Contents lists available at ScienceDirect

Technological Forecasting & Social Change

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

Combining topic modeling and SAO semantic analysis to identify technological opportunities of emerging technologies

Tingting Ma[a], Xiao Zhou[b,*], Jia Liu[c,d], Zhenkai Lou[e], Zhaoting Hua[a], Ruitao Wang[a]

[a] School of Logistics, Beijing Wuzi University, No.321 Fu He Street, Tongzhou, Beijing 101149, PR China
[b] School of Economics and Management, Xidian University, No. 266 Xinglong Section of Xifeng Road, Xi'an, Shaanxi 710126, PR China
[c] State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024, PR China
[d] School of Economics & Management, Communication University of China, Beijing 100024,PR China
[e] School of Management Science and Engineering, Anhui University of Technology, Maanshan, Anhui 243032, PR China

ARTICLE INFO

Keywords:
Technological opportunities analysis
Topic extraction
Topic modeling
Semantic SAO analysis
Dye sensitized solar cell

ABSTRACT

With the advancement of science and the emergence of new technologies, technology opportunities analysis has attracted increasing attention from both society and academia. This study proposes a hybrid approach to integrate topic modeling, semantic SAO analysis, machine learning, and expert judgment, identifying technological topics and potential development opportunities. The systematical methodology is applied to analyze a set of 9,883 Derwent Innovation Index (DII) patents related to the dye-sensitized solar cell to present its potential contribution of technical intelligence for R&D management. Also, how the approach is validated and optimized is illustrated. The main contributions of this paper are two-fold. First, an optimized topic extraction model with high accuracy is constructed, considering both the patent classification codes and term location. Second, we integrate the topic modeling, SAO technique, and machine learning to explore semantic relationships among technological topics represented as a suite of terms. The methodology overcomes some drawbacks of the current studies. It can be used as a powerful tool for technological opportunities analysis.

1. Introduction

Emerging technologies have been rapidly advanced these days and thus provide tremendous opportunities as well as the uncertainty of technological innovation (Zhu and Porter, 2002). Accordingly, accurate and timely ability to identify technological opportunities has become a key factor in ensuring innovation with competitive advantages for both countries and enterprises (Zhu and Porter, 2002). Many efforts, known as technological opportunities analysis (TOA), were made to identify promising technologies and predict the development paths by tracking recent developments of the technologies (Lee et al., 2015).

The TOA first relied on expert judgment (Lee et al., 2017). However, the expert depending approach suffered low reliability due to the increased complexity of technology, environment, and mutual interactions (Lee et al., 2017). Further, experts could not gain a consensus in evaluating and forecasting the emerging technology due to its ambiguity and uncertainty(Zhou et al., 2019). In order to overcome those limitations, text mining and bibliometric methods were explored. The methods mined technical intelligence from the Science, Technology & Innovation (ST&I) data to provide an objective argument for TOA. Also, the effective methodology and process of TOA were studied, aiming to identify and predict hot spots and technology emergence, evolving paths, technological gaps, and vacancies (Jing Ma et al., 2019). This paper focuses on identifying promising technologies through the topic information analysis and exploring paths evolution through the technical topics links.

As outputs of R&D activities, patents are embedded with rich technical information and are evaluated and generated to an international standard (Yoon and Magee, 2018). Thus, we adopted the patents as the data source for the analysis of technology opportunities. However, unlike scientific articles, patents do not include keywords to categorize their content, causing the difficulty of analyzing topical information from patents (Ma and Porter, 2015). Therefore, extracting topical information of patents becomes a key research question in patent mining. To this end, several methods have been explored. Earlier studies used the International Patent Classification (IPC) to substitute a keyword to identify the technological topics. However, it was too rough, and most of the technical information contained in the content text was lost. Recent

[*] Corresponding author.
E-mail addresses: matingtingmay@163.com (T. Ma), belinda1214@126.com (X. Zhou), louzk@ahut.edu.cn (Z. Lou).

