Contents lists available at ScienceDirect

Technological Forecasting & Social Change

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

Profiling academic-industrial collaborations in bibliometric-enhanced topic networks: A case study on digitalization research

Hongshu Chen[a,*], Qianqian Jin[a], Ximeng Wang[b], Fei Xiong[c]

[a] *School of Management and Economics, Beijing Institute of Technology, Beijing, China*
[b] *Cyber Finance Department, Postal Savings Bank of China, Beijing, China*
[c] *School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, China*

ARTICLE INFO

Keywords:
Topic networks
Topic vectorization
Topic modeling
Word2Vec
Academic-industrial collaboration

ABSTRACT

Collaborations between industry and academia provide a key pathway for innovation and serve as a stimulus for basic and applied research. The collaborative innovations of the two communities are embedded in both the collaborative networks of these organizations and the knowledge networks established by coupling among knowledge elements in the collaborative content. However, existing studies on academic-industrial collaborations have mainly been concerned with analyzing these interactions at the institutional level. To fill the gap of profiling collaborative content and to inspire related studies, this paper provides a bibliometric-enhanced method of mapping topic networks and measuring the semantic structures of academic-industrial collaboration. Via this method, topics can be extracted, vectorized, and correlated to construct a bibliometric-enhanced topic network as a representation of the collaborative content generated by these partnerships. Examining the structural properties of the topic network can provide comprehensive insights for future academic-industrial research collaborations. To showcase these insights, we conducted a case study involving both articles and patents in the field of digitalization. As the case study shows, the method provided in this paper can serve as a tool for cooperative research planning, innovation management, and problem-solving in a given target area of research.

1. Introduction

Interactions between industry and academia are an important pathway for innovation, as they provide inspiration for knowledge fusion (Mao et al., 2020). For academics, collaborating with industry is an opportunity for funding, gaining access to advanced equipment, and gleaning cutting edge insights by complementing their theories with practice (Lee, 2000; Skute et al., 2019). For industry, collaborations with academia can help to broaden knowledge, improve a firm's problem solving capabilities, and provide access to the latest theoretical breakthroughs in a technology. This all enhances innovation performance and helps a company remain competitive (Perkmann et al., 2011). As Wang et al. (2015) report, over the past decade, academic institutions have become more actively involved in commercial activities and firms have sought more research cooperation. As such, various disciplines have witnessed an upward trend in industry-academia collaboration. In kind, an increasing number of studies have focused on the phenomenon of academic-industrial collaboration in an attempt to both understand the dynamics behind these interactions and to improve the productivity of these partnerships.

Academic-industrial collaboration is characterized by complex interactions of information, technologies, knowledge, and other resources. The innovations of the two types of organizations are embedded in both knowledge networks established by coupling among knowledge elements and in social networks formed via collaborative relationships at the organizational level (Wang and Hsu, 2014; Guan and Liu, 2016). Existing research provides promising analyses of the cooperative interactions between industry and academia industry-academia collaboration using co-authored publications, patents, and questionnaires, staff numbers, grants, research contracts and geographic information (Huang et al., 2015; Teixeira and Mota, 2012; Magazinik et al., 2019). However, a recent survey highlights that studies on academic-industrial collaborations have mainly focused on processes at the institutional level. For example, the patterns and structures of participant interactions have been mapped via collaborative networks (Nsanzumuhire and Groot, 2020). Further, quantitative methods and tools based on network

[*] Corresponding author.
E-mail address: hongshu.chen@bit.edu.cn (H. Chen).

https://doi.org/10.1016/j.techfore.2021.121402
Received 30 June 2020; Received in revised form 24 November 2021; Accepted 25 November 2021
Available online 1 December 2021

analysis theory and techniques have been developed to reveal the relations among these participants, including individuals, teams, organizations and countries (Akhtar et al., 2019). But comparatively less attention has been paid to analyzing the content of these interactions between academia and industry.

The pursuit of technological innovation is a key driving force behind the academic-industry collaboration. Here, the content of the collaboration, that is, the co-addressed topics, reveals the shared focus of attention, especially for data- and technology-driven industries. This co-addressed content is converted into a knowledge network (Phelps et al., 2012; Guan and Liu, 2016) or a semantic network (Hellsten and Leydesdorff, 2020) for analysis, to explain the processes and trends of knowledge interaction and absorption. In existing research, these knowledge networks are mainly constructed from co-occurring keywords in scientific publications (Guan et al., 2017), co-occurring hashtags (Hellsten and Leydesdorff, 2020) or co-applications of international patent classification codes (IPCs) (Guan and Liu, 2016; Chang, 2017), which offers the most concise representations of the technological or knowledge features discussed. However, what gets overlooked with these keyword-based or tag-based network is the context of the textual content. It is difficult to examine knowledge structures when only considering straightforward tags, keywords, or classifications, especially when they are based on a subjective and individual taxonomy (Roth and Cointet, 2010). In addition, for a given target area of research, the number of keywords and the size of IPC co-occurrence networks can be limited if only academic-industry collaboration items are considered. The academia-dominant or industry-dominant knowledge elements are often not discussed simultaneously. Thus, a full picture of the knowledge flows is lacking.

Over the past decade, methods of topic extraction and analysis have been enhanced by the power of modern computing techniques and machine learning algorithms. From a topic extraction perspective, an increasing number of studies are identifying concrete chunks of knowledge researched by collaborating partners using appraoches such as topic modeling, word clustering, and so forth (Woltmann and Alkærsig, 2018; Suominen et al., 2019; Zhao et al., 2019). Word embedding, in recent years, has also caught researchers' eyes for its promising ability to map words into numeric vectors. With these techniques, the latent semantics in large-scale text data can be discovered while considering the context of the content (Mikolov et al., 2013). These vectors can be used to replace traditional word representations in scientific text mining. This holds great potential for topic extraction (Zhang et al., 2018). Analyzing the complex interactions of the extracted topics, however, warrants further research compared to topic extraction. There are inherent limitations to methods based on topic modeling when handling correlations between different topics, as it is hard to detect similarity ratios via direct distance measurements (Jung and Yoon, 2020).

Based on increasing interest in both academic-industry collaborations and new topic extraction techniques for examining the focus of joint attention, this paper provides a bibliometric-enhanced method of mapping topic networks to better measure the semantic characteristics of a corpus corresponding with collaborative relationships of organizations. The challenge of exploring and modeling textual content in the context of academic-industry collaboration requires that one considers both the interactions between the two communities and the topics they are addressing together. Our framework balances statistical evaluation with human interpretation when perform topic modeling, to extract three categories of topics: academia-driven topics, industry-driven topics, and collaborative topics. The featured topics are then vectorized by exploiting word embedding technique to reveal their semantic relations, which makes topic similarity detection possible. All the topic vectors are finally mapped into a topic network. The constructed network represents co-addressed, academia-driven and industry-driven topics, while at the same time highlighting the interactive relationships between these topics. Examining the structural properties of the topic network can provide comprehensive insights for future academic-industrial research collaborations. Our empirical analysis focuses on the area of digitalization, which has drawn considerable attention from both industry and academia in recent years. Using both articles and patents, the case study profiles the links between academia and industry. It also exposes the corresponding semantic structure of the collaborative content. The insights revealed can potentially be used for practical collaboration planning and innovation management.

The rest of this paper is organized as follows: We review related work in Section 2. Section 3 provides details of research design, explaining the modules of topic network construction. In Section 4, we apply our proposed method to digitalization articles and patents, and discuss the finding of the case study in serving as a reference for future academic and industrial research cooperation. The last section concludes the proposed method, explains the limitations of the study, and addresses future research directions.

2. Related work

With this study, our focus is on constructing topic networks in a given target area for academic-industrial collaboration research purpose, while, at the same time, highlighting the interactions between these topics. Hence, work relating to both collaborative networks and knowledge networks is reviewed in this section. Topic extraction and text-mining analytics are also pertinent to this work; work relevant is reviewed in Section 2.2.

2.1. Collaborative networks and knowledge networks in academic-industrial collaboration research

Existing studies on academic-industrial collaborations have provided promising outcomes. To reveal the cooperative relations between industry and academia, researchers have been focusing on the quantity of publications and patents that have involved industrial contributions. For example, Fan et al. (2015) extracted articles jointly authored by university researchers and enterprise personnel, trying to dig out the patterns of collaboration between Chinese university and industry. After analyzing a combined dataset containing hundreds of top American universities, Lin (2017) claimed that too much industry collaboration, as well as too little, is detrimental to academic innovation. Coincidentally, Wang et al. (2016) exploited a unique dataset of 61 Chinese universities to discover a U-shaped relationship when it comes to university-industry interaction and teaching performance. Giunta et al. (2016) focused on co-publications in the biopharmaceuticals in Italy and summarized several decisive factors affecting the combination of production and scientific research. However, the investigation of industry-academia research collaboration has seldom been considered from a content perspective, even though it is worthy of in-depth study.

