Scientometrics (2021) 126:969–994
https://doi.org/10.1007/s11192-020-03797-8

A deep learning framework to early identify emerging technologies in large-scale outlier patents: an empirical study of CNC machine tool

Yuan Zhou[1] · Fang Dong[1] · Yufei Liu[2] · Liang Ran[3]

Received: 5 November 2019 / Accepted: 16 November 2020 / Published online: 4 January 2021

Abstract
Radical novelty is one of the key characteristics of emerging technologies. This characteristics makes emerging technologies as a quite different from established technologies. From the perspective of radical novelty, some studies consider patents with little similarity in terms of key concepts and contents to existing patents as candidate emerging technologies. However, existing research remains in examining small-scale patents for evaluating candidate emerging technologies due to the lack of data-processing capacity—the recent rising of deep learning methods may help in this. This study, therefore, develops a novel deep learning based framework for identifying emerging technologies by combining a technological impact evaluation using patents and a social impact evaluation using website articles. Using a large scale multi-source dataset including 129,694 patents and 35,940 website articles, this paper applies the framework to investigate the case of computerized numerical control machine tool technology, through which the framework is validated. The results show that 16,131 patents out of 129,694 patents are considered as candidate emerging technologies, and 192 patents out of 16,131 patents are identified as emerging technologies through the evaluation of technology impact and social impact. This implies that these candidate emerging technologies can evolve to emerging technologies, though not all of them—we need deep learning method to scrutinize a larger scale multi-source data to identify rather a small number of potential emerging technologies. The proposed framework can also be extended to explore other disciplinary multi-source data for strategic decision support in identifying emerging technologies.

Keywords Emerging technologies · Deep learning · Outlier patents · CNC machine tool

✉ Yufei Liu
liuyufei0418@qq.com

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

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

[3] National Numerical Control Systems Engineering Research Center, School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China

Introduction

Emerging technologies will not only can change the technological paradigm that traditional industries rely on or create new industries (Day and Schoemaker 2000; Porter et al. 2002), but also can change the existing socio-economic systems and socio-economic production methods (Adner and Snow 2010; Rotolo et al. 2015). Rotolo et al. (2015) defined five characteristics of emerging technologies: radical novelty, relatively fast growth, prominent impact, coherence, uncertainty and ambiguity. According to these normative definitions, the criteria for a emerging technology are defined and used to identify such technology (Noh et al. 2016).

Early identification of emerging technologies is significant for decision-makers of governments or enterprises to grab strategic opportunities in the face of technological change. A great deal of previous studies of identification emerging technologies have focused on patent data. In these patent-based studies, some studies define a technology as a collection of patents, where an emerging technology is considered as a collection of patents as well. For example, they define an international patent classification (IPC) code or a patent group generated by clustering method to represent a technology, and then, they analyze the characteristics of relevant patents of a IPC code or a patent group to identify emerging technologies (Geum et al. 2012; Kim and Bae 2017; Kyebambe et al. 2017; Zhou et al. 2020). Other studies regard a single patent as a theoretical focal point of analysis. The purpose of these studies are focused on identifying a high impact patent rather than a potential emerging technological field. For example, a few studies predict expected forward citation frequencies of patents and use the prediction results to evaluate the impact of patents (Lee et al. 2016, 2018).

Radical novelty is one of the key characteristics of emerging technologies (Rotolo et al. 2015) which means that emerging technologies are quite different from established technologies (Halaweh 2013; Köhler and Som 2014; Rotolo et al. 2015). From the perspective of radical novelty, some recent patent based studies have identified emerging technologies early from patents with little similarity in terms of key concepts and contents to existing patents (Yoon and Kim 2012; Aharonson and Schilling 2016; Song et al. 2018), and this kind of patent is also called outlier patents (Yoon and Kim 2012). For instance, Aharonson and Schilling (2016) consider outlier patents as candidate emerging technologies and suggeste that outlier patents are more likely to to evolve into emerging technologies than other patents; Song et al. (2018) developed an outlier patents based approach for identifying emerging technologies by applying bibliographic coupling to patents.

In spite of their meaningful contributions by developing a novel outlier patents based approach to help identify emerging technologies, the existing studies are subject to several limitations. On the one hand, current studies have mainly identified emerging technologies in small-scale outlier patents, due to the data size and data types continue to increase rapidly, how to identify the emerging technologies in large-scale outlier patents has not been studied; On the other hand, current studies have mainly emphasized the technological impacts of identifying emerging technologies; few studies have focused on the social impacts of outlier patents. According to the characteristics of emerging technologies (Rotolo et al. 2015), social impacts are one of the most critical criteria to evaluate whether a technology will be an emerging technology or not (Martin 1995).

Considering these limitations, first, this study employs a deep learning model to examine large scale outlier patents. As an advanced machine learning model, deep learning has a relatively complex model structure and exhibits better performance within large scale data

than other machine learning models (Liu et al. 2019). Recent studies have explored the applications of deep learning in bibliometrics and patent analysis, such as emerging technologies forecasting (Zhou et al. 2020), patent classifcation (Li et al. 2018), citation classifcation (Hassan et al. 2018), and natural language processing (Zhang et al. 2018b). Second, we evaluate whether outlier patents are emerging technologies through two aspects: technological impact and social impact. The technological impact of emerging technologies refers to the contribution of emerging technologies to subsequent technologies (Harhoff et al. 1999; Trajtenberg 1990; Geum et al. 2012), and we measure the technological impact by the the number of patent forward citation (Lee et al. 2018). The social impact of emerging technologies refers to the impact of emerging technologies on the general public (Hiltunen 2008; Eaton et al. 2014; Kwon et al. 2017). We measure the social impact by the number of occurrences of patent keywords in the website articles, and website articles refers to all articles in a emerging technology related website, patent keywords refer to the non-virtual words in the patent title.

For the proposed framework, first, outlier patents are identified as candidate emerging technologies based on the content similarity between patents in a large number of patent data. Second, 11 patent indicators of outlier patents are chosen as early signals of emerging technologies. Multi-source data, including patents and website articles, is used to evaluate the external impacts of emerging technologies. The number of forward citations measures the technological impact levels of outlier patents, and the number of times that patent keywords appear in website articles measures social impact levels of outlier patents. Third, the deep learning model is used to fit the associated relationship between early signals and technological impact levels and the associated relationship between early signals and social impact levels, separately. Finally, the associated relationship obtained by the deep learning model is extended to the future to identify potential emerging technologies in outlier patents. The empirical analysis results in the field of CNC machine tool technology prove the effectiveness of the framework. The accuracy of the relationship between early signals and technological impact levels obtained by the deep learning model can reach 74%; of the same accuracy is obtained for the relationship between early signals and social impact levels. Compared with current outlier patents based studies, this study utilizes the deep learning model to solve the problem that identifying emerging technologies in outlier patents is limited by small samples. In addition, this study utilizes multi-source big data including website articles and patents to identify emerging technologies, which can evaluate both social impacts and technological impacts of outlier patents to increase the reliability.

This paper is organized as follows. Section 2 presents the work related to our research, and Sect. 3 explains the research process and methodology. Section 4 provides the guidelines for implementation and evaluation of our approach. Finally, Sect. 5 offers our conclusions.

Related works

Patent-based studies for identifying emerging technologies

The existing approaches for identifying emerging technologies can be classified into two types: normative methods and extrapolative methods. Normative methods are driven by expert such as the Delphi method (Cho et al. 1991; Bañuls and Salmeron 2008) and analytic hierarchy process (AHP) (Lee et al. 2014). On the contrary, extrapolative methods are

driven by data in which patent data is the most frequently used data. In early patent-based studies, most studies have emphasized the usage of patent indicators analysis (Trappey et al., 2011; Bermudez-Edo et al., 2013) and patent citation analysis (Lee et al., 2012; Jang et al. 2017). Then, there have been some studies began to explore the usage of complex network analysis (Sakata et al. 2013; Zhou et al. 2019a) and semantic analysis (Gerken and Moehrle, 2012; Guo et al., 2016). In recent years, with the rise of artificial intelligence (AI), more attention has focused on the usage of AI methods such as machine learning model and deep learning model (Kong et al. 2017; Lee et al. 2018; Kyebambe et al. 2017; Aristodemou and Tietze 2018; Zhou et al. 2020).

In these patent-based studies, some studies define a technology as a collection of patents, where an emerging technology is considered as a collection of patents as well. Specifically, they define an IPC code or a patent group generated by clustering method to represent a technology, and then, they analyzed the characteristics of patents in each of the IPC codes or a patent group to identify emerging technologies. For instance, Geum et al. (2012) used IPC codes defining one IPC code as one technology through analyzing the characteristics of patents in each of the IPC codes to identify emerging technologies; Kim and Bae (2017) clustered patent documents and analyzed patent indicators such as forward citations, triadic patent families, independent claims to assess patent groups for forecasting emerging technologies; Kyebambe et al. (2017) proposed an approach which can automatically label patent clusters to forecast emerging technologies; Zhou et al. (2020) adopted Gartner's emerging technology hype cycles to assign relevant patents and used the deep learning model to forecast technologies. Other studies regard a single patent as a theoretical focal point of analysis. The purpose of these studies are focused on identifying a high impact patent rather than a potential emerging technological field. For instance, Lee et al. (2016) developed an approach to predict expected citation frequencies of each patent to evaluate patents for identifying emerging technologies. Lee et al. (2018) proposed a machine learning approach and using multiple patent indicators to assess the value of each patents for identifying emerging technologies at early stages.