https://doi.org/10.1016/j.techfore.2021.121159
Received 30 June 2020; Received in revised form 19 August 2021; Accepted 23 August 2021
Available online 6 September 2021

studies employed the natural language processing (NLP) technique to extract terms from patent documents. Also, a screening method, such as term clumping, was employed to improve the efficiency and accuracy of topic recognition. These studies only focused on extracting patent topic information but ignored the IPC that can contribute to technological topic recognition. In addition, according to the argumentation zoning theory (Teufel et al., 2009), the terms positioned in different parts of scientific articles may have different values. For example, the terms used in the title best reflect the theme of the article. Inspired by this aspect, we attempt to design a feature selection model to figure out the following two questions. 1) whether assigning terms of title higher weight can improve the accuracy of patent topic recognition? 2) do the patent classification codes contribute to topic recognition if integrating them with patent terms? An optimized topic extraction model is proposed by answering these two questions and then applied to identify trending and promising technologies.

Description of the evolving paths by linking identified technological topics is one of the key issues for TOA. Several methods were proposed to identify the linkage between technologies, including co-word analysis, citation network analysis, similarity measurement method, and Subject-Action-Object (SAO) semantic analysis. However, each of these methods had its own limitations. The citation network analysis was only available to analyze ST&I data with reference information. The co-word analysis was suitable for studying the majority of ST&I data. However, it could only be used to link ’points,’ i.e., it was a technology targeting a very specific goal (Porter and Detampel, 1995; Zhou et al., 2013). The similarity measurement linked technological topics by using the similarity between clusters that contain a suite of related technologies. It could supplement the shortcomings of the co-word analysis. Still, it could not reveal the semantic relationships among topics (Zhang et al., 2016a). The SAO semantic analysis has become a trendy method recently due to its advantage in mining semantic relations. However, the SAO semantic analysis only finds the relationship between one ’point’ and another ’point,’ rather than identifying links between topics. To overcome these drawbacks, we integrate SAO semantic analysis with topic modeling and machine learning, exploring the hidden semantic relationships between technological topics. Then, the identified technological topics and the semantic connections are plotted to construct a technology roadmap (TRM). The constructed TRM is used to trace the development trajectory and forecast the future development directions of emerging technologies.

In summary, we propose a hybrid approach based on topic modeling and SAO semantic analysis to identify potential technologies and predict promising evolution paths of emerging technology. Also, a feature selection model is designed to improve the performance of topic recognition. The remainder of this paper is organized as follows. Section 2 presents the theoretical background of the research framework. In Section 3, a detailed research methodology and process are described. In Section 4, an illustration is proposed to approach the development of the hybrid methodology (keeping DSSCs in view as a case study). Lastly, Section 5 concludes this paper with a summary, limitations, and future works.

2. Related works and theoretical background

2.1. Theoretical foundation: how technological opportunities emerged by the process of science revolutions

Thomas Kuhn argued, in his masterpiece "The Structure of Science Revolutions" in 1962, that the scientific revolution can be viewed as a process: normal science (old paradigms) experiences anomalies culminating in crises, which lead to revolutionary science (new paradigms) (Kuhn, 1962). Typically, normal scientific progress is considered as "development-by-accumulation" of accepted facts and theories. However, science is an episodic model in which periods of conceptual continuity are interrupted by periods of revolutionary science. The discovery of "anomalies" leads to scientific revolutions and new paradigms that direct new research in the future (Kuhn, 1962). After 50 years of development as a guide for identifying future technology opportunities by the process of scientific evolution, this theory has been applied to a wide range of scientific fields – sociology, medical science, and many others.

This research aims to identify anomalous that show significant promise for future problem-solving and, hence, future technology opportunities. As mentioned, Kuhn’s theory focused on the emergence of scientific discoveries, contributing to the progress of scientific revolutions (Kuhn, 1962). Kuhn’s theory provides an excellent theoretical roadmap for this research.

