescales. *ACM Transactions on Knowledge and Discovery from Data*, *5*, 19:1–19:27. http://dx.doi.org/10.1145/2086737.2086739
Jarić, I., Knežević-Jarić, J., & Lenhardt, M. (2014). Relative age of references as a tool to identify emerging research fields with an application to the field of ecology and environmental sciences. *Scientometrics*, *100*, 519–529. http://dx.doi.org/10.1007/s11192-014-1268-9
Jia, L.-T., Chen, S.-Y., & Yang, A.-G. (2012). Cancer gene therapy targeting cellular apoptosis machinery. *Cancer Treatment Reviews*, *38*, 868–876. http://dx.doi.org/10.1016/j.ctrv.2012.06.008
Jiang, Z., Liu, X., & Gao, L. (2015). Chronological citation recommendation with information-need shifting. *Proceedings of the 24th ACM international conference on information and knowledge management*, 1291–1300. http://dx.doi.org/10.1145/2806416.2806567
Jordan, M., Grhahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, *37*, 183–233. http://dx.doi.org/10.1023/A:1007665907178
Kawamae, N., & Higashinaka, R. (2010). Trend detection model. In M. Rappa, P. Jones, J. Freire, & S. Chakrabarti (Eds.), *Proceedings of the 19th international conference on world wide web* (pp. 1129–1130). New York, NY, USA: ACM.
Kleinberg, J. (2002). Bursty and hierarchical structure in streams. In *Proceedings of the 8th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 91–101).
Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*, 79–86. http://dx.doi.org/10.1214/aoms/1177729694
Li, X.-L., Li, G.-H., Fu, J., Fu, Y.-W., Zhang, L., Chen, W., et al. (2018). Highly efficient genome editing via CRISPR-Cas9 in human pluripotent stem cells is achived by transient BCL-XL overexpression. *Nucleic Acids Research*, *46*, 10195–10215. http://dx.doi.org/10.1093/nar/gky804
Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, *37*, 145–151. http://dx.doi.org/10.1109/18.61115
Liu, Z., Yin, Y., Liu, W., & Dunford, M. (2015). Visualizing the intellectual structure and evolution of innovation systems research: A bibliometric analysis. *Scientometrics*, *103*, 135–158. http://dx.doi.org/10.1007/s11192-014-1517-y
Ma, R. (2012). Author bibliographic coupling analysis: A test based on a Chinese academic database. *Journal of Informetrics*, *6*, 532–542. http://dx.doi.org/10.1016/j.joi.2012.04.006
Moqtaderi, Z., & Geisberg, J. V. (2013). Construction of mutant alleles in *Saccharomyces cerevisiae* without cloning: Overview and the *Delitto Perfetto* method. *Current Protocols in Molecular Biology*, *104* http://dx.doi.org/10.1002/0471142727.mb1310cs104, pp. 13.10C.1–13.10C.17
Morris, S. A., Yen, G., Wu, Z., & Asnake, B. (2003). Time line visualization of research fronts. *Journal of the Association for Information Science and Technology*, *54*, 413–422. http://dx.doi.org/10.1002/asi.10227
Ohniwa, R. L., Hibino, A., & Takeyasu, K. (2010). Trends in research foci in life science fields over the last 30 years monitored by emerging topics. *Scientometrics*, *85*, 111–127. http://dx.doi.org/10.1007/s11192-010-0252-2
Porter, A. L., Garner, J., Carley, S. F., & Newman, N. C. (2019). Emergence scoring to identify R&D topics and key players. *Technological Forecasting & Social Change*, *146*, 628–643. http://dx.doi.org/10.1016/j.techfore.2018.04.016
Reiss, T., Vignola-Gagné, E., Kukk, P., Glänzel, W., & Thijs, B. (2013). *ERACEP – Emerging research areas and their coverage by ERC-supported projects*. Technical Report European Research Council.
Robinson, D. K. R., Ruivenkamp, M., & Rip, A. (2007). Tracking the evolution of new and emerging S&T via statement-linkages: Vision assessment in molecular machines. *Scientometrics*, *70*, 831–858. http://dx.doi.org/10.1007/s11192-007-0314-2
Roche, I., Besagni, D., François, C., Hörlesberger, M., & Schiebel, E. (2010). Identification and characterisation of technological topics in the field of molecular biology. *Scientometrics*, *82*, 663–676. http://dx.doi.org/10.1007/s11192-010-0178-8
Rotolo, D., Hicks, D., & Martin, B. R. (2015). What is an emerging technology? *Research Policy*, *44*, 1827–1843. http://dx.doi.org/10.1016/j.respol.2015.06.006
Sætre, R., Yoshida, K., Yakushiji, A., Miyao, Y., Matsubayashi, Y., & Ohta, T. (2007). AKANE system: Protein–protein interaction pairs in the BioCreAtIvE2 challenge, PPI-IPS subtask. *Proceedings of the 2nd biocreative challenge evaluation workshop*, 209–212.
Schultz, L. I., & Joutz, F. L. (2010). Methods for identifying emerging general purpose technologies: A case study of nanotechnologies. *Scientometrics*, *85*, 155–170. http://dx.doi.org/10.1007/s11192-010-0160-5