Radical novelty is one of the key characteristics of emerging technologies which means that emerging technologies are quite different from that of established technology (Halaweh, 2013; Köhler and Som, 2014; Rotolo et al., 2015). From the perspective of radical novelty, the latest research has given more attention to the outlier patents(Yoon and Kim 2012; Aharonson and Schilling 2016; Song et al. 2018). Outlier patents refer to patents that have recently published but have little similarity with the existing patents in terms of their contents. For instance, Yoon and Kim (2012) identified outlier patents using semantic similarity and regarded outlier patents as potential emerging technologies with the great possibilities of technological paradigm change. Aharonson and Schilling (2016) considered outlier patents were as candidate emerging technologies, and also argued that outlier patents to be more likely to evolve into emerging technologies than other patents. Song et al. (2018) suggested an outlier patents based methodology for identifying emerging technologies by applying bibliographic coupling to patents.

Though the meaningful contributions by developing novel approach based outlier patents. Little effort has been made to investigate whether the technologies evaluated as emerging technologies have actually become emerging. According to the characteristics of an emerging technology, emerging technologies will not only have a significant impact on the technological field, but also have a significant impact on society (Martin 1995). At the same time, some technologies have had a great impact on the technological field, but have failed to produce a significant impact on society. As a result, such technologies are not emerging technologies. It is difficult to evaluate and forecast whether a technology

is a potential emerging technology based on a single patent data. Current research indicates that website articles is effective in reflecting the impact of emerging technologies on society (Aral et al. 2013; Kalampokis et al. 2013; Breitzman and Thomas 2015; Injadat et al. 2016; Askitas and Zimmermann 2015). In this study, website articles and patent data, and these two multi-source data are applied to evaluate emerging technologies more comprehensively, with consideration of both technological impacts and social impacts.

Machine learning studies for identifying emerging technologies

Machine learning is considered to be an effective tool for discovering implicit information in big data and is applied to the identification of emerging technologies (Aristodemou and Tietze 2018). Machine learning models can be classified into supervised learning models and unsupervised learning models, according to whether external labeled data is required (Bishop 2006). The unsupervised learning model is mainly used to discover the inherent distribution of data, and the supervised learning model is mainly used to discover hidden relationships.

Previous studies using machine learning models to identify emerging technologies mostly use unsupervised learning machine learning models, including the k-means clustering model, topological clustering model, and topic model. For instance, Choi and Jun (2014) developed a Bayesian model for patent clustering to forecast emerging technologies, and Zhou et al. (2019a) developed a framework through the citation network and topology clustering to reveal the convergence process of scientific knowledge to forecast emerging technologies. Zhou et al. (2019b) developed a semi-supervised topic clustering model, and generated a sentence-level semantic technological topic description to identify emerging technologies. The goal of the unsupervised learning model is to discover the distribution characteristics of the data itself, without the need for external, manually labeled data. However, the unsupervised learning model generates results based on mathematical optimality, which makes it difficult to directly meet the actual needs in practice. The results obtained by the unsupervised learning model require a large amount of external domain knowledge for identification and screening. A remedy for this is the supervised learning method, which can directly generate prediction or regression results by embedding external knowledge into the model using external, manually labeled data. The aim of the supervised learning model is to discover the relationship between different data features. Some recent studies have employed supervised learning models to identify emerging technologies; these models include the SVM, ANN, naive Bayesian model, and random forest model. For example, Kreuchauff and Korzinov (2017) developed a support vector machine model based on the robotics patents to detect the early development of an emerging technology in patent data. Kyebambe et al. (2017) used labeled data based on new classes established in the United States Patent Classification (USPC) system to train supervised learners to forecast emerging technologies. Meanwhile, Lee et al. (2018) employed a feed-forward multilayer neural network to capture the complex nonlinear relationships between input and output indicators to identify emerging technologies in the early stages. Studies based on supervised learning models mostly employ statistical supervised learning models to verify the effectiveness of the method. Since data size and data types increase rapidly, and due to their simple structure, these models have difficulty accurately fitting the implicit relationship between emerging technologies and early signals.

As a typical form of supervised learning, deep learning has a complex model structure and better performance (Liu et al. 2019). Deep learning, developed by Geoffrey Hinton

(Hinton and Salakhutdinov 2006), has become the key technology of big data intelligence (Zhuang et al. 2017) and has led to major breakthroughs in many fields. Recent studies have also explored the value of deep learning in bibliometrics. Zhou et al. (2020) proposes a novel approach that integrates data augmentation and deep learning method to overcome the problem of lacking training samples for forecasting emerging technologies. Li et al. (2018) proposed an effective patent classification algorithm based on deep learning to solve the large-scale and multiclass patent classification problem, and suggested that deep learning has several advantages in large-scale patent classification, including free of hand crafted features, straightforward models. Moreover, deep learning is easy to implement without tedious feature engineering compared with traditional supervised learning algorithms. Hassan et al. (2018) compared deep learning and the classical statistical supervised learning models for classifying the importance of a citation using the same dataset, and the results showed that deep learning with all 64 features has higher accuracy than SVMs and a RF with the 29 best features. This study also proved the powerful modeling power of deep learning with complex features and large-scale data. Zhang et al. (2018b) utilized word embedding, an application of deep learning in natural language processing, mapped words from vocabulary to vectors, and created a way to discover the latent semantics in large-scale text. This study showed the superior performance of deep learning in handling topic extraction tasks in large-scale text data.

In this study, the deep learning model is used to fit the implicit relationship between emerging technologies and early signals. The deep learning model consists of an input layer and an output layer (Schmidhuber 2015). Early signals emerging from emerging technologies serve as input features for the input layer, and emerging technology evaluation indicators serve as labels for the output layer. In previous studies, many patent indicators were designed to describe the characteristics of emerging technologies in different aspects (Porter and Detampel 1995; Ernst 1998; Ernst 2003; Ernst and Omland 2011; Geum et al. 2013; Song et al. 2016). These indicators can be classified into static indicators that do not change over time and dynamic indicators that do change over time. Static patents indicators mainly indicate technological characteristics, such as the number of IPCs (Tong and Frame 1994; Lanjouw and Schankerman 1997), the number of inventors (Balconi et al. 2004; Sternitzke et al. 2008), and the number of non-patent references (Narin et al. 1997; Meyer 2000). Dynamic patent indicators mainly indicate the value of technology, such as the number of forward citations (Lerner 1994; Narin et al. 1987), and the number of patents being renewed (Pakes and Schankerman 1984; Lanjouw et al. 1998). In the proposed framework, 11 static patent indicators are used as early signals for emerging technologies, and the number of forward citations as a dynamic patent indicator is used to evaluate the technological impact levels of outlier patents. In addition, the number of times that patent keywords appear in website articles is used to evaluate the social impact levels of outlier patents. Two different dimensions of the technological impact levels and social impact levels are used to determine whether outlier patents are potential emerging technologies.

Methodology

Overall process

Figure 1 shows the overall process of the proposed approach. Given the complexities involved, the proposed approach is designed to be executed in 5 major steps: (1) collecting outlier patents as candidate emerging technologies; (2) assessing the social impacts and technological impacts of candidate emerging technologies; (3) developing static patent indicators as early signals of candidate emerging technologies; (4) fitting relationships between early signals and candidate emerging technologies' impacts; and (5) identifying emerging technologies in the future among candidate emerging technologies. A detailed discussion of how we performed the proposed approach follows.

Outlier patents collection

Patents are retrieved from the patent database, and then, outlier patents are collected in; these patents. Once a technology field of interest is chosen, the relevant patents are collected based on certain search criteria. The patent documents are parsed according to the types of information, including attribute information and citation information, for outlier patents collection and patent indicators extraction.

Outlier patents refer to isolated nodes that do not form a connection with any of the patent nodes in the patent similarity network. The nodes of the patent similarity network are patents, and whether there is a connection between patents is determined by the similarity

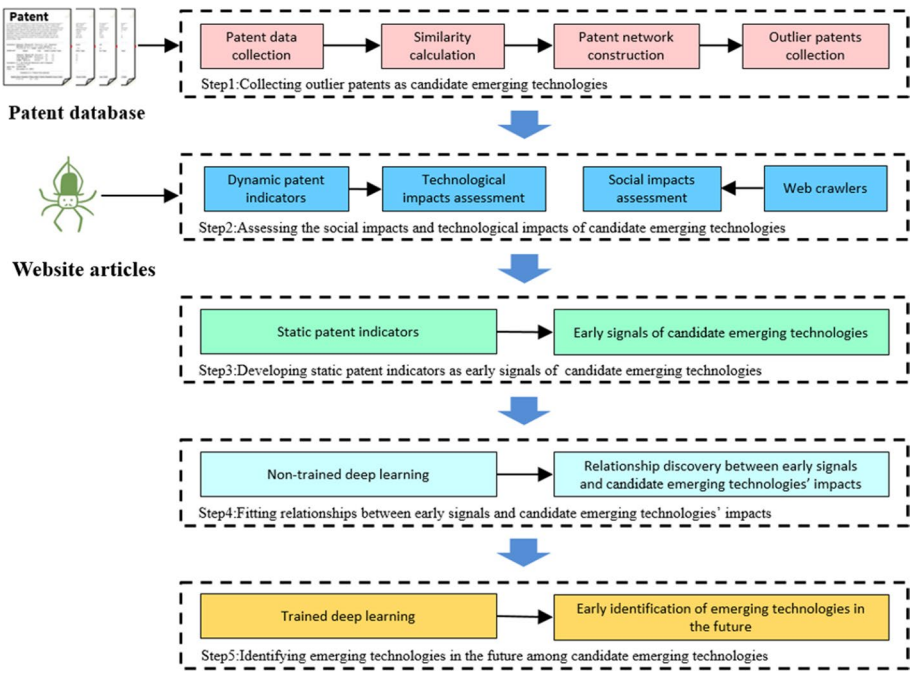

Fig. 1 Overall process of the proposed approach

between them. As shown in Fig. 2, when the similarity between two patents is greater than the set domain value, there is a connection, and vice versa. In this study, bibliographic coupling is used to calculate the similarity between patents. Bibliographic coupling is based on the concept that patents with more co-citations will be from a similar technological area (Kessler 1963). Since patent citation data are created by an inventor or an examiner who is generally an expert in the field, bibliographic coupling can calculate the similarity between patents relatively accurately (Boyack and Klavans 2010). The process of calculating the similarity between patents using patented coupling involves two steps. First, a bibliographic coupling matrix is developed to present the number of co-citations between patents. Second, the similarities of patents are measured by applying a cosine similarity to the matrix.