2.2. Technological opportunities analysis (TOA)

Technological opportunity is defined as a set of possibilities for technology advances to enhance product functions or production (Olsson, 2005). Many researchers have been devoted to developing effective methods to identify and predict technological opportunities. Earlier studies mainly used qualitative methods, such as Delphi and Workshop, to obtain experts’ opinions and judgments on technologies. However, expert judgment was often limited and biased by personal knowledge. Further, experts sometimes cannot reach a consensus. To overcome these drawbacks, Alan Porter et al. firstly introduced bibliometrics to analyze the technological opportunities of emerging technologies (Porter and Detampel, 1995). Bibliometrics quantitatively measured the number of scientific articles, patents, and citations and interpreted technological advances to assess research and development (R&D) activity levels (Guo et al., 2012; Porter and Detampel, 1995). It also ascertained important links by analyzing organizations that published scientific articles and patents, co-occurrence of topics, and citations to identify R&D activity patterns (Porter and Detampel, 1995). Bibliometrics provided quantitative data and an objective basis for experts to assess technological opportunities and reach a consensus.

However, bibliometrics is limited in analyzing the contents of documents. Text mining was introduced to analyze the contents due to its significance in understanding the characteristics of technical documents (Yoon and Magee, 2018). Text mining was initially used to analyze technological topics information by extracting terms from the contents of ST&I documents. Quantitative indicators(e.g., frequency and TFIDF) were used to measure core terms considered as technologies (Ma et al., 2014; Yoon and Park, 2005; Zhou et al., 2013). Also, the terms were used in the co-word analysis(Lee and Su, 2011) and morphology analysis (Yoon and Park, 2005; Yoon and Park, 2007) to explore relationships among the technologies. However, the selection of core terms is labor-intensive and time-consuming, and the co-word analysis (Lee and Su, 2011) and morphology analysis only identified the links between ’points’ rather than topics. More recent studies explored statistic and data mining methods (Chiu and Hong, 2015; Zhang et al., 2016b), such as Principle Component Analysis (PCA) and Text Clustering(Zhou et al., 2019), to extract topic information by clustering terms or documents. Then, the topic similarity measurement was employed to identify linkage among technogical topics (Zhang et al., 2016b). Recently, with the development of machine learning and NLP techniques, topic modeling was introduced to analyze topic information from technical text, which provided an advantage in solving polysemy problems (Pavlinek and Podgorelec, 2017). Besides, SAO semantic analysis (Wang et al., 2017) was employed to discover semantic relationships among technologies. These two advanced methods have attracted increasing attention in recent studies.

Scientific articles and patents have been two major data sources in TOA because they are public and well-organized. Patents were more often used in the technological opportunity analysis due to their abundant technical information (Yoon and Magee, 2018). In contrast, scientific articles contain information scientific rather than technological (Ma et al., 2017). Further, patent documents follow a global standard format, which makes the analysis easier. The World International Property Organization (WIPO) developed a procedure for standardizing the bibliographical data in patent documents, using a minimum set of data followed by every member country of WIPO. One of the most important is the International Patent Classifications (IPC), which groups patents according to the technologies they pertain to. Accordingly, the IPC can be used as a key indicator of technological topics.

The United States Patent and Trademark Office (USPTO) and the Derwent Innovation Index (DII) are the most popular patent databases in the TOA field. In this study, DII patents are selected because of the following considerations. First, the DII database collects patents from more than 40 patent offices worldwide, while the USPTO database only collects patents applied in the United States. Second, most descriptions in titles & abstracts of the USPTO patents are obscure due to technology protection. Although the USPTO patent claims can provide more detailed technical information, more complicated and confusing attorney language is used. Also, it is preferably used in a legal context to analyze patent infringement cases rather than technological opportunities. In contrast, the titles and abstracts of the DII patents are rewritten by domain experts who read the full texts of the original patents. These contents are written in easy-to-understand language, which supports the idea to draw information from semantic sections. The rewritten title highlights the novelty of the invention disclosed in the patent description, while the rewritten abstract summarizes the main claims as well as the novelties, detailed description, use cases, advantages of the invention. These characteristics of DII patents make up for the disadvantage of the ambiguity in most patent data. Accordingly, the titles and abstracts were often used to analyze topic information (Ma and Porter, 2015; Ma et al., 2014). Third, an invention patented in multiple nations/regions may have varied IPCs due to the different use and understanding of IPC in patent offices of nations/regions. In order to solve this problem, the DII database creates the basic patent to describe its patent family and assigns appropriate Derwent Manual Codes (DMCs) to the basic patent for unifying the classification of its patent family. Specifically, the DMCs are assigned to DII patents by professional indexing staff, indicating the technological innovation and application of an invention. Thus, the accuracy of patent retrieval can be significantly improved by the DMCs, implying that the DMCs, similar to the IPCs, also contains information that can reveal the technological topic.