In understanding and modeling collaborative content, a number of empirical studies have demonstrated that the innovations of the two types of organizations are embedded in both knowledge networks established by coupling among knowledge elements and in social networks formed by collaborative relationships of participants, at individual level or organizational level (Guan and Liu, 2016). Although promising analyses on actor interactions have been conducted via collaborative networks (Nsanzumuhire and Groot, 2020), the content of the interactions between academia and industry has only been systematically modeled in recent years. Focusing on content-based analysis in social relationships among individuals or higher level collectives, Guan and Liu (2016) constructed knowledge networks from the co-application of IPCs. Co-occurring keywords from academic papers (Guan et al., 2017) and co-occurrences of hashtags (Hellsten and Leydesdorff, 2020) have also served as representations of the knowledge structured in the collaborations. When considering the context of the content, however, it is not sufficient to adopt only straightforward tags, keywords or classifications to form the knowledge networks. modeling with the context considered warrants further research.

2.2. Topic extraction and analytics in scientific text mining

Over the past decade, topic models have been providing new inspiration for turning bibliometric and textual data into valuable and insightful text mining research. As such, these have served as an efficient tool for discovering latent and potentially useful content (Chen et al., 2021). As one of the most well-known topic modeling methods, latent Dirichlet allocation (LDA) discovers topics in large collections of documents (Blei et al., 2003). LDA presumes that each document is a mixture of topics and that each topic is a mixture of words (Blei, 2012).

LDA has been used as a powerful tool in various fields, including information and knowledge management, innovation management, technology forecasting and many others (Lee and Kang, 2018; Yang et al., 2013). For example, using LDA, Chen et al. (2015) extracted topics from a massive corpus of patent documents published by the United States Patent and Trademark Office (USPTO). From this corpus, they identified topics changes over years. Lamba and Madhusudhan (2019) used LDA to analyze articles retrieved from a journal in the field of library and information technology to explore the main topics discussed in this area. Song and Suh (2019) combined LDA with a network analysis method to forecast technological innovation trends and integration in industrial safety patents. As another benefit, co-word-based approaches have difficulty processing technological synonyms, especially in emerging sectors (Peters and van Raan, 1993), whereas LDA provides word distributions and topic distributions to reveal latent semantic structures, thus avoiding these difficulties.

As with all topic extraction/clustering approaches, LDA suffers from the problem of having to pre-identify the number of topics to be extracted. Usually, this is done through a priori knowledge. However, that raises the chicken and egg syndrome of having to understand the thematic structure of the corpus before applying the LDA model so as to understand the thematic structure of the corpus (Suominen and Toivanen, 2016). In existing research, likelihood (Griffiths and Steyvers, 2004) and perplexity (De Battisti et al., 2015) are the two most common approaches for deciding the number of topics. But statistical approaches tend to lead to a comparatively large number of topics, which is challenging for humans to interpret. In addition, there are inherent limitations to topic modeling methods when handling topic correlations between different, dynamic topics. The topics discovered via LDA are made up of words and corresponding distributions. Similarity measures of topics are limited to the ratio of word similarity, yet a set of topics from a document collection is bound to be disparate (Jung and Yoon, 2020). For this reason, it can be hard to detect topic correlation ratios through direct distance measurements. This inherent limitation can make it difficult to detect and compare evolving topics, map semantic networks, and perform other types of content analysis based on similarity measures (Chen et al., 2017; Jung and Yoon, 2020). Thus, methods for capturing multiple connections between topics warrant further research.

From a topic extraction perspective, word embedding has captured immense interest from scientific text mining researchers in the past two years. This is because has shown great capabilities at mapping terms and concepts in vector space and in discovering latent semantic correlations. Zhang et al. (2018) proposed a novel kernel k-means clustering method that effectively extracts topics from bibliometric data by incorporating word embedding. Greiner-Petter et al. (2020) applied a word embedding method to retrieve math information from scientific articles. Lee et al. (2020) combined and learned word embeddings with WordNet for semantic relatedness and similarity measurement. In general, word embedding aims to map words from vocabularies to form numeric vectors with the basic assumption that words with similar contexts will have similar meanings (Firth, 1957). Among the techniques that apply this approach, Word2Vec is arguably the most accepted and used method (Mikolov et al., 2013), which have two specific models: Skip-gram model and continuous bag-of-word model (CBOW) model (Le and Mikolov, 2014). There has been many research comparing the two models, but according to the independent benchmarking conducted by Levy et al. (2015), there is no fundamental performance difference between the two models. Overall, word embedding has and will continue to show great potential in text mining research.

3. Methodology

The methodology consists of four steps: data preprocessing, bibliometric-enhanced topic modeling, topic vectorization, and topic network construction with analytics. Each step is described in more detail in the following.

3.1. Data pre-processing

The process begins by extracting both scientific articles from the Web of Science (WoS) and patent data from the Derwent Innovation Index database (DII). The title and abstract fields of the articles are combined first as a corpus, and all the other typical bibliometric fields are stored separately. The same happens to the patents – the titles and abstracts are merged into a corpus and the other information, such as patent number, authors and assignees, is stored in a separate file. Words in the corpora are then lemmatized, then all the text is cleaned and consolidate before topic modeling. The lemmatization strategy includes: (1) returning plural nouns to single form; (2) turning inflected forms of verbs back to their stem; (3) returning comparative adjectives back to their basic form. In addition, terms that provide limited contribution to research topics are eliminated by removing all the punctuation, non-alphabetic characters, stop words and words commonly used in scientific literature (Chen et al., 2021).

To identify the entities in the academic and industrial communities, we further pre-processed the author affiliations in the articles and the assignee of the patents. Then the academic institutions and industrial organizations were recognized and tagged. We provide two definitions for natural language processing (NLP) and data cleaning purposes, as follows:

Definition 1. industrial organizations are private firms, public enterprises, or non-profit organizations that provide products or services to society. Academic organizations include universities, higher education providers, research institutions, and academic research laboratories that are dedicated to education and research.

We first tagged academic organizations from common terms that identifying this as their sector – for example, “University”, “Institution”, “School”, “College”, “Faculty” and so forth. Organizations were tagged as industrial if their names contained “Ltd” (Limited), “Co.” (Company), and so on. Some of the addresses were not written in English; others were abbreviated or had other problems.[1] These records were not tagged by the NLP module but rather were passed to a batch for manual checking.

Definition 2. After tagging based on *Definition 1*, the articles and patents were categorized into following main types:

- Academia-dominant item: all of the authors/assignees are with academic organizations.
- Industry-dominant item: all of the authors/assignees are with industrial organizations.
- Academic-industry collaboration item: at least one author/assignee is from an industrial organization, regardless of order of authorship.

[1] If the address information provided by WoS is not sufficient to recognize the type an organization, it will be tagged as “Error”. Ideally, a prepared academic institution list will be helpful for this NLP process, however, building this list is beyond the scope of this paper.

- Error record: one or more of the affiliations in the document could not be determined.

3.2. Topic modeling and parameters setting

Topics were extracted using one of the most accepted topic modeling techniques – latent Dirichlet allocation (LDA). LDA generates a discrete distribution of words for each topic and also a distribution of topics for each document. This generation process can be represented by a joint distribution of random variables (Blei, 2012):

$$p\left(\overrightarrow{w}_d, \overrightarrow{z}_d, \overrightarrow{\vartheta}_d, \varphi \middle| \overrightarrow{\alpha}, \overrightarrow{\beta}\right) = \Pi_{(n=1)}^{(N_d)} p\left(w_{(d,n)} \middle| \overrightarrow{\phi}_{z_{d,n}}\right) p\left(z_{(d,n)} \middle| \overrightarrow{\vartheta}_d\right) p\left(\overrightarrow{\vartheta}_d \middle| \overrightarrow{\alpha}\right) p\left(\varphi \middle| \overrightarrow{\beta}\right), \tag{1}$$

where the observable term number of the d^{th} document in the overall dataset D is defined as N_d, and $w_{d,n}$ is the n^{th} word in the document d. $\overrightarrow{Z}_d$ represents the topic assignment for the document d, $\overrightarrow{\theta}_d$ is the corresponding topic ratio, and $\mathrm{Z}_{d,n}$ represents the topic assignment of the n^{th} word in the d^{th} document. We assume there are K topics that can be denoted with $\overrightarrow{\phi}_{1:k}$, and each $\overrightarrow{\phi}_k$ is a distribution of words. α and β are hyperparameters that determine the amount of smoothing applied to the topic and word distributions (Heinrich, 2005).