Assessing the impacts of outlier patents

Assessing the social impacts

The social impact of emerging technologies refers to the impact of emerging technologies on the general public (Hiltunen 2008; Eaton et al. 2014; Kwon et al. 2017). Specifically the social impact include the impact on the general public's employment, general public's health and general public's living environment (Kwon et al. 2017). Existing studies suggest that emerging technology-related websites can effectively reflect the social impact of emerging technologies (Raford 2015; Kwon et al. 2017; Li et al. 2019), since the website articles on the emerging technologie related website are mainly for the general public and the main content of website articles is the impact of emerging technologies on the general public (Kwon et al. 2017). Kwon et al. (2017) suggests that the more articles on the emerging technology related websites, the higher the impact of emerging technologies on the general public. In this study, we measure the social impact by the number of occurrences of patent keywords in the website articles, and website articles refers to all articles in a

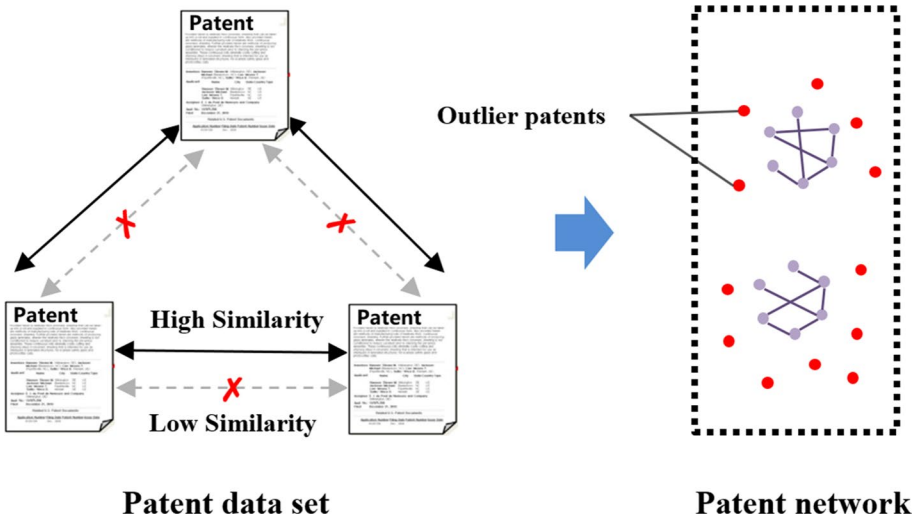

Fig. 2 The process to collect outlier patents

emerging technology related website, patent keywords refer to the non-virtual words in the patent title.

Website articles is obtained through web crawlers. The web crawler collects all the articles in a website and puts them into a local database. These articles are parsed according to the title, issuing time, and article body. The same article may appear in different website columns, and the raw data obtained by the web crawler needs to remove duplicate data. The result of the social impacts of the technological innovation is a continuous integer variable. The precise description of the number of times the patent keywords appear in the website articles is not our focus. Considering this issue, the original results of social impacts are classified into 3 different levels, from low to high, namely, SL1, SL2, and SL3. The original results are arranged from small to large, the first 40% of the original results correspond to SL1, the middle 40% of the original results correspond to SL2, and the last 20% of the original results correspond to SL3.

Assessing the technological impacts

The number of forward citations of a patent is used to assess the technological impacts of technological innovations in the outlier patent. Although the previous bibliometric research has proposed many valuable methods to assess the technological impacts, the number of forward citations is one of the most popular means to measure technological impacts. This allows us to evaluate the degree that a technology has contributed to subsequent technologies, and technologies that are more frequently and widely used in subsequent technologies are regarded as having a higher technological impact (Harhoff et al. 1999; Trajtenberg 1990; Geum et al. 2012). As the way social impacts, the original results of technological impacts are classified into 3 different levels from low to high, TL1 with 0 backward citations, TL2 with 1 backward citations, and TL3 with more than 1 backward citations.

Developing patent indicators as early signals

The bibliometric literature has developed a variety of static patent indicators to assess the characteristics of potential emerging technologies. This study employs a total of 11 static patent indicators as early signals of emerging technologies, which are divided into five subcategories: (1) novelty, (2) science intensity, (3) growth speed, (4) scope and coverage, and (5) development effort and capabilities. These static patent indicators are immediately determined when the relevant patents are issued and will not change over time. The description and measurement method of each indicator is shown in Table 1.

(1) Novelty. Two indicators are employed to represent the novelty of patents. The first indicator is technological originality, which captures a combination of existing ideas and how much the patented invention draws from different sets of technologies. The existing literature has suggested that patents based on a wider set of ideas may indicate more valuable knowledge (Jaffe and Trajtenberg 2002; Fernández-Ribas 2010; Bessen 2008). The second indicator is prior knowledge, which captures the relationship between the target patent and prior arts. The existing literature has identified that patents with large numbers of backward citations have a relatively low novelty and low monetary value (Harhoff et al. 2003; Haupt et al. 2007).

Table 1 Static patent indicators employed in this study

Categories	Indicator	Description
Novelty	Technological originality (TO)	Herfindahl index on classes of cited patents
	Prior knowledge (PK)	Number of backward citations
Growth speed	Technology cycle time (TCT)	Median age of cited patents
Science intensity	Scientific knowledge (SK)	Number of non-patent literature references
Scope and coverage	Technological scope (TS)	Number of classes to which a patent belongs
	Commercial scope (CS)	Number of patents registered in multiple countries with the coverage of the same invention
	Protection coverage described in independent claims (PCID)	Number of independent claims
	Protection coverage described in dependent claims (PCD)	Number of dependent claims
Development effort and capabilities	Collaboration (COL)	1 if a patent has more than one assignee; otherwise, 0
	Inventors (INV)	Number of inventors
	Total know-how (TKH)	Number of patents issued by an assignee

(2) Growth speed. Technology cycle time is employed to represent the growth speed of technological development. This indicator is widely used in the existing literature and can capture the degree of newness of prior knowledge or the pace of technological progress (Bierly and Chakrabarti 1996; Kayal and Waters 1999).

(3) Science intensity. Scientific knowledge is employed to represent the science intensity of patents. The basic assumption of this indicator is that more scientific knowledge contained in the patented invention may lead to the development of more innovative and influential technology (Cozzens et al. 2010; Day and Schoemaker 2000).

(4) Scope and coverage. Four indicators are employed to represent the scope and coverage of patents. The first indicator is technological scope, which captures the scope of the technological fields of a patent by the number of classes to which a patent belongs (Lee et al. 2009). The second indicator is commercial scope, which captures the scope of the commercial fields of a patent by the size of the patent family (OuYang and Weng 2011). It has been suggested that there are positive relationships between the size of the patent family and the economic value of a patent (Guellec and de la Potterie 2000). The third indicator is the protection coverage described in independent claims, and captures the essential technological innovation of a patent by the number of independent claims (Lanjouw and Schankerman 2001; Ernst 2003; Jeong et al. 2016). The fourth indicator is the protection coverage described in dependent claims, and captures the additional technological innovation of a patent by the number of dependent claims (Lanjouw and Schankerman 2001; Ernst 2003; Jeong et al. 2016).

(5) Development effort and capabilities. Three indicators are employed to represent development capabilities. The first indicator is collaboration, which captures collaboration on the assignee side of patents; the literature suggests that there is a significant positive relationship between the coassignee and the value of a patent (Ma and Lee 2008; Meyer 2006). The second indicator is inventors, which captures collaboration on the inventor side of patents, and it has been suggested that patents by multiple inventors are more valuable than those by a single inventor (Ernst 2003; Ma and Lee 2008). The

third indicator is total know-how, which captures the level of assignees' knowledge stocks (Meyer 2006).

Deep learning approach to finding relationships

The key to early identification of emerging technologies is to discover the relationship between early signals and external impacts of every outlier patent. However, a technology will be influenced by many factors in its evolution, which makes it impossible to directly measure the technological impacts and social impacts of every outlier patent in the future. In this study, we will push the time forward to replace the future technological impacts and social impacts by evaluating the current technological impacts and social impacts of the outlier patents, as shown in Fig. 3. First, all outlier patents of a chosen technology in a certain year in history (Year T-t) are identified, and patent indicators of each outlier patents are chosen as early signals of emerging technologies. Then, the current (Year T) technological impacts and social impacts of these outlier patents are calculated. Finally, the deep learning model is used to fit the relationship between early signals of each historical outlier patents and future external impact levels.

The technological impacts and social impacts of outlier patents have been classified into 3 different levels. Specifically, the deep learning model is used to fit the relationship between the early signals and the different levels of external impacts. Thus, the deep neural network (DNN) as the deep learning classifier is employed in the proposed framework. The impacts of outlier patents include both technological impacts and social impacts, and two DNN are used to fit the relationship between early signals and technological impact levels and the relationship between early signal and social impact levels, respectively. The operation of the DNN involves two steps: training and testing. In the training step, the two DNN use the training set to formulate connection weights for each node in the network. In the training process, when the loss function of the DNN converges, the training process of the DNN model ends, and the connection weights for each node of the DNN model are obtained. In the testing step, the performances of the two trained DNN are tested by the test set.