To sum up, the title, abstract, IPC, and DMC are four types of tech-related data features of the DII patent used for technological topic identification. However, in the previous studies, these features were separately used to extract topic information. Combining these features was not explored, and thus further study is needed.

2.3. Topic extraction and modeling

Accurate topic extraction from ST&I data is significant for exploring valuable technologies. The widely-used tools for topic extraction include text clustering and topic modeling (Zhang et al., 2016b). The text clustering methods, such as K-means clustering and hierarchical agglomerative clustering (HAC), identify topics based on the similarity between documents, emphasizing statistical properties and connections of words or phrases (Zhang et al., 2016b). Under the assumption that each document is related to only one topic, it outputs a solid clustering result for grouping documents (Chen et al., 2019). However, text clustering is not suitable for documents with highly coupled topics.

The topic modeling was originally developed and branched off from the subject area of 'generative probabilistic modeling' (Lin et al., 2016). It assumes that each document in the collection incorporates a small number of topics (Chen et al., 2015). It is a soft clustering method that can provide the probabilities of how various documents belong to different topics (Blei, 2012). Hence, topic modeling has the advantage of handling the collection of documents with highly coupled topics. Also, it has an advantage in solving polysemy problems (Pavlinek and Podgorelec, 2017).

Due to these advantages, probabilistic-based topic modeling has been widely used for topic extraction. The most popular topic modeling techniques are probabilistic Latent Semantic Indexing (pLSI) (Hofmann, 1999) and Latent Dirichlet Allocation (LDA) (Blei et al., 2003). The pLSI modeled each word in a document as a sample from a mixture model, where the mixture components were multinomial random variables that can be viewed as representations of topics. However, the number of parameters in the pLSI model was linearly increased according to the topic number, often leading to over-fitting (Yau et al., 2014). To overcome this drawback, the LDA model was proposed based on a three-layer Bayesian model (Blei et al., 2003). The LDA has been widely used in various applications, such as text mining, bioinformatics, and image processing. LDA is also used to extract topics from large volumes for textual data (Xing and Croft, 2007). For instance, Chen et al. (Chen et al., 2017) proposed an LDA-based model to analyze technological topic changes in Australia during 2000-2014. Erzurumlu and Pachamanova (Erzurumlu and Pachamanova, 2020) applied the LDA algorithm to discover topics from patents to forecast promising technologies in the healthcare application. Griffiths and Steyvers (Griffiths and Steyvers, 2004) employed LDA-based topic modeling to find the hot topics covered by the scientific journal of PNAS. Inspired by its success in the topic information extraction, we adopt the LDA to detect latent topics from DII patents.

2.4. Technology roadmap and SAO semantic analysis

The TRM is defined as a future-oriented strategic planning approach to connect technologies, products, and markets over time (Phaal et al., 2004; Zhang et al., 2016a). The traditional TRM depended on domain experts, which was time-consuming and labor-intensive (Zhang et al., 2014b). In order to improve the efficiency, several traditional bibliometrics and text mining methods were proposed to construct an intelligent TRM composing models, for examples, citation network, co-occurrence map, and similarity measurement (Chaomei, 2006; Shibata et al., 2011; Waltman et al., 2010; Zhang et al., 2016b; Zhu and Porter, 2002). Bibliometrics and text mining methods were used as data pre-processing followed by experts' evaluation and refinement (Zhang et al., 2016b). The citation network is only available for analyzing data with citation information, and the co-occurrence map is used to find linkages among 'points,' limiting it to a specific target (Porter and Detampel, 1995; Zhou et al., 2013). In contrast, the similarity measurement identified critical linkages between topics, avoiding the above shortcomings. However, it could not reveal the semantic relationship among topics (Zhang et al., 2016a).