When applying LDA to large-scale scientific text mining, one of the most important concerns is the tuning of parameter K. This is especially so when prior knowledge of the target area is limited. Simply guessing the setting for K will reduce the reliability of the result. Further, statistical approaches, such as perplexity or likelihood tend, to lead to a comparatively large number of topics, which is challenging for humans to interpret (De Battisti et al., 2015). Addressing both these concerns, we balanced statistical evaluation with human interpretation by measuring how well the trained model fit the dataset with the perplexity scores, as shown in Eq. (2). At the same time, we applied an exponential function to describe the interpretation complexity of the model, as shown in Eq. (3).

$$Perplexity(D) = exp - \frac{\sum_{d=1}^{M} \log(p(w))}{\sum_{d=1}^{M} N_d} \tag{2}$$

$$complexity(K) = exp\left(\frac{K - \min(K)}{\max(K) - \min(K)}\right) \tag{3}$$

Perplexity is defined as the reciprocal geometric mean of the likelihood (Huang et al., 2018). N_d in (2) stands for the document length of d in D, which has M documents in total, and $\sum \log(p(w))$ represents the likelihood of the corpus given the trained model with a specific value of K. A lower perplexity score indicates a lower misrepresentation of the words in the dataset (De Battisti et al., 2015). We preferred a model with comparatively smaller complexity score to make the results easier to interpret. The final setting for the parameter K was selected when the sum value of normalized perplexity and complexity score reached its minimum:

$$\arg\min_K f(K) = \frac{perplexity(K) - \min\, perplexity(K)}{\max\, perplexity\ (K) - \min\, perplexity(K)} + \frac{complexity(K) - \min\, complexity(K)}{\max\, complexity(K) - \min complexity(K)} \tag{4}$$

The extracted K topics are meaningful decompositions of the corpus. The topic distribution is presented as a matrix Θ, in which $\vartheta_{i,j}$ represents the proportion of document d_i that contributes to topic j . The tagged 'academic-industrial collaborative' and 'pure academic' types in Definition 2 were considered as category labels. We then computed the average 'pure academic' participation per topic, denoted as $A_j(0 \le j \ge K)$, which shows "how academia-driven the topic is". Next, we calculated the average industrial contribution per topic, denoted as $I_j(0 \le j \ge K)$, which represents "to what degree industry has participated the research of this topic". Finally, we calculated the average 'academic-industry collaboration' for each topic, denoted as $C_j(0 \le j \ge K)$. This effectively measured collaborative effort, as per Eq. (5).

$$A_j = \frac{\sum_{i=1} \gamma \vartheta_{ij}}{\sum \gamma}; I_j = \frac{\sum_{i=1} \tau \vartheta_{ij}}{\sum \tau}; C_j = \frac{\sum_{i=1} \omega \vartheta_{ij}}{\sum \omega}; \tag{5}$$

$$\gamma = \begin{cases} 1, d_i \in academa - driven\ items \\ 0,\ \ d_i \notin academa - driven\ items \end{cases}, 0 \le j \ge K$$

$$\tau = \begin{cases} 1, d_i \in industry - driven\ items \\ 0,\ \ d_i \notin industry - driven\ items \end{cases}, 0 \le j \ge K$$

$$\omega = \begin{cases} 1, d_i \in academic - industrial\ collaborative\ items \\ 0,\ \ d_i \notin academic - industrial\ collaborative\ items \end{cases}, 0 \le j \ge K$$

3.3. Topic vectorization with word embedding

Word embedding has attracted increasing attention in scientific text mining research for its capability of capturing contextual and relational semantics from large-scale unstructured text data (Greiner-Petter et al., 2020). In this paper, we used Word2Vec to vectorize the extracted topics into fixed-length numeric vectors. Word2Vec is one of the most efficient and effective word embedding methods (Mikolov et al., 2013). We use the Skip-gram model via the Genism toolkit in this paper (Rehurek and Sojka, 2010).

Skip-gram predicts a surrounding context given a target word. Given a sequence of words D, $D = \{w_{i-k}, \ldots, w_{i-1}, w_i, w_{i+1}, \ldots, w_{i+k}\}$. Here, w_i stands for a target word and k is the context size (window size) of the target word. The window size is denoted as $S = 2k + 1$. Summarizing the method proposed by Mikolov et al. (2013) and based on the notations given by Zhang et al. (2018), the main objective is to maximize the average log probability $L(D)$ of the Skip-gram model, where the probability $Pr(w_{i+c}|w_i)$ is formulated with a softmax function:

$$L(D) = \frac{1}{\phi} \sum_{i=1}^{\phi} \sum_{-k \le c \le k, c \ne 0} \log Pr(w_{i+c}|w_i) \tag{6}$$

$$Pr(w_{i+c}|w_i) = \frac{\exp(x_{i+c} \cdot x_i)}{\sum_{w \in W} \exp(x \cdot x_i)} \tag{7}$$

where ϕ is the size of corpus, and x_{i+c} is the vector representation of a context word for the target word w_i.

We set the dimensions of the word vectors with a parameter γ, and adopt a negative sampling technique to train $L(D)$. This technique was also taken from the Genism toolkit. For each unique term within the corpus (in total there were ϕ), the model returns a γ-dimensional vector. The output of this step is denoted as a $\gamma \times \phi$ matrix that keeps all contextual and relational semantics of the target corpus.

Topics in this study are bibliometric-featured word distributions providing semantically meaningful decompositions of the target corpus, which preserves the global semantics. The word vectors reflect the co-occurrence of words in the context of the content to retain detailed contextual semantics. To map the topics to vectors for further topic network construction, the $\gamma \times \phi$ matrix is integrated with the results of the topic modeling. For each topic, the top n words with the highest proportions are deemed to be the ones that have contributed most to the topic. As such, these terms represent the main content of the topic. As shown in Fig. 1, we calculated the weighted average value of all vectors of the "contributor" words to obtain a unique vector representation for the topic. The weight vector was set to the normalized word proportions of LDA output. This topic vectorization method finally generates K topic vectors from the textual data, laying a solid foundation for topic network construction in the next step.

3.4. Topic network construction and analysis

After topic modeling and vectorization, we have K topic vectors derived from the core constituents of the featured topics. Their correlation implies semantic relation in a vector space. This study then constructs a topic network for quantitatively measuring semantic structure and characteristics in academic and industrial collaborations.

Fig. 2 shows the topic network construction. The topic network has bibliometric-enhanced topics as nodes and semantic correlations of topics as ties. Pearson's correlation coefficient was used to build a topic-correlated matrix TC of different topics via Eq. (8), where T_p and T_q are the vectors of topics p and q. When constructing the network, we emphasized the strong links by maintaining ties with correlated values larger than the upper quartile of the normalized matrix TC.

$$TC\left(T_p, T_q\right) = \frac{\sum_{i=1}^{\gamma}\left(T_{pi} - \overline{T_p}\right)\left(T_{qi} - \overline{T_q}\right)}{\sqrt{\sum_{i=1}^{\gamma}\left(T_{pi} - \overline{T_p}\right)^2}\sqrt{\sum_{i=1}^{\gamma}\left(T_{qi} - \overline{T_q}\right)^2}} \tag{8}$$

Examining the structural properties of this topic network yields comprehensive insights for future academic-industrial research collaborations. In this research, we mostly analyze the degree and betweenness centrality to extract such insights. Guan et al. (2017) provide sound reasons for selecting the above two metrics to indicate the potential of a focal node combined with other elements and illustrate its combinatorial opportunities. Inspired by their work, we used degree and betweenness centrality to measure the structural characteristics and topic interactions of the constructed networks (Kong et al., 2019; Wasserman and Faust, 1994).

4. Empirical study: academic-industrial collaborations in digitalization research

4.1. Data

To demonstrate the feasibility of the methodology, we select digitalization as a target area to conduct an empirical case study. Digitalization is not only an increasingly important skill for industrial communities to have to improve their problem-solving capabilities and overall flexibility, it also serve as a stimulus for both basic and applied research (Sousa and Rocha, 2019; Hess et al., 2016). Thus, digitalization has drawn considerable attention by both industry and academia in recent years.

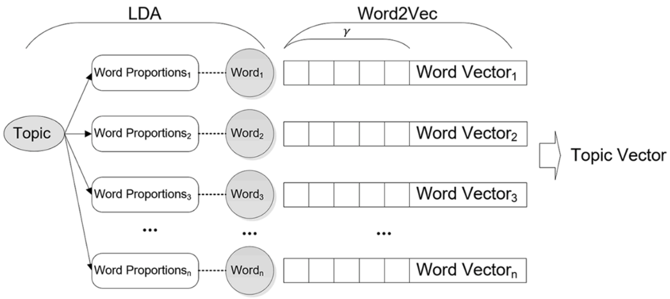


Fig. 1. Topic vectorization process.

The concept of digitalization broadly covers the technical process of converting information into digital form (digitization), the business process of the technological change within organizations or industries (digitalization) and the overall societal effect, measure, and patterns (digital transformation) (Khan, 2016). Based on the generalized definition above, we retrieved a 10-year article dataset from the Web of Science (WoS), and a 10-year patent dataset with a strategy of searching title, abstract and keywords fields for terms of "digitalization", "digitalization", "digitization", "digitization", "digital business transformation" and "digital transformation". The time spanned the years 2010 to 2019.[2] To ensure the best quality and integrity of the affiliation data, we focused on papers with document type of "article" only. With these parameters, we retrieved 4166 valid articles and 3132 valid patents to make up our corpus.[3] As shown in Fig. 3, the number of articles and patents on digitalization research and application has both been increasing in general, which confirms the growing interest in this field in recent years. Notably, the number of research articles has seen sharper growth than patents over the last 3 years.

Based on *Definition 1* in Section 3, the NLP software tagged the academic and industrial affiliations. From a traditional bibliographic coupling perspective, we investigated the main collaborative relationships of the WoS article corpus and the DII patent corpus in Figs. 4(a) and (b), respectively. This revealed a relatively full picture of the academic-industrial engagement of digitalization research and applications. All academic institutions are marked in green, and the industrial ones are marked in red. The node size reflects the number of links with other items.[4]

As illustrated in Fig. 4(a), the majority of organizations participating in digitalization research are academic institutions. The collaborations of participants in academia and industry generally follow patterns of: (1) industrial organizations like "VTT Tech Res Ctr Finland Ltd" on the left upper corner who have developed extensive cooperation with different universities; (2) a number of non-profit (industrial) organizations leading by "Hat Hist Museum" working together closely as a red cluster; and (3) enterprises like "Siemens AG" choosing to "work alone" on digitalization research, as shown in the right lower corner. Patents in this area show completely different collaborative patterns at the institutional level. Fig. 4(b) presents "State Grid Corp China" as the most collaborative actor in the network, while universities and research institution like "Univ Hangzhou Dianzi" and "China Electric Power Res Inst" are mainly working with several focal industrial organizations stably, with no extensive cooperation.