The architecture of two DNN include an input layer, an output layer, and some hidden layers. For the architecture of DNN that fit the associated relationship between early signals and technological impact levels, the 11 static patent indicators of each outlier patent in the training set are used as the input feature vector for the input layer; The technological impact levels (TL1, TL2, TL3) are used as the category label for the

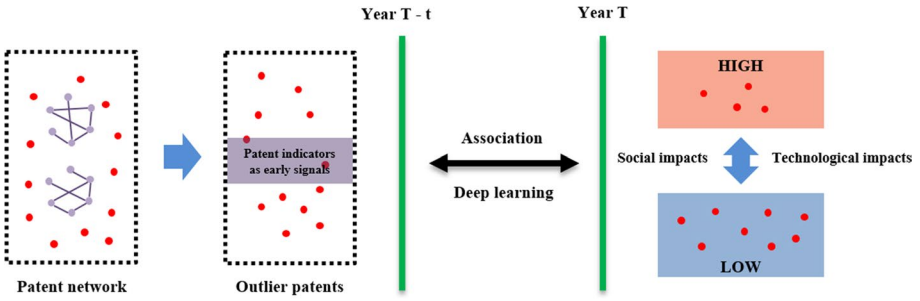

Fig. 3 The process to find relationships by deep learning

output layer; Hidden layers which are both rectified linear units (Relu) connect the input layer and the output layer. For the architecture of DNN that fit the associated relationship between early signals and social impact levels, the 11 static patent indicators of each outlier patent in the training set are also used as the input feature vector for the input layer; But the social impact levels (SL1, SL2, SL3) are used as the category label for the output layer; Hidden layers which are also both Relu units connect the input layer and the output layer.

The performance of the DNN directly reflects the quality of the discovered relationships. Accuracy, F-measure, and G-mean are used to evaluate the performance of the DNN. Accuracy denote the proportion of predictions that are correct, the F-measure represente the harmonic mean of precision and recall (e Sousa et al. 2017), and the G-mean indicate the geometric mean of recall (Sun et al. 2007). These 3 evaluation metrics are based on the confusion matrix shown as Fig. 4, and are defined in Eqs. 1–3. M represents the number of classes. Ri and Pi indicate the recall and precision of class Li, respectively, and are defined in Eqs. 4 and 5. nii and nij denote the number of class Li samples that are correctly predicted as class Li and incorrectly predicted as class Lj, respectively.

$$Accuracy = \frac{\sum_{i=1}^{M} n_{ii}}{\sum_{i=1,j=1}^{M} n_{ij}} \tag{1}$$

$$F\text{-}measure = \frac{2}{L} \cdot \frac{\sum_{i=1}^{M} R_i \sum_{i=1}^{M} P_i}{\sum_{i=1}^{M} R_i + \sum_{i=1}^{M} P_i} \tag{2}$$

$$G - mean = \left(\prod_{i=1}^{M} R_i \right)^{\frac{1}{M}} \tag{3}$$

$$R_i = \frac{n_{ii}}{\sum_{j=1}^{M} n_{ij}} \tag{4}$$

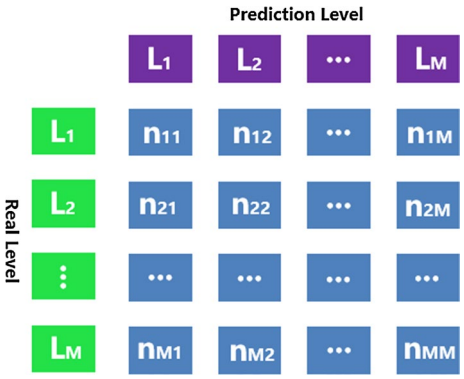

Fig. 4 Diagram of multi-classification confusion matrix

$$P_i = \frac{n_{ii}}{\sum_{j=1}^{M} n_{ji}} \tag{5}$$

Identifying emerging technologies

The identification of emerging technologies involves assessing the future technological impacts and future social impacts of all current outlier patents through technological evolution. Outlier patents of a technology field are candidate emerging technologies, and only those outlier patents that have both high technological impacts and high social impacts in the future can be considered as emerging technologies. The future technological impacts and future social impacts of current outlier patents cannot be directly measured. We can assess future technological impacts and future social impacts of current outlier patents based on the early signals of current outlier patents and the discovered relationship between early signals of current outlier patents and future external impacts. Although emerging technologies will be influenced by many factors, according to the existing research, they will still have some characteristics in common with other technologies during the initial stage of development. Moreover, the evolution path of emerging technologies will follow certain patterns. Thus, we consider that the relationship discovered by the deep learning model in historical data can be extended to the future.

Early identification of emerging technologies in outlier patents is based on the future technological and social impact levels obtained by the deep learning classifier. After training the deep learning classifier in historical data, the early signals of each current outlier patents are inputted to the deep learning classifier. Then, the deep learning classifier can directly output the future technological impact levels (TL1, TL2, TL3) and the future social impact levels (SL1, SL2, SL3) of these outlier patents. In this study, in order to increase the precision of the early identification of emerging technologies, we only select outlier patents that have both the highest level (TL3–SL3) of technological impact and the highest level of social impact in the future as emerging technologies.

Empirical analysis and results

A case study is conducted on CNC machine tool technology. For the CNC machine tool technology, emerging technologies such as intelligent control systems and intelligent machine tools have emerged one after another, which provides a valuable opportunity to analyze the relationship between emerging technologies that have already emerged and their early signals. Moreover, CNC machine tool technology as the basic equipment technology of modern industrial manufacturing, identifying potential emerging technologies in this technology field is of great significance for governments to catch technology development opportunities and reduce R&D risks.

Regarding the framework, both patent data and website articles are required to identify potential emerging technologies. The Thomson Reuters (TI) patent database is employed to obtain patent data. It is one of the most representative databases for academic research and contains comprehensive and high-quality patent data. We develop a patent search formula based on the keywords of CNC machine tool technology to retrieve patent data, and

the search formula is shown in "Appendix 1". Then, a total of 129,694 patent data, dated up to the end of 2018, are retrieved from the patent database. We select a CNC machine tool technology portal website www.industryweek.com to obtain website articles. At the end of 2018, we employ a web crawler to crawl all the articles on this website; a total of 35,940 articles are obtained after removing the duplicate articles from the different website columns.

Extracting indicators and impacts of outlier patents

The historical outlier patents are collected as candidate emerging technologies, and these candidate emerging technologies are evaluated for current technological impact levels and social impact levels. At present, the emerging technologies in the field of CNC machine tool technology are mainly due to the outbreak of artificial intelligence and big data 10 years ago. Thus, we collected all outlier patents in the patent data of CNC machine tool technology 10 years ago (year 2008). A total of 28,265 outlier patents are obtained, and these outlier patents are used as candidates emerging technologies.

For all the collected outlier patents, 11 static patent indicators are extracted as early signals of emerging technologies, and 2 indicators are extracted from patent data and website articles to evaluate the current (year 2018) technological impact levels and current (year 2018) social impact levels for each outlier patent. The two evaluation indicators are used to determine whether outlier patents are potential emerging technologies. The resulting matrix is a 28,265 by 13 indicator matrix (11 static patent indicators and 2 evaluation indicators), which is used to train and test the DNN. The matrix is not reported here in its entirety owing to lack of space, but part of the matrix is shown as Table 4 in "Appendix 2".

Based on all the current (year 2018) outlier patents at different technological and social impact levels, we utilize two box plots to analyze the 11 static patent indicators, and the data of the box plots is standardized data, as shown in Figs. 5 and 6. For different technological impact levels, TS, CS, PCID, INV, TKH increase as the level of technological impact increases and the remaining static patent indicators do not have this trend; for different social impact levels, PK, TCT, TS, PCD, INV, TKH increase as the level of social impact increases and the remaining static patent indicators do not have this trend. However, from the distribution of the static patent indicators in the box plots, it is still difficult to distinguish different impact levels by a single static patent indicator or a set of static patent indicators.

Further investigation on the distribution of static patent indicators is conducted. We use the principal component analysis (PCA) method to analyze the 11 static patent indicators, and select top two contributing rate PCA scores (PC1 and PC2) as the two-dimensional coordinates for visualization, as shown in Figs. 7 and 8. The PCA scores plots show that different impact levels have overlapping areas, which makes it difficult to distinguish different impact levels by PCA scores plots. Thus, it is necessary to apply the deep learning model to discover the relationship between static patent indicators and future impact levels. Although the deep learning model cannot intuitively show the difference distribution of static patent indicators at different impact levels, it can give accurate classification results.