Sentmanat, M. F., Peters, S. T., Florian, C. P., Connelly, J. P., & Pruett-Miller, S. M. (2018). A survey of validation strategies for CRISPR-Cas9 editing. *Scientific Reports*, 8. http://dx.doi.org/10.1038/s41598-018-19441-8
Shibata, N., Kajikawa, Y., Takeda, Y., & Matsushima, K. (2008). Detecting emerging research fronts based on topological measures in citation networks of scientific publications. *Technovation*, *28*, 758–775. http://dx.doi.org/10.1016/j.technovation.2008.03.009
Small, H., & Griffith, B. C. (1974). The structure of scientific literatures I: Identifying and graphing specialties. *Science Studies*, *4*, 17–40.
Small, H., Boyack, K. W., & Klavans, R. (2014). Identifying emerging topics in science and technology. *Research Policy*, *43*, 1450–1467. http://dx.doi.org/10.1016/j.respol.2014.02.005
Soriano, A. S., Álvarez, C. L., & Valdés, R. M. T. (2018). Bibliometric analysis to identify an emerging research area: Public relations intelligence-a challenge to strengthen technological observatories in the network society. *Scientometrics*, *115*, 1591–1614. http://dx.doi.org/10.1007/s11192-018-2651-8
Suykens, J. A. K., Van Gestel, T., De Brabanter, J., De Moor, B., & Vandewalle, J. (Eds.). (2002). *Least squares support vector machines*. World Scientific Pub. Co.
Tsuruoka, Y., Tateishi, Y., Kim, J.-D., Ohta, T., McNaught, J., Ananiadou, S., et al. (2005). Developing a robust part-of-speech tagger for biomedical text. In P. Bozanis, & E. N. Houstis (Eds.), *Proceedings of the 10th panhellenic conference on informatics* (pp. 382–392). http://dx.doi.org/10.1007/11573036_36
Tu, Y.-N., & Seng, J.-L. (2012). Indices of novelty for emerging topic detection. *Information Processing & Management*, *48*, 303–325. http://dx.doi.org/10.1016/j.ipm.2011.07.006
Waltman, L., & van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the Association for Information Science and Technology*, *63*, 2378–2392. http://dx.doi.org/10.1002/asi.22748
Wang, X., & McCallum, A. (2006). Topics over time: A non-Markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining*. pp. 424–433. New York, NY, USA: ACM. http://dx.doi.org/10.1145/1150402.1150450
Wang, Q. (2018). A bibliometric model for identifying emerging research topics. *Journal of the Association for Information Science and Technology*, *69*, 290–304. http://dx.doi.org/10.1002/asi.23930
Wang, C., Blei, D., & Heckerman, D. (2008). Continuous time dynamic topic models. *Proceedings of the 24th international conference in uncertainty in artificial intelligence*, 579–586.
Xu, S., An, X., Qiao, X., Zhu, L., & Li, L. (2013). Multi-output least-squares support vector regression machines. *Pattern Recognition Letters*, *34*, 1078–1084. http://dx.doi.org/10.1016/j.patrec.2013.01.015
Xu, S., An, X., Qiao, X., & Zhu, L. (2014). Multi-task least-squares support vector machines. *Multimedia Tools and Applications*, *71*, 699–715. http://dx.doi.org/10.1007/s11042-013-1526-5
Xu, S., Shi, Q., Qiao, X., Zhu, L., Zhang, H., Jung, H., et al. (2014). A dynamic users' interest discovery model with distributed inference algorithm. *International Journal of Distributed Sensor Networks*, 1–11. http://dx.doi.org/10.1155/2014/280892
Xu, S., Liu, J., Zhai, D., An, X., Wang, Z., & Pang, H. (2018). Overlapping thematic structures extraction with mixed-membership stochastic blockmodel. *Scientometrics*, *117*, 61–84. http://dx.doi.org/10.1007/s11192-018-2841-4
Xu, S., Hao, L., An, X., Pang, H., & Li, T. (2019). Review on emerging research topics with key-route main path analysis. *Scientometrics*.

Xu, S., Hao, L., An, X., Zhai, D., & Pang, H. (2019). Types of doi errors of cited references in Web of Science with a cleaning method. *Scientometrics*, *120*, 1427–1437. http://dx.doi.org/10.1007/s11192-019-03162-4

Xu, S., Zhai, D., Wang, F., An, X., Pang, H., & Sun, Y. (2019). A novel method for topic linkages between the scientific publications and patents. *Journal of the Association for Information Science and Technology*, *70*, 1026–1042. http://dx.doi.org/10.1002/asi.24175

Yan, E. (2014). Research dynamics: Measuring the continuity and popularity of research topics. *Journal of Informetrics*, *8*, 98–110. http://dx.doi.org/10.1016/j.joi.2013.10.010

Zhao, D., & Strotmann, A. (2008). Evolution of research activities and intellectual influences in information science 1996-2005: Introducing author bibliographic-coupling analysis. *Journal of the Association for Information Science and Technology*, *59*, 2070–2086. http://dx.doi.org/10.1002/asi.20910

Zhu, X., Turney, P., Lemire, D., & Vellino, A. (2014). Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology*, *66*, 408–427. http://dx.doi.org/10.1002/asi.23179