4.2. Bibliometric-enhanced topic modeling

We then lemmatized words in the corpus and cleaned and consolidated the dataset. Following *Definition 1* and *Definition 2* given in Section 3, we tagged the records in both the article and the patent corpus into academia-dominant items, industry-dominant items, and academic-industrial collaborative items.

Aiming to receive an interpretable topic modeling result, we ran multiple LDA experiments with the number of total topics, setting K to

[2] The field of timespan setting we applied for patent retrieval is Basic Patent Year, which is the first time an invention is collected in the DII.

[3] 4172 articles are first retrieved followed the search strategy, in which 33 articles have no affiliations; 3134 patents are first retrieved, 2 of them are duplicated.

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

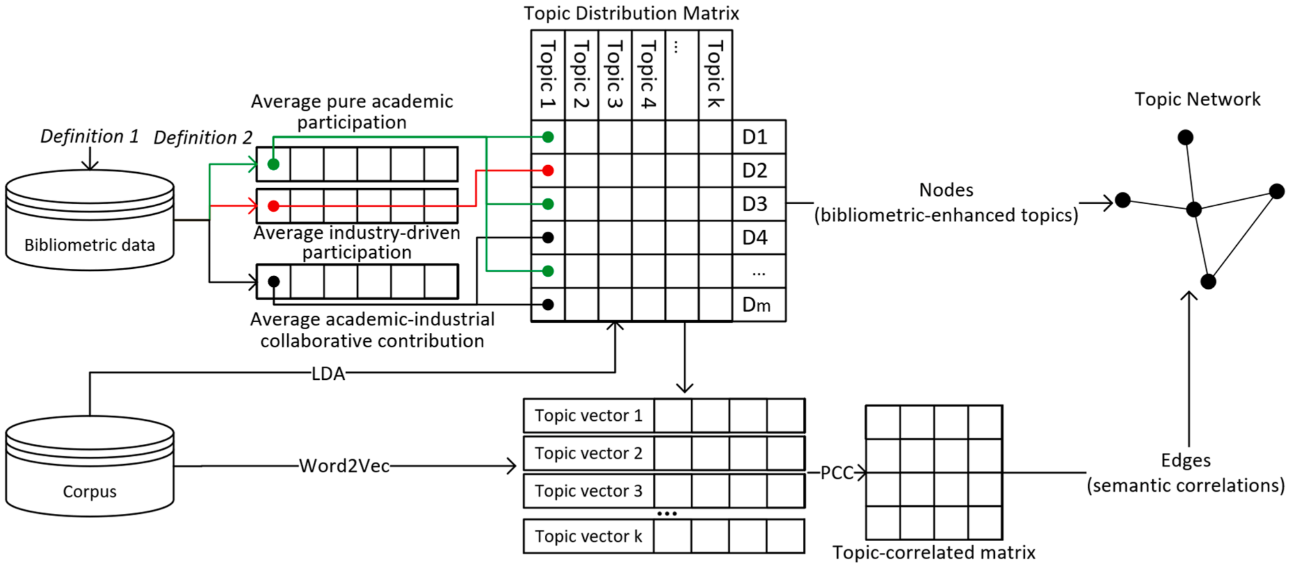

Topic Distribution Matrix
Topic 1
Topic 2
Topic 3
Topic 4
…
Topic k
D1
D2
D3
D4
…
Dm
Average pure academic participation
Definition 1
Definition 2
Bibliometric data
Average industry-driven participation
Average academic-industrial collaborative contribution
LDA
Corpus
Word2Vec
Topic vector 1
Topic vector 2
Topic vector 3
Topic vector k
PCC
Topic-correlated matrix
Nodes (bibliometric-enhanced topics)
Topic Network
Edges (semantic correlations)

Fig. 2. The schematic diagram of topic network construction process.

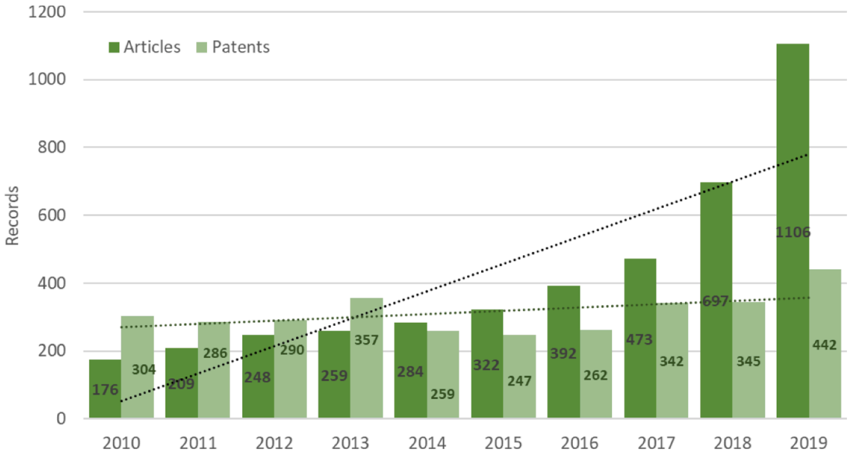


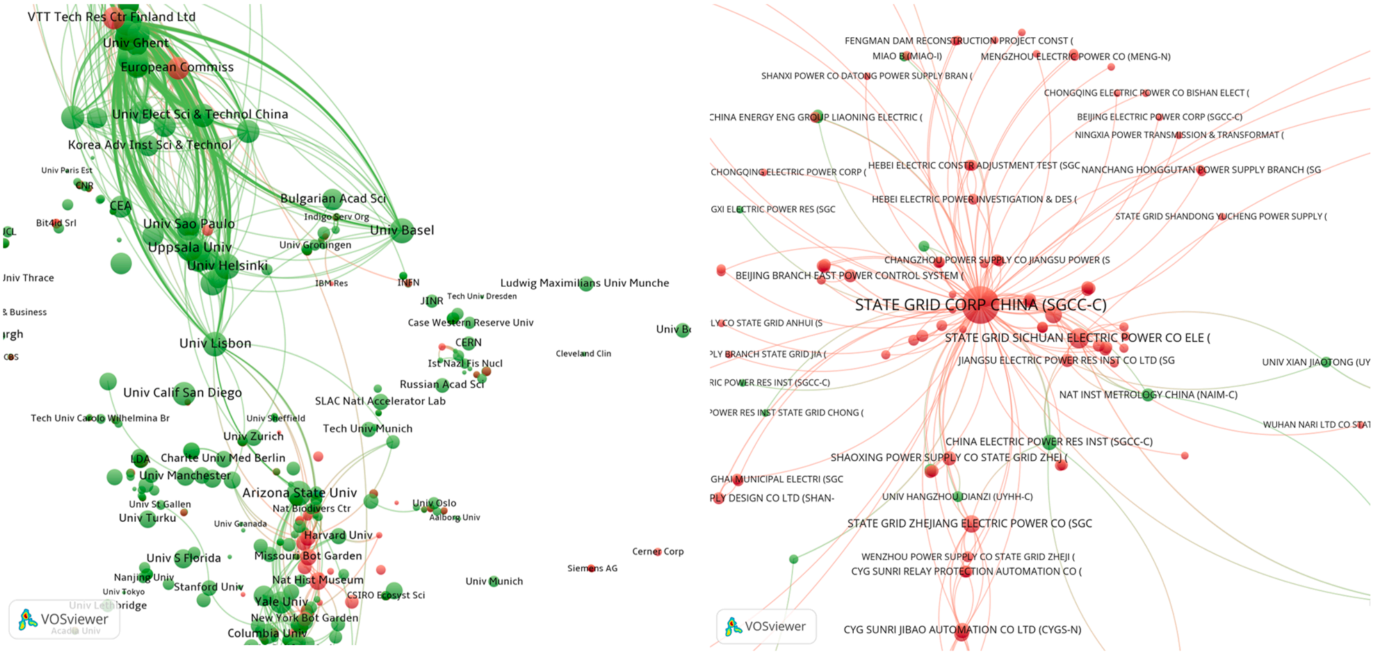
Fig. 3. Number of "digitalization" articles and patents published on between 2010 and 2019.

different values. To balance statistical evaluation with interpretability, we ultimately set the parameter K to its minimum of 25 for both corpora, as shown in Figs. 5(a2) and (b2). To provide a fine-grained decomposition of the document collection, we follow the common settings for hyperparameters α and β in existing research, and set $\alpha = 0.5$ and $\beta = 0.01$ (Griffiths and Steyvers, 2004). We then proceeded to apply 10, 000 iterations of Gibbs sampling to infer the latent variables and distributions (Heinrich, 2005).