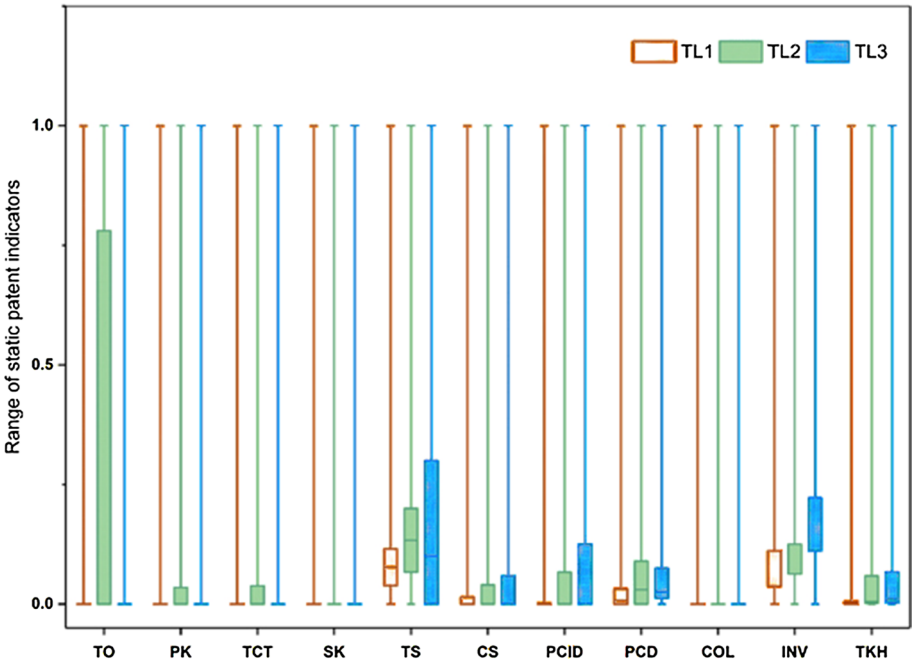

Fig. 5 Distribution of static patent indicators at each technological impact level

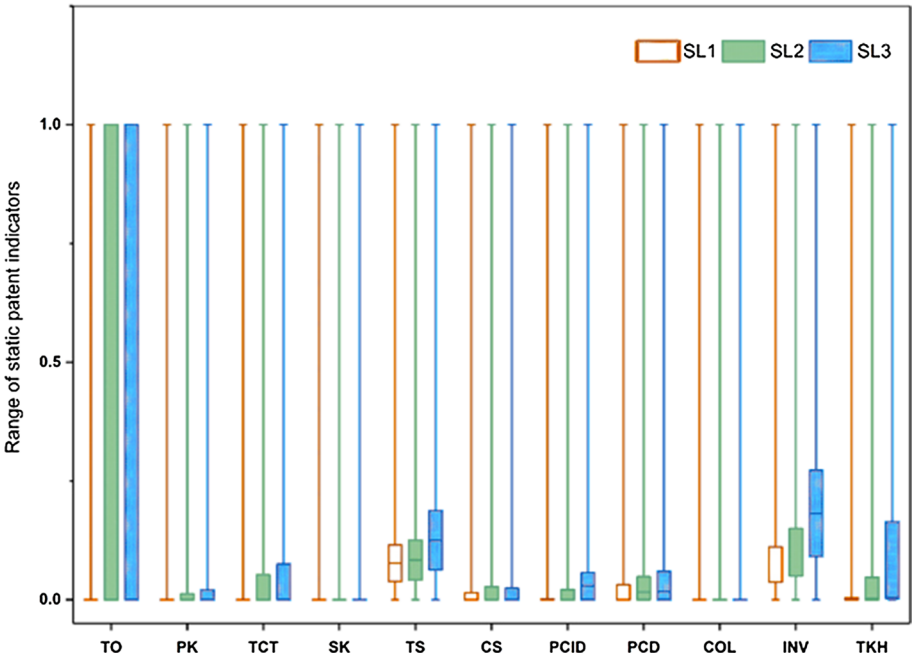

Fig. 6 Distribution of static patent indicators at each social impact level

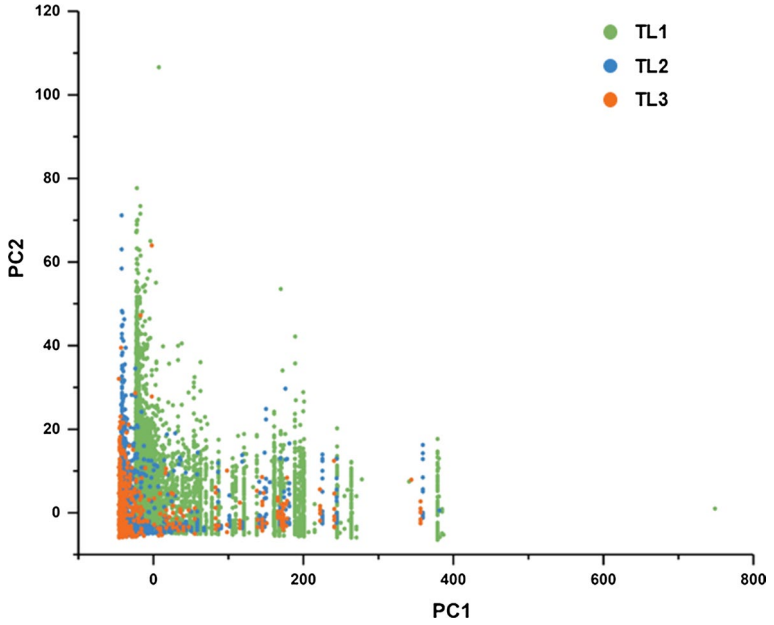

Fig. 7 PCA scores of static patent indicators at each technological impact level

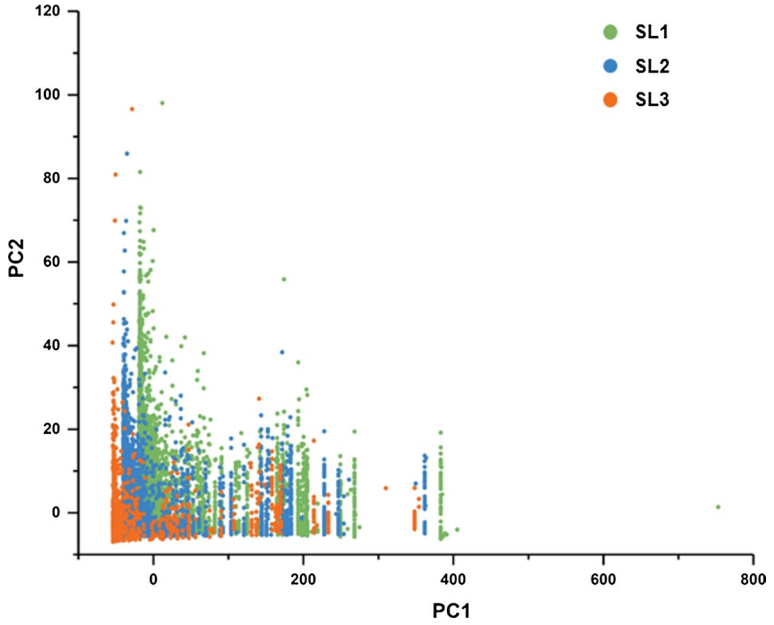

Fig. 8 PCA scores of static patent indicators at each social impact level

Relationship results

Two DNN are developed using Python and TensorFlow package to fit the associated relationship. 70% of historical outlier patents are randomly selected as the training set, and the remaining 30% of historical outlier patents are randomly selected as the testing set. To evaluate the efectiveness of DNN, we used random forests (RF) and logistic regression (LR) for comparison experiments.

First, parameters for each DNN are determined by experimental analysis. To start with, the number of layers and the number of nodes for each layer in the DNN are fundamental parameters. Too few layers will hinder the ability of a network to build a representation at a level of abstraction that is appropriate to adequately capture data complexity; conversely, too many layers will substantially complicate training and likely cause overfitting (Fiore et al. 2017). Since training and tuning a DNN is an expensive operation, we conduct a limited number of experiments, in which 8, 16, 32, 64, 128, 256 and 512 nodes with 2 to 16 layers are tested, respectively. The optimal parameters of the two DNN are determined after a series of experimental analysis. These 2 DNN have same parameters, 11 input units are used as the output layer and 3 softmax units are used as the output layer, the classifier has 10 hidden layers, the first 4 hidden layer contains 512 ReLU units and the last 6 hidden layer contains 256 ReLU units; cross entropy is used as the loss function, and the maximum number of iterations is set to 50,000.

Second, parameters of RF include the number of trees, maximum depth of trees and number of attributes used by each tree. The more trees there are in RF, the better the result and the longer the training time will be. However, when the result is optimized to some extent, the effect of the increase in the number of trees will disappear (Wang and Xu 2018). In the process of optimizing parameters, the number of trees for each RF is sampled from the range of 1 to 10 by a step of 1. At the same time, an appropriate maximum depth of a tree can effectively avoid over-fitting and underfitting problems. We evaluate the maximum depth from the range of 2–20 by a step of 1. Moreover, the number of attributes used by each tree is generally set as the root of the total number of attributes, and we evaluate the parameter from 2 to 11. Experimental results prove that 2 RF models have same parameters. Specifically, 3 trees whose maximum depth is set as 10 and 3 attributes used by each tree achieve the best performance.

Third, for the ordinary LR model, since it is a simple linear classifier, there are no priori parameters that need to be determined by experimental analysis like DNN and RF. However, the ordinary LR model can only deal with the binary classification task. Thus, a method is needed to enable the LR model to deal with the multinomial classification task. According to existing research, we adopt a method of training multiple independent LR models to deal with the multinomial classification task. Specifically, the number of independent LR models is equal to the number of classification categories minus 1. In this study, the classification categories of the multinomial classification task are both 3. For each multi-classification task, two corresponding LR models need to be trained, and the category with the largest number of samples is used as the base category.

The experimental results are shown in Table 2. The accuracy of the two DNN are higher than the RF and the LR, which means that DNN is stronger in ability to fitting the relationship between early signal and external impact levels. The F-measure and G-mean of the two DNN are not as high as the accuracy, which is mainly due to the unbalanced sample size of each category, but the F-measure and G-mean of the two DNN exceed 70%, which indicate that the unbalanced sample size has no significant impact on the classification

Table 2 The performances of DNN, RF and LR

Model	Accuracy	F-measure	G-mean
DNN for technological impacts	0.744	0.740	0.724
DNN for social impacts	0.744	0.746	0.736
RF for technological impacts	0.622	0.629	0.616
RF for social impacts	0.678	0.671	0.649
LR for technological impacts	0.456	0.455	0.403
LR for social impacts	0.589	0.598	0.541

quality and the classifier has good classification performance for each category. Moreover, the F-measure and G-mean of the two DNN are also higher than the RF and the LR, which means that DNN perform better than RF and LR when dealing with the unbalanced sample size of each category. In general, DNN performs better than RF, and RF performs better than LR. The results of comparative experiments suggest that DNN, as the most complex classification model in the comparison experiment, can achieve excellent performance after training with large-scale data and high computational costs, while LR, as the simplest classification model in the comparison experiment, is difficult to obtain good performance when dealing with the large-scale data. This finding is also consistent with existing research conclusions (Goodfellow et al. 2016; Liu et al. 2019). Taking these results together, it is confirmed that these two DNN models are effective in fitting the relationship.