With the development of NLP technologies, SAO semantic analysis has become a popular method to support intelligent TRM construction with its ability in semantic structure extraction. In this method, the subject (S) and the object (O) are nouns or phrases that represent the components, and the action (A) is the verb or verb phrase that reflects the relationship between components in the invention (Wang et al., 2017). For example, in the SAO structure of "battery contains an anode," 'battery' (S) and 'anode' (O) denote technologies, and 'contains' denotes the relationship between the two technologies, indicating that anode is part of the battery. In this way, SAO semantic analysis explores the semantic relationship between technologies, enabling researchers to glean a complete understanding of the connection and evolution between technologies (Zhou et al., 2020). However, existing SAO semantic analysis methods only find the relationship between two 'points' rather than identifying linkages between topics that contain suites of related technologies (Zhang et al., 2014a).

Theoretically, the semantic relationship between topics can be identified by summarizing the relationship type (A: verbs or verb phrases) between two technical term groups (S&O: nouns or phrases). For this, the following two problems should be addressed. 1) Determining types of semantic relationships between technologies, and 2) classifying and grouping SAO structures according to the types of

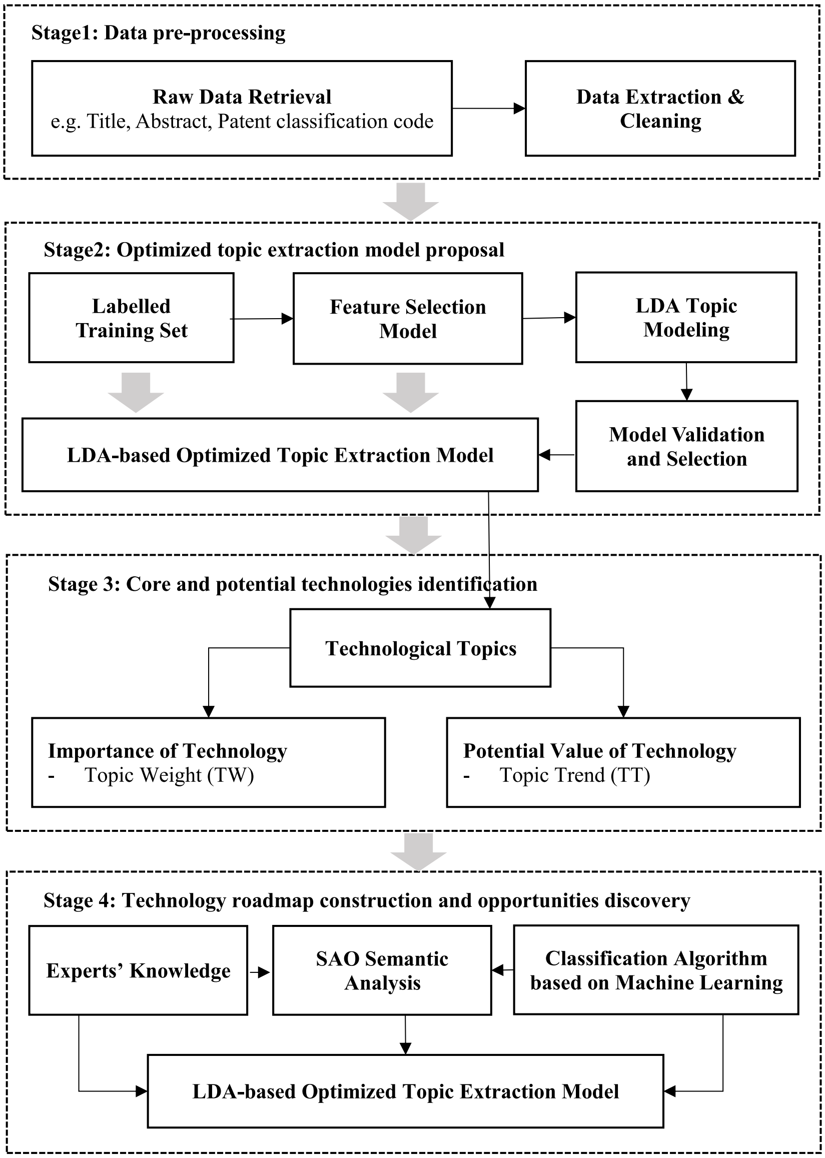

Stage1: Data pre-processing
Raw Data Retrieval
e.g. Title, Abstract, Patent classification code
Data Extraction & Cleaning
Stage2: Optimized topic extraction model proposal
Labelled Training Set
Feature Selection Model
LDA Topic Modeling
Model Validation and Selection
LDA-based Optimized Topic Extraction Model
Stage 3: Core and potential technologies identification
Technological Topics
Importance of Technology
- Topic Weight (TW)
Potential Value of Technology
- Topic Trend (TT)
Stage 4: Technology roadmap construction and opportunities discovery
Experts' Knowledge
SAO Semantic Analysis
Classification Algorithm based on Machine Learning
LDA-based Optimized Topic Extraction Model

Fig. 1. Research framework for technological opportunities analysis.

semantic relationships. Inspired by the previous works (Wang et al., 2017; Xin et al., 2008), we select four main types of semantic relationships between technologies to classify SAO semantic structures. Also, we construct a supervised classifier using labeled SAO structures to identify the key relationships for each SAO group (the topic pair). Here, we select a naïve Bayesian algorithm to conduct machine learning due to its higher accuracy than Decision Tree, Maximum Entropy algorithms, and Logistic Regression (Zhou et al., 2020).