Following the balanced parameter settings, we extracted topics for the article and patent corpus respectively. Table 1 summarizes all the topics discovered in the WOS articles and the DII patents with the topic numbers for the article-derived topics (A-Topic#), the topic numbers for patent-derived topics (P-Topic#), the Pearson Correlation Coefficient (PCC) values, topic labels and their main content,[5] plus the total topic proportion to present weight (total topic proportion) (Chen et al., 2021). These two groups of topics highlight the main themes of digitalization studies from a research perspective and an application perspective. Each has a different emphasis, but they do have underlying topic connections. To examine whether research articles and patents shared same topics in the target area, we followed the topic vectorization module of our proposed methodology and trained the word vectors on a combined corpus of articles and patents via Genism toolkit. From this, we constructed a unified vector space. Considering the configurations used in previous studies (Zhang et al., 2018), we set the window size to 10 and the dimension of the vectors to 100. The identified topics were word distributions providing semantically meaningful decompositions of a target corpus. These topics were then vectorized into corresponding vectors in the unified vector space. This makes topic similarity detection possible. We received 25 vectors for article-derived topics and 25 vectors for patent-derived topics, both with 100 dimensions (latent features). For the purposes of the measuring the topic correlations, we calculated Pearson's correlation coefficients and built a matrix of article-derived and patent-derived topics. From this we kept only correlated values greater than the upper quartile to highlight strong correlations.

The article-derived topics cover the main concerns, concepts, skills, tools, and applications of digitalization research. Hotspots like industry 4.0 and technologies inherent in transforming images, motions, and signals into cyber systems have drawn considerable attention in recent years – especially for their ability to connect the physical and digital world. The patent-derived topics mainly focused on the detailed techniques and devices of transforming inputs into digitalized outputs. The topic correlations revealed seven pairs of topics common to both the

[5] Due to the space limitation, Table 1 only shows 5 of the top 10 words that contribute most to each topic.


VTT Tech Res Ctr Finland Ltd
Univ Ghent
European Commiss
Univ Elect Sci & Technol China
Korea Adv Inst Sci & Technol
Univ Paris Est
CNR
CEA
Bulgarian Acad Sci
Indigo Serv Org
Bit4id Srl
Univ Sao Paulo
Univ Basel
Uppsala Univ
Univ Groningen
Univ Helsinki
Univ Thrace
IBM Res
INFN
Ludwig Maximilians Univ Munche
JINR
Tech Univ Dresden
& Business
Case Western Reserve Univ
Univ Lisbon
CERN
Cleveland Clin
Ist Nazl Fis Nucl
Russian Acad Sci
Univ Calif San Diego
SLAC Natl Accelerator Lab
Tech Univ Carolo Wilhelmina Br
Univ Sheffield
Tech Univ Munich
Univ Zurich
Charite Univ Med Berlin
Univ Manchester
Arizona State Univ
Univ St Gallen
Nat Biodivers Ctr
Univ Oslo
Univ Turku
Aalborg Univ
Univ Granada
Harvard Univ
Cerner Corp
Univ S Florida
Missouri Bot Garden
Siemens AG
Nanjing Univ
Nat Hist Museum
Univ Munich
Univ Tokyo
Stanford Univ
CSIRO Ecosyst Sci
Univ Lethbridge
Yale Univ
New York Bot Garden
Columbia Univ
VOSviewer
FENGMAN DAM RECONSTRUCTION PROJECT CONST (
MIAO B (MIAO-I)
MENGZHOU ELECTRIC POWER CO (MENG-N)
SHANXI POWER CO DATONG POWER SUPPLY BRAN (
CHONGQING ELECTRIC POWER CO BISHAN ELECT (
CHINA ENERGY ENG GROUP LIAONING ELECTRIC (
BEIJING ELECTRIC POWER CORP (SGCC-C)
NINGXIA POWER TRANSMISSION & TRANSFORMAT (
CHONGQING ELECTRIC POWER CORP (
HEBEI ELECTRIC CONSTR ADJUSTMENT TEST (SGC
NANCHANG HONGGUTAN POWER SUPPLY BRANCH (SG
HEBEI ELECTRIC POWER INVESTIGATION & DES (
GXI ELECTRIC POWER RES (SGC
STATE GRID SHANDONG YUCHENG POWER SUPPLY (
CHANGZHOU POWER SUPPLY CO JIANGSU POWER (S
BEIJING BRANCH EAST POWER CONTROL SYSTEM (
STATE GRID CORP CHINA (SGCC-C)
LYCO STATE GRID ANHUI (S
STATE GRID SICHUAN ELECTRIC POWER CO ELE (
PLY BRANCH STATE GRID JIA (
JIANGSU ELECTRIC POWER RES INST CO LTD (SG
UNIV XIAN JIAOTONG (UY
RIC POWER RES INST (SGCC-C)
NAT INST METROLOGY CHINA (NAIM-C)
POWER RES INST STATE GRID CHONG (
WUHAN NARI LTD CO STAT
CHINA ELECTRIC POWER RES INST (SGCC-C)
SHAOXING POWER SUPPLY CO STATE GRID ZHEJ (
GHAI MUNICIPAL ELECTRI (SGC
PLY DESIGN CO LTD (SHAN-
UNIV HANGZHOU DIANZI (UYHH-C)
STATE GRID ZHEJIANG ELECTRIC POWER CO (SGC
WENZHOU POWER SUPPLY CO STATE GRID ZHEJI (
CYG SUNRI RELAY PROTECTION AUTOMATION CO (
VOSviewer
CYG SUNRI JIBAO AUTOMATION CO LTD (CYGS-N)

(a) (b)

Fig. 4. Organizational level academic-industrial collaborations in digitalization research and application; (a) shows the main collaborative relationships of organizations in the WOS articles corpus; (b) illustrates the main collaborative relationships of organizations in DII patents corpus. The above figures were drawn using VOSviewer.

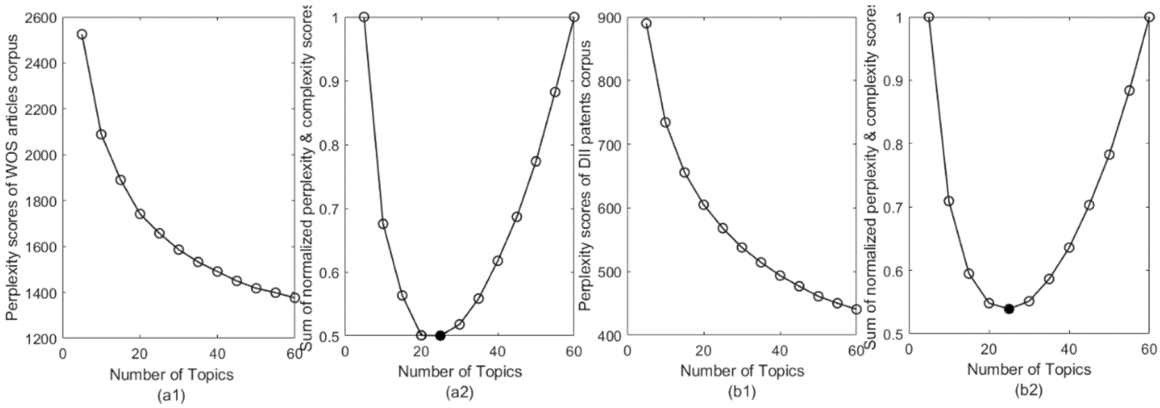


Fig. 5. Parameter setting for topic modeling: (a1) and (b1) shows the perplexity values of WOS article corpus and DII patent corpus respectively; (a2) and (b2) shows the sum values of normalized perplexity and complexity scores, with different setting of topic number *K,* for two corpora respectively.

articles and patents. These included digital service (A-T08, P-T23), signal frequency (A-T12, P-T01), 3D (A-T14, P-T11), sensors (A-T15, P-T10), image recognition (A-T18, P-T16), diagnosis (A-T21, P-T18) and measurement (A-T22, P-T08). Although each pair of topics shares similar content, they have different focuses from perspectives of research or application. For example, the article-derived topic of digital service (A-T08) is heavily weighted toward business innovation, highlighting company-related services, while the patent-derived topic of digital service (P-T23) mainly focuses on products, such as “block” and “payment”; the article-derived topic of image recognition place emphasis on algorithms of recognition, while patent-derived image recognition cares more about tools of image capture.

We then featured all topics using method provided in subSection 3.2, and quantitatively measured “how academia-driven/industry-driven/collaborative a topic is” after computing the average academic participation, industrial participation, and academic-industrial collaborative participation for each topic. The last column of Table 1 shows the collaborative characteristics of extracted topics, in which “A” indicates academia-driven, “I” stands for industry-driven, and “C” means collaborative. As shown in Table 1, article-derived topics are mainly academia-driven or collaborative, while the majority of the patent-derived topics are industry-driven or academia-driven. Although the research articles and patents share some common topics in this area, industry and academia have participated these topics to different degrees, and their main concerns are diverse. For example, “diagnosis” is one the most important themes of both digitalization academic research and applications. From a research perspective, it is a collaborative topic with main content “image” being highlighted, yet, from an applications perspective, it is an academia-driven topic with main content “ultrasonic” as the focus. Another example is topic “image recognition”. From research perspective, it is a collaborative topic with main content being “recognition algorithm”. Yet, from patent-derived perspective, it is an

Table 1
Topics extracted from WOS articles and DII patents.