Identifying emerging technologies

Assessing the impacts of outlier patents

Two deep learning classifiers trained in historical outlier patents are used to evaluate the potential technological impact levels and social impact levels of current outlier patents in the field of CNC machine tool technology. A total of 16,131 outlier patents are obtained in 2018, and these outlier patents are used as current emerging technologies candidates. The 11 static patent indicators for each current outlier patent are extracted as the input feature vector for the input layer of the deep learning classifier. The two trained deep learning classifiers can directly generate the technological impact levels (TL1, TL2, TL3) and social impact levels (SL1, SL2, SL3) of each outlier patent after 10 years based on the input feature vector. The evaluation results of the current outlier patents are shown in Fig. 9. We find that 192 current outlier patents have both the highest level (L3-level) technological impacts and social impacts in the future. In this study, we consider that the technological innovations contained in these outlier patents are more likely to have high technological impacts and social impacts in the future. That is, these outlier patents are potential emerging technologies.

Analysis of emerging technologies

Natural language processing (NLP) is used to analyze the content of the identified outlier patents. NLP can describe potential emerging technologies in the field of CNC machine tool technology, and the results can provide guidance for governments and companies to

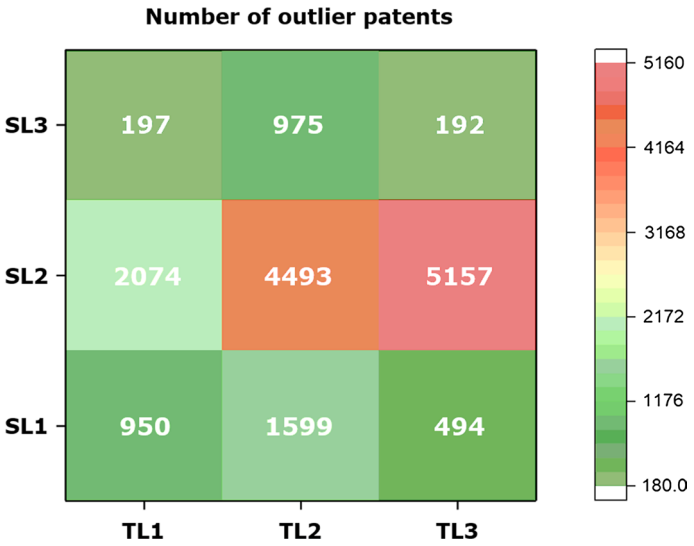

Fig. 9 Future impact map of outlier patents in 2018

develop relevant technology development policies. More specifically, the Latent Dirichlet Allocation (LDA) model is employed to analyze the text content of these identified outlier patents; the analyzed text includes titles and abstracts. The LDA model is commonly used for processing text content, and can automatically cluster text content and extract topic words for each category (Yau et al. 2014). In this study, we analyze the topics of potential emerging technologies based on the topic words obtained by the LDA model. The identified outlier patents are clustered into 5 categories (K = 5) according to the topics by the LDA model. The parameters of the LDA model are determined based on the empirical parameters, $\alpha = 50/K = 10$ and $\beta = 0.01$, and the number of iterations is 5000. The topic words of each category extracted by the LDA model are shown in Table 3.

As a result, we determine the potential emerging technologies corresponding to each topic based on these topic words and the background knowledge of CNC machine tools. (1) Topic 1: Autonomous sensing and connection. Autonomous sensing and connection is the frontier for the CNC machine tools (Chen et al. 2019). During the operation of CNC machine tools, a large amount of real-time in-process electronic data will be produced,

Table 3 Topic words of outlier patents

Topic id	Emerging technology	Topic words
1	Autonomous sensing and connection	Laser controller data computer signal wireless automatic connected
2	Ultrasonic cutting main shaft	Ultrasonic detection machine tool cutter main shaft measuring
3	Numerical control chip	Numerical control chip module unit operation arranged protective
4	Surface precision processing	Surface processing piece material cutter precision product
5	Electric main shaft	Electric main shaft servo transmission motor roller screw

which is a quantitative and precise description of the working condition. By automatically aggregating the electronic data to the data center which is connected to the CNC machine tool in real time, autonomous sensing and connection is realized (Zhou 2015; Zhou et al. 2018; Chen et al. 2019). (2) Topic 2: Ultrasonic cutting main shaft. Ultrasonic cutting main shaft is a new CNC processing technology. Ultrasonic cutting main shaft changes the traditional cutting mode by avoiding continuous touching between the cutter and the work piece, can reduce the processing temperature, increase the cutter life, and improve the processing precision (Zhang et al. 2018a). (3) Topic 3: Numerical control chip. Numerical control chip is the computing core of a CNC machine tool. With the development of high-speed, high-precision and intelligent CNC machine tools, numerical control chip also need to be continuously improved to meet the emergent requirements (Martinov and Kozak 2016). (4) Topic 4: Surface precision processing. Surface precision processing is a new and complex application scenario faced by CNC machine tools. With the increasing complexity of mechanical parts, various special-shaped surfaces have been produced. Exploring the usage of CNC machine tools to process these special-shaped surfaces has become a new research area (Chen et al. 2019). (5) Topic 5: Electric main shaft. Electric main shaft is a new and core component for the CNC machine tools. Compared with the traditional main shaft, the electric main shaft eliminates all mechanical transmission such as gears and belts from the motor to the main shaft, which can overcome the shortcomings of traditional main shaft such as slippage, vibration, and noise at high speeds, and improve the precision and efficiency of main shaft (Albertelli 2017; Li et al. 2017).

The findings from the potential emerging technologies provide some useful insights on the upcoming innovation in the CNC machine tool. First, the innovation of CNC machine tool technology includes two aspects, one is the body of CNC machine tool, such as autonomous sensing and connection, ultrasonic cutting main shaft, numerical control chip and electric main shaft, the other is the application of CNC machine tools, such as surface precision processing. Second, these identified potential emerging technologies are all newly appeared technologies except numerical control chip. As the core and fundamental technology of CNC machine tools, numerical control chip is a traditional research field. As the proposed framework emphasized the use of outlier patents, we consider that although numerical control chip is not new, some novel methods or applications are emerging in numerical control chip, and it is necessary to pay attention to the development and innovation in this field. Finally, it should be noted that the final decision as to whether such potential emerging technologies are worth pursuing in an organization should be made after careful consideration by the domain experts with various knowledge on relevant technological fields

Discussions

There were several issues raised during the analysis processes. First, it is necessary to discuss the identified potential emerging technologies of CNC machine tool. The CNC machine tool is the foundation of the manufacturing industry. At present, as the main focus of the intelligent manufacturing, it has deeply integrated with advanced information technology, and is undergoing a paradigm change. There are some valuable studies consider that the evolution of the CNC machine tool can be divided into three stages: the CNC machine tool, the smart machine tool, and the intelligent machine tool (Zhou 2015; Zhou et al. 2018; Chen et al. 2019). Intelligent machine tools are the development direction of CNC machine tool, it is taken for granted that current innovations are focused on intelligent

machine tools. However, according to the identified potential emerging technologies and the characteristics of intelligent machine tools, only the autonomous sensing and connection belong to the intelligent machine tools. We consider that it is necessary to pay attention to the key foundational technologies and emerging application fields of CNC machine tools, while focus on the innovation of intelligent machine tools. The CNC machine tool in each stage has a progressive relationship, good foundation is the guarantee of developing intelligent machine tool. Therefore, the governments or enterprises should consider their own condition including technological basis, application field, etc. for reasonable decision making and strategic planning.

Second, this study adopted a deep learning model to construct a framework for identifying emerging technologies. It should be noted that though the deep learning model is powerful and effective with large-scale data training, the application scenarios of deep learning model should be noted in the research. Since the deep learning model recombines the input patent features, the deep learning model in this study cannot tell us the specific relationship between the input patent features and the external impacts of the patent. For example, the degree and direction of a certain patent feature's influence on the external impacts. Therefore, this research is based on the deep learning model to obtain an effective prediction model, which can effectively predict the external impacts of the patent based on the patent features. In subsequent research, we will continue to explore the specific relationship between patent features and patent external impacts. It is difficult to achieve this research purpose only using deep learning models, and some additional methods are necessary, such as regression model and feature engineering method.

Conclusions

This study employs a deep learning model and multi-source data evaluation to propose a systematic framework for identifying potential emerging technologies in large-scale outlier patents. The central tenet of the framework is that two deep learning classifiers are used to fit the relationship between external impacts of every outlier patent and 11 selected static patent indicators. The empirical analysis results in the field of CNC machine tool technology prove the validity of the two deep learning classifiers. The accuracy of the classifier for technological impact levels is 74% and the accuracy of the classifier for social impact levels is 74%. Extending the relationship obtained by the deep learning model to the future, potential emerging technologies in the field of CNC machine tools are identified in large-scale outlier patents, including autonomous sensing and connection, ultrasonic cutting main shaft, surface machining, numerical control chip, surface precision processing, electric main shaft. These findings are valuable for governments and enterprises to identify technology development opportunities and reduce R&D risks.