3. Methodology

Kuhn's theory focuses on the emergence of scientific discoveries that contribute to the progress of scientific revolutions. Given this research aims to predict future technology opportunities, the theory is an excellent guide for setting out the decomposable steps to identify technology opportunities and their evolving ways. Accordingly, in this paper, we develop a hybrid approach to identify core and potential technologies with the exploration of their evolving paths. The proposed framework consists of four stages, as illustrated in Fig. 1: (1) Data pre-processing; (2) Optimized topic extraction model proposal; (3) Core and potential technologies identification; (4) Technology roadmap construction and opportunities discovery.

(1) Stage 1 - Data pre-processing

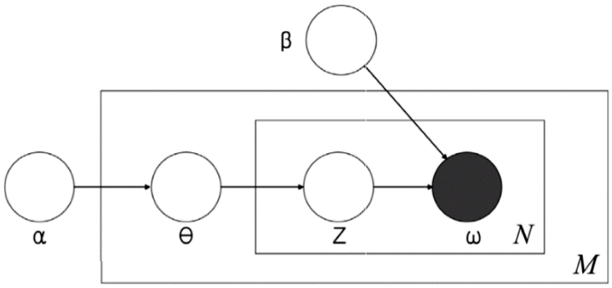
We select the DII database as our data source. It contains all published patents regardless of whether it is granted or under review. In this stage, the tech-related data features are extracted and cleaned from raw data of the DII patent, including title, abstract, IPC, and DMC. First, we retrieve raw patent data from the DII database in Web of Science using a multi-step Boolean search algorithm. Second, the raw data is imported into Vantagepoint software to extract the specific data features (terms of Title and Abstract, IPCs, and DMCs). Third, the term clumping process based on the ClusterSuite [program developed by J.J. O'Brien, with Stephen J. Carley, at Georgia Tech –available at www.VPInstitute.org] is applied to remove meaningless terms (Zhang et al. 2014a). Lastly, the low frequent (< 2) terms, IPCs, and DMCs are removed to improve operational efficiency.

(2) Stage 2 - Optimized topic extraction model proposal

In this stage, we firstly construct a feature selection model and a

Table 1
Feature Selection Model.

Group	No.	Model Structure
A	#1	Title + Abstract Terms
	#2	Title + Abstract Terms + IPC
	#3	Title + Abstract Terms + DMC
	#4	Title + Abstract Terms + IPC+DMC
B	#1	Weighed Title + Abstract Terms
	#2	Weighed Title + Abstract Terms + IPC
	#3	Weighed Title + Abstract Terms +DMC
	#4	Weighed Title + Abstract Terms + IPC+DMC



- ✓ Choose $\theta_i \sim Dir(\alpha)$, where $i \in \{1, \ldots, M\}$
- ✓ Choose $\varphi_k \sim Dir(\beta)$, where $k \in \{1, \ldots, K\}$
- ✓ For each word position i, j where $j \in \{1, \ldots, N_i\}$, and $i \in \{1, \ldots, M\}$
  - Choose a topic $z_{ij} \sim Multinomial(\theta_i)$.
  - Choose a word $w_{ij} \sim Multinomial(\varphi_{z_{ij}})$,