A-topic#	P-topic#	PCC	Topic Label	Main Content	Weight	Type
A-T01	–	–	motion measure	muscle, angle, bone, motion, measure, facial	123.80	C
A-T02	–	–	digital copyright	digitization, digital, music, copyright, law	138.52	A
A-T03	–	–	online marketing	online, market, consumer, share, mobile	150.74	A
A-T04	–	–	security	network, control, security, architecture, risk	165.83	C
A-T05	–	–	electronics	detector, readout, pulse, resolution, electronics	175.04	C
A-T06	–	–	specimen collection	collection, specimen, specie, digitization, herbarium	154.21	I
A-T07	–	–	sample	sample, material, surface, flow, temperature	138.39	C
A-T08	P-T23	0.78	digital service (business)	business, innovation, service, digital, company	262.92	A
A-T09	–	–	digital education	medium, digital, education, learn, communication	163.74	A
A-T10	–	–	algorithm	model, digitization, structure, space, algorithm	150.44	A
A-T11	–	–	industry-4.0	manufacture, production, smart, industry-4.0, big-data	197.06	I
A-T12	P-T01	0.96	signal frequency (noise)	signal, frequency, digitization, noise, optical	156.16	C
A-T13	–	–	digital scan	scan, digital, impression, scanner, accuracy	129.13	C
A-T14	P-T11	0.77	3D (image)	3d, image, surface, scan, digitization, reconstruction	180.49	C
A-T15	P-T10	0.76	sensor (circuit)	sensor, power, circuit, converter, analog	168.73	C
A-T16	–	–	public service	public, development, service, country, economy	200.58	A
A-T17	–	–	map	map, land, spatial, site, satellite	147.74	C
A-T18	P-T16	0.95	image recognition (algorithm)	image, algorithm, automate, document, recognition	176.52	C
A-T19	–	–	pattern recognition	pattern, shape, human, tree, size	113.78	C
A-T20	–	–	social practice	digital, social, practice, technology, digitalization	235.32	A
A-T21	P-T18	0.78	diagnosis (image)	image, patient, tissue, cancer, diagnosis	107.64	C
A-T22	P-T08	0.78	measurement (performance)	result, measure, test, error, performance	194.94	C
A-T23	–	–	software application	tool, software, model, development, application	204.63	C
A-T24	–	–	electronic healthcare	health, care, patient, medical, electronic	140.86	C
A-T25	–	–	digital library	library, heritage, digitization, historical, preservation	188.80	A
P-T01	A-T12	0.96	signal frequency (receiver)	signal, frequency, radio, receiver, wave	131.72	I
P-T02	–	–	utility device	plate, box, board, utility, device	152.23	I
P-T03	–	–	signal O and I	signal, digital, output, input, analog	171.34	I
P-T04	–	–	circuit	circuit, power, voltage, control, supply	151.72	I
P-T05	–	–	digitization	sample, detector, pulse, digitization, digitize	114.03	I
P-T06	–	–	document digitize	document, digitize, medium, digitization, description	125.01	I
P-T07	–	–	equipment	unit, device, equipment, receive, storage	136.08	I
P-T08	A-T22	0.78	measurement (calculation)	measure, measurement, calculate, determine, range	93.36	A
P-T09	–	–	digital transformer	transformer, digital, substation, protection, network	164.22	C
P-T10	A-T15	0.76	sensor (device)	sensor, pressure, temperature, water, device	116.03	A
P-T11	A-T14	0.77	3D (virtual)	model, three-dimensional, weld, construction, virtual	114.02	A
P-T12	–	–	vehicle	motor, machine, vehicle, mechanism, control	135.72	I
P-T13	–	–	electric detect	electric, test, energy, digital, detection	108.61	C
P-T14	–	–	electronic-file management	electronic, file, management, server, medical	114.39	I
P-T15	–	–	platform	module, management, platform, intelligent, control	136.09	A
P-T16	A-T18	0.95	image recognition (capture)	image, camera, pixel, color, capture	124.96	I
P-T17	–	–	video processor	video, processor, card, screen, audio	106.24	I
P-T18	A-T21	0.78	diagnosis (ultrasonic)	patient, cell, diagnosis, ultrasonic, tissue	79.94	A
P-T19	–	–	language detection	apparatus, digitization, text, detect, language	147.05	I
P-T20	–	–	communication	monitor, communication, network, terminal, wireless	137.74	I
P-T21	–	–	laser	light, optical, surface, source, laser	109.20	A
P-T22	–	–	digital print	print, layer, material, transfer, digital	99.58	I
P-T23	A-T08	0.78	digital service (product)	digital, service, product, block, payment	113.00	I
P-T24	–	–	parameter sensor	parameter, sense, structure, environment, schematic	89.30	A
P-T25	–	–	text	involve, text, language, non-english, perform	160.43	A

industrial topic with main content “image capture” and “camera pixel” being highlighted. These bibliometric-enhanced topics not only shows the research and application hotspots in the target area as a full picture, but also highlight the collaborative characteristics of these focal elements.

4.3. *Topic network construction and analytics*

To further analyze co-addressed topics of the academic and industrial communities, we mapped the topics into topic networks and quantitatively measured their semantic structures and characteristics. To provide a more comprehensive view of the knowledge flows at play, we did not limit the scale of the networks to the academic-industry collaboration items. We also included papers written by only academic authors or industrial authors.

Featured topics are recognized as vertices of the topic networks. Emphasizing semantic correlations among each group of topics that reveal the main themes of digitalization studies, we trained the word vectors on the article corpus and the patent corpus separately. Following the configurations used in previous studies, we set the window size to 10 and the dimension of vectors to 100. For each corpus, we received 25 vectors of 100 dimensions (latent features), and further applied Pearson’s correlation coefficients to calculate the semantic correlations between different topics as the network’s edges. To highlight strong links, we only kept links with correlated values larger than the upper quartile of the normalized matrix.

Fig. 6 presents the topic networks derived from the corpora. Here, the node color reflects the collaborative characteristics of the topic. The colors and shapes convey information about what type of topic each node represents. Academia-driven topics are colored in green, industry-driven ones are colored in red and collaborative ones are colored in black. A triangle node shape indicates that the topic is shared across the two corpora. The links mean the topics are correlated.[6]

We then calculated degree and betweenness centrality to measure

[6] http://www.analytictech.com/Netdraw/netdraw.htm

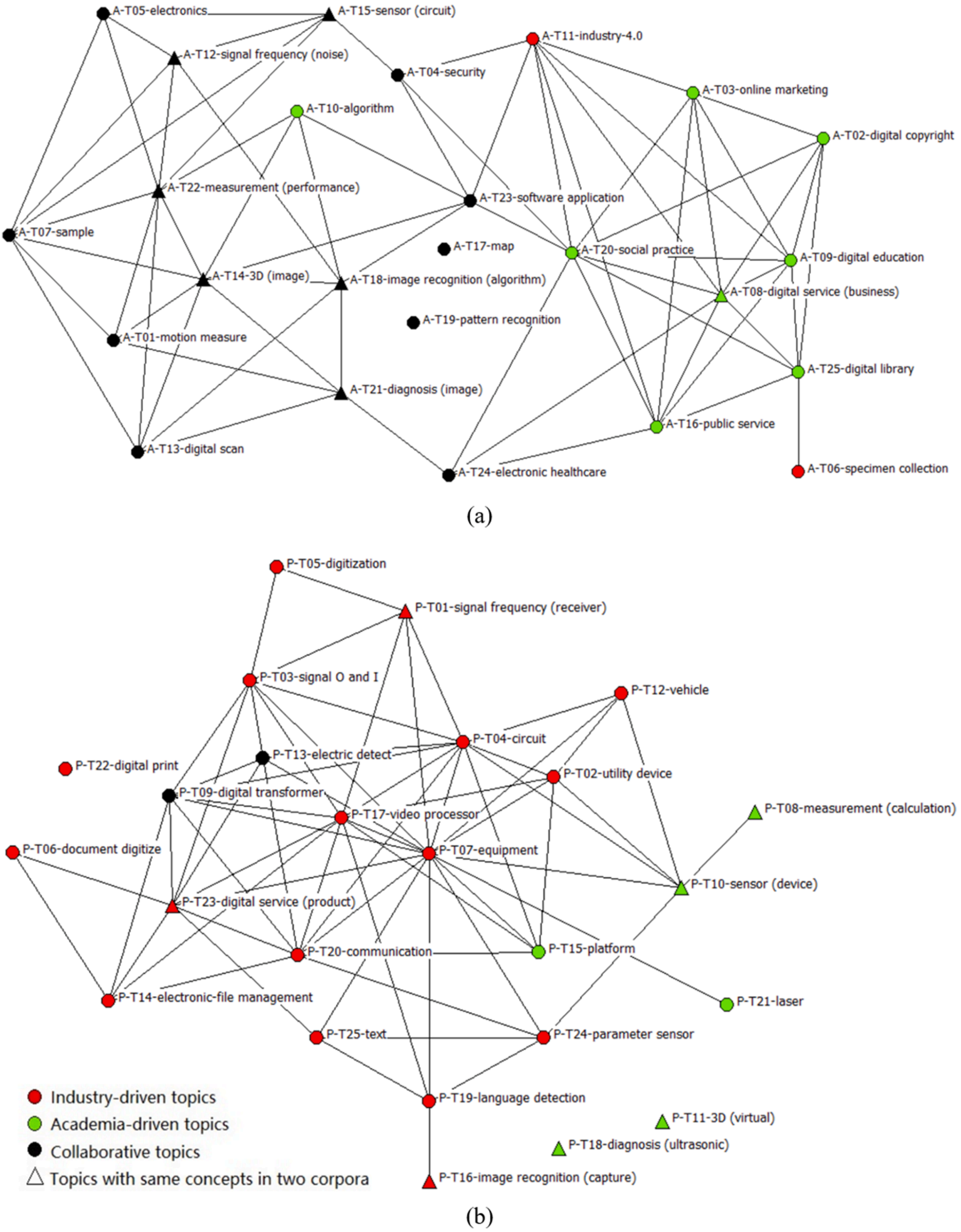


Fig. 6. Topic networks derived from (a) WOS article and (b) DII patent corpus. The above figures were drawn using NetDraw.