The contributions of this research are twofold. First, from an academic perspective, this study utilizes the deep learning model to solve the problem that identifying emerging technologies in outlier patents is limited by small samples. Through the usage of the deep learning model, it is possible to discover the complex non-linear relationship between the patent indicators and future impacts in large scale outlier patents, and to assess whether an outlier patent is potential emerging technology based on its future impacts. This study by using the deep learning model to investigate the large scale outlier patens enriches the theory of identifying emerging technologies and extends the application of the deep learning model in the research of innovation management. Furthermore, this study utilizes

multi-source big data including website articles and patent data to identify emerging technologies, which can evaluate both social impacts and technological impacts of outlier patents to increase the reliability. It is important to emphasize that the framework is extensible and other researchers can employ this framework to explore more early signals and evaluation indicators based other multi-source big data for better decision making and strategic planning.

Second, from a practical standpoint, the results provide information about the future prospects of CNC machine tool technology, enabling the quick analysis of wide ranging technologies and supporting decision making, at acceptable levels of time and cost. Moreover, the proposed framework is of more practical use than previous studies, due to the deep learning model, it does not require any distribution assumptions on the input patent indicators which are difficult to selection at early stages of technology development. We expect the proposed framework could be useful as a complementary tool to support governments or enterprises for better decision making and strategic planning.

In spite of meaningful contributions, this study also has limitations, and further study is required. First, though the deep larning model outperforms the RF and LR, more experimentation is needed such as exploring additional early signals to further improve the performance of the framework. Second, deep learning can effectively discover implicit relationships, but due to the complex structure of the model, it is hard to analyze these relationships in depth. Further research needs to discover new causality and causal mechanisms based on these relationships to enrich the relevant theories of emerging technologies.

Acknowledgements This research was supported by the National Natural Science Foundation of China (71974107, 91646102, L1924058, L1824039, L1724034, L1624045 and L1524015), the Ministry of Education in China Project of Humanities and Social Sciences (Engineering and Technology Talent Cultivation) (16JDGC011), the UK-China Industry Academia Partnership Programme (UK-CIAPP\260), the Volvo-supported Green Economy and Sustainable Development Tsinghua University (20153000181), as well as the Chinese Academy of Engineering's China Knowledge Centre for Engineering Sciences an Technology Project (CKCEST-2020-2-5).

References

Adner, R., & Snow, D. (2010). Old technology responses to new technology threats: Demand heterogeneity and technology retreats. *Industrial and Corporate Change, 19*(5), 1655–1675.

Aharonson, B. S., & Schilling, M. A. (2016). Mapping the technological landscape: Measuring technology distance, technological footprints, and technology evolution. *Research Policy, 45*(1), 81–96.

Albertelli, P. (2017). Energy saving opportunities in direct drive machine tool spindles. *Journal of cleaner production, 165*, 855–873.

Aral, S., Dellarocas, C., & Godes, D. (2013). Social media and business transformation: A framework for research. *Information Systems Research, 24*(1), 3–13.

Aristodemou, L., & Tietze, F. (2018). The state-of-the-art on Intellectual Property Analytics (IPA): A literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (IP) data. *World Patent Information, 55*, 37–51.

Askitas, N., & Zimmermann, K. F. (2015). The internet as a data source for advancement in social sciences. *International Journal of Manpower, 36*(1), 2–12.

Balconi, M., Breschi, S., & Lissoni, F. (2004). Networks of inventors and the role of academia: An exploration of Italian patent data. *Research Policy, 33*(1), 127–145.

Bañuls, V. A., & Salmeron, J. L. (2008). Foresighting key areas in the information technology industry. *Technovation*, *28*(3), 103–111.

Bermudez-Edo, M., Noguera, M., Hurtado-Torres, N., Hurtado, M. V., & Garrido, J. L. (2013). Analyzing a firm's international portfolio of technological knowledge: A declarative ontology-based owl approach for patent documents. *Advanced Engineering Informatics, 27*(3), 358–365.
Bessen, J. (2008). The value of US patents by owner and patent characteristics. *Research Policy, 37*(5), 932–945.
Bierly, P., & Chakrabarti, A. (1996). Determinants of technology cycle time in the US pharmaceutical industry'. *R&D Management, 26*(2), 115–126.
Bishop, C. M. (2006). *Pattern recognition and machine learning (information science and statistics)*. New York: Springer.
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 and Technology, 61*(12), 2389–2404.
Breitzman, A., & Thomas, P. (2015). The emerging clusters model: A tool for identifying emerging technologies across multiple patent systems. *Research policy, 44*(1), 195–205.
Chen, J., Hu, P., Zhou, H., Yang, J., Xie, J., Jiang, Y., Gao, Z., & Zhang, C. (2019). Toward Intelligent Machine Tool. *Engineering, 5*(4), 679–690.
Cho, Y. Y., Jeong, G. H., & Kim, S. H. (1991). A Delphi technology forecasting approach using a semi-Markov concept. *Technological Forecasting and Social Change, 40*(3), 273–287.
Choi, S., & Jun, S. (2014). Vacant technology forecasting using new bayesian patent clustering. *Technology Analysis & Strategic Management, 26*(3), 241–251.
Cozzens, S., Gatchair, S., Kang, J., Kim, K. S., Lee, H. J., Ordóñez, G., & Porter, A. (2010). Emerging technologies: quantitative identification and measurement. *Technology Analysis & Strategic Management, 22*(3), 361–376.
Day, G. S., & Schoemaker, P. J. (2000). Avoiding the pitfalls of emerging technologies. *California Management Review, 42*(2), 8–33.
Eaton, W., Wright, W., Whyte, K., Gasteyer, S. P., & Gehrke, P. J. (2014). Engagement and uncertainty: emerging technologies challenge the work of engagement. *Journal of Higher Education Outreach and Engagement, 18*(2), 151–178.
Ernst, H. (1998). Patent portfolios for strategic R&D planning. *Journal of Engineering and Technology Management, 15*(4), 279–308.
Ernst, H. (2003). Patent information for strategic technology management. *World Patent Information, 25*(3), 233–242.
Ernst, H., & Omland, N. (2011). The Patent Asset Index–A new approach to benchmark patent portfolios. *World Patent Information, 33*(1), 34–41.
e Sousa, L. R., Miranda, T., e Sousa, R. L., & Tinoco, J. (2017). The use of data mining techniques in rockburst risk assessment. *Engineering, 3*(4), 552–558.
Fernández-Ribas, A. (2010). International patent strategies of small and large firms: An empirical study of nanotechnology. *Review of Policy Research, 27*(4), 457–473.
Fiore, U., De Santis, A., Perla, F., Zanetti, P., & Palmieri, F. (2017). Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. *Information Sciences, 479,* 448–455.
Gerken, J. M., & Moehrle, M. G. (2012). A new instrument for technology monitoring: Novelty in patents measured by semantic patent analysis. *Scientometrics, 91*(3), 645–670.
Geum, Y., Kim, C., Lee, S., & Kim, M. S. (2012). Technological convergence of IT and BT: Evidence from patent analysis. *Etri Journal, 34*(3), 439–449.
Geum, Y., Lee, S., Yoon, B., & Park, Y. (2013). Identifying and evaluating strategic partners for collaborative R&D: Index-based approach using patents and publications. *Technovation, 33*(6–7), 211–224.
Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge: MIT press.
Guellec, D., & de la Potterie, B. V. P. (2000). Applications, grants and the value of patent. *Economics Letters, 69*(1), 109–114.
Guo, J., Wang, X., Li, Q., & Zhu, D. (2016). Subject–action–object-based morphology analysis for determining the direction of technological change. *Technological Forecasting and Social Change, 105,* 27–40.
Halaweh, M. (2013). Emerging technology: What is it. *Journal of Technology Management & Innovation, 8*(3), 108–115.
Harhoff, D., Narin, F., Scherer, F. M., & Vopel, K. (1999). Citation frequency and the value of patented inventions. *Review of Economics and statistics, 81*(3), 511–515.
Harhoff, D., Scherer, F. M., & Vopel, K. (2003). Citations, family size, opposition and the value of patent rights. *Research Policy, 32*(8), 1343–1363.