Fig. 2. Concept and generative process of LDA model (Blei et al., 2003).

training dataset. Then, for each of these assemble sets, LDA is applied to generate topics from the labeled patents, and each patent is assigned to the topic with the highest probability. The input of the LDA model is a 'Document-Feature' Matrix in which each cell indicates the frequency of feature data in each document. Then, we compare the assigned topics with actual categories of the patents. Through this process, the topic identification accuracies of different combinations are evaluated and compared. Finally, the best combination of features is selected. The detailed descriptions for the feature selection model and validation method are given in the following.

1) The feature selection model

Title, abstract, and IPC were typically used as tech-related features of patents in the existing technical topic identification methods. Title and abstract contain information on technical novelties and details, while IPC reveals technology category. In addition to those features, DII patents provide a particular field – DMC, a kind of classification code that contains information on technical novelties and applications. Similar to IPC, DMC can also significantly improve the accuracy of patent retrieval. Accordingly, in the proposed model, title, abstract, IPC, and DMC are used as features.

We first construct the A-#1 model using the most commonly used features, the terms of title and abstract (as shown in Table 1). Then, we construct A-#2, A-#3, and A-#4 models to evaluate 1) the accuracy improvement by technical classification codes and 2) the influence of classification code type on the topic recognition accuracy. In addition, four corresponding weighted models: B-#1, B-#2, B-#3, and B-#4 are constructed in group B to evaluate 3) the relative significance between title and abstract. Following Cao and Jia (2013) and Li and Ma (2008) researches, we set the weight ratio of title and abstract as 2:1 in our study.

2) Latent Dirichlet Allocation (LDA)

LDA (Blei et al., 2003) is adopted to generate topics and cluster patents due to its ability to process large-scale documents and interpret latent topics (Jeong et al., 2019). The process of the LDA consists of three steps, as depicted in Fig. 2. Let denote D be a corpus consisting of M documents, each of length N_i, K be the topic numbers for D. α and β are the parameters of the Dirichlet prior to the per-document topic distribution and the per-topic word distribution, respectively. θ_i is the topic distribution for document i, ϕ_k is the word distribution for topic k, Z_{ij} is the topic for the j^{th} word in document i, and w_{ij} is the specific word.

The goal is to find the θ_i and ϕ_k, targeting the topic distribution for each document and the content of topics. We use Gibbs sampling (Griffiths and Steyvers, 2004) to sample the latent variable z. Based on z, we infer the topic distribution for each document and the content of each topic (Yau et al., 2014). In addition, the appropriate number of topics is determined based on the perplexity (Blei et al., 2003). The perplexity is a popular indicator to evaluate the probability model, defined as the reciprocal geometric mean of the likelihood of a test corpus (Glenisson et al., 2005; Huang et al., 2018). A lower perplexity score indicates a lower misrepresentation of the words in the corpus (Zhang et al., 2017), mathematically formulated as follows:

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

where N_d is the document length of document d in corpus D, including M documents, and log(p(w)) represents the likelihood of a corpus given the model.

3) The validation method

The latent topics are generated from the labeled training data via LDA and then manually assigned into specific categories based on the most common words in each topic. The aggregated distribution of latent topics associated with the same category provides the correlations of documents with each category. Based on this, we assign each document to the category with the highest proportion. Finally, the total precision (TP) (Zhang et al., 2016b) is measured on all documents connected to categories to evaluate the topic recognition performance. In such a way, the best feature combination is determined to support the subsequent identification of core and potential technologies. The total precision is defined as follows:

$$\text{Total Precision} = \frac{\text{Number of records clustered to correct topic}}{\text{Total number of records}} \tag{2}$$

(3) Stage 3 - Core and potential technologies identification

In this stage, meaningful core and potential technologies are identified through the automatic extraction of technological topics from patents by the optimized LDA-based topic model.