Table 2
Degree and betweenness centrality of article-derived and patent-derived topics.

A-topics Label	Degree	Betweenness	P-topics label	Degree	Betweenness
A-T01-motion measure	1.98	2.66	P-T01-signal frequency (receiver)	2.91	6.92
A-T02-digital copyright	2.66	0.40	P-T02-utility device	2.64	2.35
A-T03-online marketing	3.55	0.45	P-T03-signal O and I	4.39	13.69
A-T04-security	1.94	31.47	P-T04-circuit	5.71	16.85
A-T05-electronics	2.44	0.00	P-T05-digitization	0.96	0.00
A-T06-specimen collection	0.53	0.00	P-T06-document digitize	0.94	0.00
A-T07-sample	3.46	6.43	P-T07-equipment	8.42	75.84
A-T08-digital service (business)	4.72	7.72	P-T08-measurement (calculation)	0.4	0.00
A-T09-digital education	3.96	1.35	P-T09-digital transformer	3.94	3.28
A-T10-algorithm	2.08	3.73	P-T10-sensor (device)	2.62	21.33
A-T11-industry-4.0	3.91	19.34	P-T11–3D (virtual)	0.00	0.00
A-T12-signal frequency (noise)	2.99	3.19	P-T12-vehicle	1.92	0.00
A-T13-digital scan	2.69	3.25	P-T13-electric detect	2.84	0.56
A-T14–3D (image)	4.44	23.22	P-T14-electronic-file management	2.29	3.57
A-T15-sensor (circuit)	3.00	25.64	P-T15-platform	2.66	0.25
A-T16-public service	4.28	5.69	P-T16-image recognition (capture)	0.43	0.00
A-T17-map	0.00	0.00	P-T17-video processor	5.44	23.47
A-T18-image recognition (algorithm)	3.21	14.48	P-T18-diagnosis (ultrasonic)	0.00	0.00
A-T19-pattern recognition	0.00	0.00	P-T19-language detection	2.48	20.67
A-T20-social practice	5.94	62.99	P-T20-communication	4.39	8.46
A-T21-diagnosis (image)	2.73	23.31	P-T21-laser	0.43	0.00
A-T22-measurement (performance)	3.97	10.47	P-T22-digital print	0.00	0.00
A-T23-software application	3.06	52.07	P-T23-digital service (product)	4.26	21.14
A-T24-electronic healthcare	1.85	22.95	P-T24-parameter sensor	2.2	4.89
A-T25-digital library	3.27	21.20	P-T25-text	1.89	1.75

the structural characteristics of the network using UCINET.[7] Degree centrality represents the interconnectedness of the nodes in terms of communication activities. Betweenness centrality indicates how much a topic controls information flow (Kong et al., 2019). Table 2 presents the degree and betweenness centrality values of all topics. Topics with high degree centrality and comparatively lower betweenness centrality, such as A-T08-digital service in Fig. 6(a) and P-T20-communication in Fig. 6 (b), serve as key elements in the networks, as they hold semantic relations with many other nodes. However, they do not really serve as bridges of the information flow. Topics like A-T20-social practice in Fig. 6(a) and P-T07-equipment in Fig. 6(b) on the other hand, play an important role in linking disparate topics. Their ability to control the information flow is stronger than other elements. In addition, although not many nodes are connected to topics like A-T23-software application in Fig. 6(a) or P-T10-sensor in Fig. 6(b), they both have high betweenness centrality, which means they are key pathway and are potentially pivotal in topic splits or mergers.

In terms of the article-derived topics, there are mainly two “communities” as shown in Fig. 6(a): a collaborative topic cluster, and an academia-driven topic cluster. Collaborative topics, marked in black, have comparatively larger betweenness centrality. As shown in Fig. 6 (b), the patent-derived topic network is dominated by industry-driven nodes. The academia-driven topics have much lower betweenness centrality, except for P-T10-sensors.

5. Conclusions and future research

The significance of the collaborations between industry and academia has not only been highlighted by the academic and industrial communities, but also by international organizations and governments. The innovations of these two types of organizations are embedded in both knowledge networks established by coupling among knowledge elements and in collaboration networks of organizations. Yet existing studies have mainly focused on analyzing collaborations at the institutional level; little research has been conducted at the content level. Moreover, much of the research that has been done relies on co-occurring keywords in articles and/or co-application of IPCs to form knowledge network. But these methods can blur the concept of “topics” with “keywords”, “categories” or “tags”. Additionally, simply considering keywords or classifications does not capture the context of the textual content. What we find is that while academia and industry often discuss the same subjects, they do so in a different context.

In this paper, we focused on unearthing content-based collaborative insights between the academic and industrial communities. We have presented an effective method of mapping the extracted topics to a knowledge network and revealed the interactions between those topics along with their network characteristics. The vertices in the network are featured topics, tagged as academia-driven ones, industry-driven ones, and collaborative ones. We then vectorized those topics via word embedding techniques to reveal semantic relations, which serve as the edges of the final network. This method provides a new perspective on generating knowledge networks in a given target area for academic-industrial collaboration research purpose. We discuss not only the academic-industrial collaborative items, but also academia-dominant and industry-dominant knowledge elements simultaneously in the constructed topic networks. As such our findings provide a more complete picture of knowledge flows between the two communities.

This study has several limitations that need be explored in further research. First, we did not take full consideration of the structural properties of the topic network. Examining more structural properties may potentially provide further insight into these collaborations. In addition, when tagging and cleaning organizations, a prepared academic institution list would have been very helpful for our NLP process. Much time was spent manually checking and disambiguating organization details. Moreover, both participants and content interactions present collaborative patterns that can provide a comprehensive understanding on the discovery of future collaborative insights. Although in this paper we provided a systematic method for fully revealing content-based collaborative insights between the academic and industrial communities, further analysis to systematically measure the interactions of topic network and collaboration network as a complex system was not explored. This will be a focus of our studies in our future research. Potentially, cooperative opportunity discovery in multi-source heterogeneous networks will also be one of our future research directions.

[7] http://www.analytictech.com/products.htm

Authors contribution

Hongshu Chen: Conceived and designed the analysis, performed the analysis and wrote the paper.

Qianqian Jin: Collected the data and performed the analysis.

Ximeng Wang: Contributed data and analysis tool and wrote the paper.

Fei Xiong: Conceived and designed the analysis, wrote the paper and provide other contribution.

Acknowledgement

This work was supported by: the National Natural Science Foundation of China under Grant Nos. 72004009 and 61872033; the Beijing Institute of Technology Research Fund Program for Young Scholars and Beijing Nova Program (Z201100006820015) under the Beijing Municipal Science & Technology Commission.

References

Akhtar, Pervaiz, Khan, Zaheer, Rao-Nicholson, Rekha, Zhang, Minhao, 2019. Building relationship innovation in global collaborative partnerships: big data analytics and traditional organizational powers. R&D Manag. 49 (1), 7–20.

Blei, D.M., 2012. Probabilistic topic models. Commun. ACM 55 (4), 77–84.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3 (4–5), 993–1022.

Chang, Shu-Hao, 2017. The technology networks and development trends of university-industry collaborative patents. Technol. Forecast. Soc. Change 118, 107–113.

Chen, Baitong, Tsutsui, Satoshi, Ding, Ying, Ma, Feicheng, 2017. Understanding the topic evolution in a scientific domain: an exploratory study for the field of information retrieval. J. Informetr. 11 (4), 1175–1189.

Chen, H., Wang, X., Pan, S., Xiong, F., 2021. Identify topic relations in scientific literature using topic modeling. IEEE Trans. Eng. Manage. 68 (5), 1232–1244.

Chen, H., Zhang, Y., Zhang, G., Zhu, D. & Lu, J., 2015. Modeling technological topic changes in patent claims. 2015 Portland International Conference On Management of Engineering and Technology (PICMET), 2-6 Aug. 2015, 2049–2059.

De Battisti, F., Ferrara, A., Salini, S., 2015. A decade of research in statistics: a topic model approach. Scientometrics 103 (2), 413–433.

Fan, X., Yang, X.W., Chen, L.M., 2015. Diversified resources and academic influence: patterns of university-industry collaboration in Chinese research-oriented universities. Scientometrics 104 (2), 489–509.

Firth, J.R., 1957. A synopsis of linguistic theory 1930-55. Studies in Linguistic Analysis The Philological Society.

Giunta, A., Pericoli, F.M., Pierucci, E., 2016. University-Industry collaboration in the biopharmaceuticals: the Italian case. J. Technol. Trans. 41 (4), 818–840.

Greiner-Petter, A., Youssef, A., Ruas, T., Miller, B.R., Schubotz, M., et al., 2020. Math-word embedding in math search and semantic extraction. Scientometrics 30.

Griffiths, T.L., Steyvers, M., 2004. Finding scientific topics. Proc. Natl. Acad. Sci. U.S.A. 101, 5228–5235.