Hassan, S. U., Imran, M., Iqbal, S., Aljohani, N. R., & Nawaz, R. (2018). Deep context of citations using machine-learning models in scholarly full-text articles. *Scientometrics, 117*(3), 1645–1662.
Haupt, R., Kloyer, M., & Lange, M. (2007). Patent indicators for the technology life cycle development. *Research Policy, 36*(3), 387–398.
Hiltunen, E. (2008). Good sources of weak signals: A global study of where futurists look for weak signals. *Journal of Futures Studies, 12*(4), 21–44.
Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science, 313*(5786), 504–507.
Injadat, M., Salo, F., & Nassif, A. B. (2016). Data mining techniques in social media: A survey. *Neurocomputing, 214*, 654–670.
Jaffe, A. B., & Trajtenberg, M. (2002). *Patents, citations, and innovations: A window on the knowledge economy*. Cambridge: MIT press.
Jang, H. J., Woo, H. G., & Lee, C. (2017). Hawkes process-based technology impact analysis. *Journal of Informetrics, 11*(2), 511–529.
Jeong, Y., Park, I., & Yoon, B. (2016). Forecasting technology substitution based on hazard function. *Technological Forecasting and Social Change, 104*, 259–272.
Kalampokis, E., Tambouris, E., & Tarabanis, K. (2013). Understanding the predictive power of social media. *Internet Research*, *23*(5), 544–559.
Kayal, A. A., & Waters, R. C. (1999). An empirical evaluation of the technology cycle time indicator as a measure of the pace of technological progress in superconductor technology. *IEEE Transactions on Engineering Management, 46*(2), 127–131.
Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation, 14*(1), 10–25.
Kim, G., & Bae, J. (2017). A novel approach to forecast promising technology through patent analysis. *Technological Forecasting and Social Change, 117*, 228–237.
Köhler, A. R., & Som, C. (2014). Risk preventative innovation strategies for emerging technologies the cases of nano-textiles and smart textiles. *Technovation, 34*(8), 420–430.
Kong, D., Zhou, Y., Liu, Y., & Xue, L. (2017). Using the data mining method to assess the innovation gap: A case of industrial robotics in a catching-up country. *Technological Forecasting and Social Change, 119*, 80–97.
Kreuchauff, F., & Korzinov, V. (2017). A patent search strategy based on machine learning for the emerging field of service robotics. *Scientometrics, 111*(2), 743–772.
Kwon, H., Kim, J., & Park, Y. (2017). Applying LSA text mining technique in envisioning social impacts of emerging technologies: The case of drone technology. *Technovation, 60*, 15–28.
Kyebambe, M. N., Cheng, G., Huang, Y., He, C., & Zhang, Z. (2017). Forecasting emerging technologies: A supervised learning approach through patent analysis. *Technological Forecasting and Social Change, 125*, 236–244.
Lanjouw, J. O., Pakes, A., & Putnam, J. (1998). How to count patents and value intellectual property: The uses of patent renewal and application data. *The Journal of Industrial Economics, 46*(4), 405–432.
Lanjouw, J. O., & Schankerman, M. (1997). *Stylized facts of patent litigation: Value, scope and ownership* (No. w6297). National Bureau of Economic Research.
Lanjouw, J. O., & Schankerman, M. (2001). Characteristics of patent litigation: A window on competition. *RAND Journal of Economics, 32,* 129–151.
Lee, C., Cho, Y., Seol, H., & Park, Y. (2012). A stochastic patent citation analysis approach to assessing future technological impacts. *Technological Forecasting and Social Change, 79*(1), 16–29.
Lee, C., Kim, J., Kwon, O., & Woo, H. G. (2016). Stochastic technology life cycle analysis using multiple patent indicators. *Technological Forecasting and Social Change, 106*, 53–64.
Lee, C., Kwon, O., Kim, M., & Kwon, D. (2018). Early identification of emerging technologies: A machine learning approach using multiple patent indicators. *Technological Forecasting and Social Change, 127*, 291–303.
Lee, S., Kim, W., Kim, Y. M., Lee, H. Y., & Oh, K. J. (2014). The prioritization and verification of IT emerging technologies using an analytic hierarchy process and cluster analysis. *Technological Forecasting and Social Change, 87*, 292–304.
Lee, S., Yoon, B., Lee, C., & Park, J. (2009). Business planning based on technological capabilities: Patent analysis for technology-driven roadmapping. *Technological Forecasting and Social Change, 76*(6), 769–786.
Lerner, J. (1994). The importance of patent scope: An empirical analysis. *The RAND Journal of Economics, 25,* 319–333.
Li, A., Zhang, G., Zhang, Z., Zhang, Y., Yang, K., & Qiang, H. (2017). Recent patents on design and simulation of dual-driving electric spindles. *Recent Patents on Mechanical Engineering, 10*(4), 326–335.

Li, S., Hu, J., Cui, Y., & Hu, J. (2018). DeepPatent: Patent classification with convolutional neural networks and word embedding. *Scientometrics, 117*(2), 721–744.
Li, X., Xie, Q., Jiang, J., Zhou, Y., & Huang, L. (2019). Identifying and monitoring the development trends of emerging technologies using patent analysis and Twitter data mining: The case of perovskite solar cell technology. *Technological Forecasting and Social Change, 146*, 687–705.
Liu, Y., Zhou, Y., Liu, X., Dong, F., Wang, C., & Wang, Z. (2019). Wasserstein GAN-based small-sample augmentation for new-generation artificial intelligence: A case study of cancer-staging data in biology. *Engineering, 5*(1), 156–163.
Ma, Z., & Lee, Y. (2008). Patent application and technological collaboration in inventive activities: 1980–2005. *Technovation, 28*(6), 379–390.
Martin, B. R. (1995). Foresight in science and technology. *Technology Analysis & Strategic Management, 7*(2), 139–168.
Martinov, G. M., & Kozak, N. V. (2016). Specialized numerical control system for five-axis planing and milling center. *Russian Engineering Research, 36*(3), 218–222.
Meyer, M. (2000). Does science push technology? Patents citing scientific literature. *Research policy, 29*(3), 409–434.
Meyer, M. (2006). Are patenting scientists the better scholars?: An exploratory comparison of inventor-authors with their non-inventing peers in nano-science and technology. *Research Policy, 35*(10), 1646–1662.
Narin, F., Hamilton, K. S., & Olivastro, D. (1997). The increasing linkage between US technology and public science. *Research policy, 26*(3), 317–330.
Narin, F., Noma, E., & Perry, R. (1987). Patents as indicators of corporate technological strength. *Research policy, 16*(2–4), 143–155.
Noh, H., Song, Y. K., & Lee, S. (2016). Identifying emerging core technologies for the future: Case study of patents published by leading telecommunication organizations. *Telecommunications Policy, 40*(10–11), 956–970.
OuYang, K., & Weng, C. S. (2011). A new comprehensive patent analysis approach for new product design in mechanical engineering. *Technological Forecasting and Social Change, 78*(7), 1183–1199.
Pakes, A., & Schankerman, M. (1984). The rate of obsolescence of patents, research gestation lags, and the private rate of return to research resources. In *R&D, patents, and productivity* (pp. 73–88). University of Chicago Press.
Porter, A. L., & Detampel, M. J. (1995). Technology opportunities analysis. *Technological Forecasting and Social Change, 49*(3), 237–255.
Porter, A. L., Roessner, J. D., Jin, X. Y., & Newman, N. C. (2002). Measuring national 'emerging technology'capabilities. *Science and Public Policy, 29*(3), 189–200.
Raford, N. (2015). Online foresight platforms: Evidence for their impact on scenario planning & strategic foresight. *Technological Forecasting and Social Change, 97*, 65–76.
Rotolo, D., Hicks, D., & Martin, B. R. (2015). What is an emerging technology? *Research Policy, 44*(10), 1827–1843.
Sakata, I., Sasaki, H., Akiyama, M., Sawatani, Y., Shibata, N., & Kajikawa, Y. (2013). Bibliometric analysis of service innovation research: Identifying knowledge domain and global network of knowledge. *Technological Forecasting and Social Change, 80*(6), 1085–1093.
Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks, 61,* 85–117.
Song, B., Seol, H., & Park, Y. (2016). A patent portfolio-based approach for assessing potential R&D partners: An application of the Shapley value. *Technological Forecasting and Social Change, 103*, 156–165.
Song, K., Kim, K., & Lee, S. (2018). Identifying promising technologies using patents: A retrospective feature analysis and a prospective needs analysis on outlier patents. *Technological Forecasting and Social Change, 128*, 118–132.
Sternitzke, C., Bartkowski, A., & Schramm, R. (2008). Visualizing patent statistics by means of social network analysis tools. *World Patent Information, 30*(2), 115–131.
Sun, Y., Kamel, M. S., Wong, A. K. C., & Wang, Y. (2007). Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition, 40*(12), 3358–3378.
Tong, X., & Frame, J. D. (1994). Measuring national technological performance with patent claims data. *Research Policy, 23*(2), 133–141.
Trajtenberg, M. (1990). *Economic analysis of product innovation: The case of CT scanners (Vol. 16)*. Cambridge: Harvard University Press.
Trappey, C. V., Wu, H. Y., Taghaboni-Dutta, F., & Trappey, A. J. (2011). Using patent data for technology forecasting: China RFID patent analysis. *Advanced Engineering Informatics, 25*(1), 53–64.

Wang, Y., & Xu, W. (2018). Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. *Decision Support Systems, 105*, 87–95.

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

Yoon, J., & Kim, K. (2012). Detecting signals of new technological opportunities using semantic patent analysis and outlier detection. *Scientometrics, 90*(2), 445–461.

Zhang, X., Sui, H., Zhang, D., & Jiang, X. (2018a). Study on the separation effect of high-speed ultrasonic vibration cutting. *Ultrasonics, 87*, 166–181.

Zhang, Y., Lu, J., Liu, F., Liu, Q., Porter, A., Chen, H., et al. (2018b). Does deep learning help topic extraction? A kernel k-means clustering method with word embedding. *Journal of Informetrics, 12*(4), 1099–1117.

Zhou, J. (2015). Intelligent manufacturing—main direction of "made in China 2025". *China Mechanical Engineering, 26*(17), 2273–2284.

Zhou, J., Li, P., Zhou, Y., Wang, B., Zang, J., & Meng, L. (2018). Toward new-generation intelligent manufacturing. *Engineering, 4*(1), 11–20.

Zhou, Y., Dong, F., Kong, D., & Liu, Y. (2019a). Unfolding the convergence process of scientific knowledge for the early identification of emerging technologies. *Technological Forecasting and Social Change, 144*, 205–220.

Zhou, Y., Dong, F., Liu, Y., Li, Z., Du, J., & Zhang, L. (2020). Forecasting emerging technologies using data augmentation and deep learning. *Scientometrics*, *123*(1), 1–29.

Zhou, Y., Lin, H., Liu, Y., & Ding, W. (2019b). A novel method to identify emerging technologies using a semi-supervised topic clustering model: a case of 3D printing industry. *Scientometrics*, *120*(1), 167–185.

Zhuang, Y. T., Wu, F., Chen, C., & Pan, Y. H. (2017). Challenges and opportunities: from big data to knowledge in AI 2.0. *Frontiers of Information Technology & Electronic Engineering, 18*(1), 3–14.