Guan, Jiancheng, Liu, Na, 2016. Exploitative and exploratory innovations in knowledge network and collaboration network: a patent analysis in the technological field of nano-energy. Res. Policy 45 (1), 97–112.

Guan, Jiancheng, Yan, Yan, Zhang, Jing Jing, 2017. The impact of collaboration and knowledge networks on citations. J. Informetr 11 (2), 407–422.

Heinrich, Gregor, 2005. Parameter estimation for text analysis: technical Report.

Hellsten, Iina, Leydesdorff, Loet, 2020. Automated analysis of actor–topic networks on twitter: new approaches to the analysis of socio-semantic networks. J. Assoc. Inf. Sci. Technol. 71 (1), 3–15.

Hess, T., Matt, C., Benlian, A., Wiesbock, F., 2016. Options for formulating a digital transformation strategy. MIS Q. Exec 15 (2), 123–139.

Huang, A.H., Lehavy, R., Zang, A.Y., Zheng, R., 2018. Analyst information discovery and interpretation roles: a topic modeling approach. Manage. Sci. 64 (6), 2833–2855.

Huang, M.H., Yang, H.W., Chen, D.Z., 2015. Industry-academia collaboration in fuel cells: a perspective from paper and patent analysis. Scientometrics 105 (2), 1301–1318.

Jung, Sukhwan, Yoon, Wan Chul, 2020. An alternative topic model based on common interest authors for topic evolution analysis. J. Informetr. 14 (3), 101040.

Khan, Shahyan., 2016. Leadership in the digital age: a study on the effects of digitalisation on top management leadership. Independent Thesis Advanced level (degree of Master (Two Years)) Student thesis.

Kong, Xiangjie, Shi, Yajie, Yu, Shuo, Liu, Jiaying, Xia, Feng, 2019. Academic social networks: modeling, analysis, mining and applications. J. Netw. Comput. Appl. 132, 86–103.

Lamba, M., Madhusudhan, M., 2019. Mapping of topics in DESIDOC journal of library and information technology, India: a study. Scientometrics 120 (2), 477–505.

Le, Quoc & Mikolov, Tomas., 2014. Distributed representations of sentences and documents. Proceedings of the 31st International Conference On Machine Learning (ICML-14), 1188–1196.

Lee, H., Kang, P., 2018. Identifying core topics in technology and innovation management studies: a topic model approach. J. Technol. Transf. 43 (5), 1291–1317.

Lee, Y.Y., Ke, H., Yen, T.Y., Huang, H.H., Chen, H.H., 2020. Combining and learning word embedding with WordNet for semantic relatedness and similarity measurement. J. Assoc. Inf. Sci. Technol. 71 (6), 657–670.

Lee, Yong S, 2000. The sustainability of university-industry research collaboration: an empirical assessment. J Technol Transf 25 (2), 111–133.

Levy, Omer, Goldberg, Yoav, Dagan, Ido, 2015. Improving distributional similarity with lessons learned from word embeddings. Trans. Assoc. Comput. 3, 211–225.

Lin, J.Y., 2017. Balancing industry collaboration and academic innovation: the contingent role of collaboration-specific attributes. Technol. Forecast. Soc. Change 123, 216–228.

Magazinik, A., Mäkinen, S.J., Lasheras, N.C., Bedolla, J.S. & Saari, U., 2019. Research-industry collaboration: a review of the literature on evaluation methods and motivations. 2019 Portland International Conference On Management of Engineering and Technology (PICMET), 25-29 Aug. 2019, 1–19.

Mao, C.F., Yu, X.Y., Zhou, Q., Harms, R., Fang, G., 2020. Knowledge growth in university-industry innovation networks - Results from a simulation study. Technol. Forecast. Soc. Change 151, 119746.

Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg S & Dean, Jeff, 2013. Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems, 3111–3119.

Nsanzumuhire, Silas U., Groot, Wim, 2020. Context perspective on University-Industry Collaboration processes: a systematic review of literature. J. Clean. Prod. 258, 120861.

Perkmann, Markus, Neely, Andy, Walsh, Kathryn, 2011. How should firms evaluate success in university–industry alliances? A performance measurement system. R&D Manag. 41 (2), 202–216.

Peters, Hpf, Van Raan, Anthony Fj, 1993. Co-word-based science maps of chemical engineering. Part I: representations by direct multidimensional scaling. Res. Policy 22 (1), 23–45.

Phelps, Corey, Heidl, Ralph, Wadhwa, Anu, 2012. Knowledge, networks, and knowledge networks. J. Manage. 38 (4), 1115–1166.

Rehurek, Radim, Sojka, Petr., 2010. Software framework for topic modelling with large corpora. Proceedings of the LREC 2010 Workshop on New Challenges For NLP Frameworks. Citeseer.

Roth, Camille, Cointet, Jean-Philippe, 2010. Social and semantic coevolution in knowledge networks. Soc. Networks 32 (1), 16–29.

Skute, Igors, Zalewska-Kurek, Kasia, Hatak, Isabella, De Weerd-Nederhof, Petra, 2019. Mapping the field: a bibliometric analysis of the literature on university–industry collaborations. J. Technol. Transf. 44 (3), 916–947.

Song, B., Suh, Y., 2019. Identifying convergence fields and technologies for industrial safety: lDA-based network analysis. Technol. Forecast. Soc. Change 138, 115–126.

Sousa, M.J., Rocha, A., 2019. Digital learning: developing skills for digital transformation of organizations. Future Generation Computer Systems-the International Journal of Escience 91, 327–334.

Suominen, A., Toivanen, H., 2016. Map of science with topic modeling: comparison of unsupervised learning and human-assigned subject classification. J. Assoc. Inf. Sci. Technol. 67 (10), 2464–2476.

Suominen, Arho, Ranaei, Samira, Dedehayir, Ozgur, 2019. Exploration of science and technology interaction: a case study on Taxol. IEEE Trans. Eng. Manage.

Teixeira, Aurora A.C., Mota, Luisa, 2012. A bibliometric portrait of the evolution, scientific roots and influence of the literature on university–industry links. Scientometrics 93 (3), 719–743.

Wang, Chao-Hung, Hsu, Li-Chang, 2014. Building exploration and exploitation in the high-tech industry: the role of relationship learning. Technol. Forecast. Soc. Change 81, 331–340.

Wang, Y.D., Hu, D., Li, W.P., Li, Y.W., Li, Q., 2015. Collaboration strategies and effects on university research: evidence from Chinese universities. Scientometrics 103 (2), 725–749.

Wang, Y.D., Hu, R.F., Li, W.P., Pan, X.F., 2016. Does teaching benefit from university-industry collaboration? investigating the role of academic commercialization and engagement. Scientometrics 106 (3), 1037–1055.

Wasserman, Stanley & Faust, Katherine, 1994. Social network analysis: methods and applications.

Woltmann, Sabrina L., Alkærsig, Lars, 2018. Tracing university–industry knowledge transfer through a text mining approach. Scientometrics 117 (1), 449–472.

Yang, Liu, Qiu, Minghui, Gottipati, Swapna, Zhu, Feida, Jiang, Jing, et al., 2013. CQARank: jointly model topics and expertise in community question Answering. Proceedings of the 22nd ACM International Conference On Conference on Information & Knowledge Management. ACM, pp. 99–108.

Zhang, Yi, Lu, Jie, Liu, Feng, Liu, Qian, Porter, Alan, et al., 2018. Does deep learning help topic extraction? a kernel k-means clustering method with word embedding. J. Informetr. 12 (4), 1099–1117.

Zhao, R.Y., Li, X.L., Liang, Z.S., Li, D.Y., 2019. Development strategy and collaboration preference in S&T of enterprises based on funded papers: a case study of Google. Scientometrics 121 (1), 323–347.

Hongshu Chen received the first Ph.D. degree in management science and engineering from the Beijing Institute of Technology, Beijing, China, in 2015, and the second Ph.D. degree in software engineering from the University of Technology Sydney, Ultimo, NSW, Australia, in 2016. She is currently an Assistant Professor with the School of Management and Economics, Beijing Institute of Technology. Her research interests include bibliometrics, scientific text mining and knowledge discovery. She has published more than 20 papers in refereed journals and conference proceedings in above fields.

Qianqian Jin is currently a Ph.D. student in School of Management and Economics, Beijing Institute of Technology, Beijing, China. Her theme of study has involved bibliometrics, scientific text mining and technology forecasting.

Ximeng Wang received the B.E. degree in communication engineering from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2011, and received the M. E. degree in software engineering and the Ph.D. degree in communication and information systems from Beijing Jiaotong University, Beijing, China, in 2013 and 2020, respectively. He is currently with the Postal Savings Bank of China, Beijing, China. He has published over 10 papers in refereed journals and conference proceedings. His-research interests include machine learning and its applications in user preference modeling, recommender systems and financial anti-fraud systems.

Fei Xiong received the B.E. degree and the Ph.D. degree both in communication and information systems from Beijing Jiaotong University, Beijing, China, in 2007 and 2013, respectively. He is currently an Associate Professor with the School of Electronic and Information Engineering, Beijing Jiaotong University. From 2011 to 2012, he was a Visiting Scholar with Carnegie Mellon University, USA. He has published over 60 papers in refereed journals and conference proceedings. His-current research interests include Web mining, complex networks and complex systems. Dr. Xiong has been a recipient of the National Natural Science Foundations of China and several other research grants.