We introduce the indicator of Topic Weight (TW) to measure the importance of topics, computed by the sum of the distribution probabilities of each topic in the collected corpus. The LDA model outputs latent topics and their distribution for each patent, representing a "Document - Topic" Matrix. The patent is assigned to multiple topics with a certain probability rather than matched only one technological topic. The probability distribution of a topic in a patent reflects the frequency that the topic is referred to in the patent. Also, the frequency indicates relative significance. Accordingly, the importance of the technological topic can be measured by the sum of the distribution probabilities of each technological topic in the collected corpus (Jeong et al., 2019), defined as Topic Weight (TW) (Eq. (3)).

$$\begin{array}{r|ccccc|l}
 & Topic1 & Topic2 & Topic3 & \cdots & \boldsymbol{Topic}\ \mathrm{K} & \\
Document1 & 0.0027 & 0.0382 & 0.0398 & \cdots & 0.0938 & \multirow{3}{*}{\left.\right\} Year1} \\
Document2 & 0.1903 & 0.0943 & 0 & \cdots & 0.0483 & \\
\vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \\
Document401 & 0.0982 & 0.0763 & 0.0387 & \cdots & 0.0273 & \multirow{3}{*}{\left.\right\} Year2} \\
Document402 & 0.0058 & 0.1983 & 0.2934 & \cdots & 0 & \\
\vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \\
Document851 & 0 & 0.0395 & 0.0964 & \cdots & 0.0498 & \vdots \\
\vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \left.\right\} Year\ T \\
Document\ \mathrm{M} & 0.0049 & 0 & 0.0984 & \cdots & 0.1928 &
\end{array}$$

Fig. 3. An example of a topic distribution matrix in chronological order.

Table 2
Search terms of DSSC.

Set	Search Terms	Records
#1	TS= (((dye-sensiti*) or (dye* same sensiti*) or (pigment-sensiti*) or (pigment same sensiti*) or (dye* adj sense)) same ((solar or Photovoltaic or photoelectr* or (photo-electr*)) same (cell or cells or batter* or pool*)))	9,662
#2	TS= (((dye- Photosensiti*) or (dye same Photosensiti*) or (pigment- Photosensiti*) or (pigment same Photosensiti*)) same ((solar or Photovoltaic or photoelectr* or (photo-electr*)) same (cell or cells or batter* or pool*))) and IP=(H01G* or H01M* or H01L* or G03C*)	740
#3	TS= (((dye- optoelectri*) or (dye same optoelectri*) or (pigment-optoelectri*) or (pigment same optoelectri*) or (dye- opto-electri*) or (dye same opto-electri*) or (pigment- opto-electri*) or (pigment same opto-electri*)) same ((solar or Photovoltaic or photoelectr* or (photo-electr*)) same (cell or cells or batter* or pool*))) and IP=(H01G* or H01M* or H01L* or G03C*)	312
#4	#1 or #2 or #3	9,883

Table 3
The Total Precision of the eight models in the DSSC case (%).

	#1	#2	#3	#4
A	67.36	68.77 (70.18)	67.36 (70.29)	69.21 (71.64)
B	69.97	70.38 (73.06)	70.04 (73.95)	71.06 (**74.73**)

$$\text{Importance of topic } \mathrm{j} = \mathrm{TW}_j = \sum_{i=1}^{i=m} p_{ij}, i = (1, 2, \cdots, M), j = (1, 2, \cdots, K) \quad (3)$$

where p_{ij} denotes the probability distribution of i^{th} document on j^{th} topic. M and K represent the total numbers of documents in the corpus and generated technological topics, respectively.

The developing trend of the discovered topics is also significant, which evaluates how a topic is keeping up with the time (Chen et al., 2019). To consider this, we introduce the Topic Trend (TT) as an indicator to measure its potential, computed by the average growth rate of the annual weight of the topic. Also, the change of TW reflects the trend of research attitude and focus (Chen et al. (2017). The growth of topic weight indicates the corresponding topic is favored and receives increasing attention. Accordingly, we use the average growth rate of the annual weight of the topic to measure its potential. The annual weight of the topic is calculated as: the "Document-Topic" distribution Matrix is firstly ranked by application years of patents, as shown in Fig. 3. Then we sum a group of elements in a column associated with patents applicated in the same year, and the total amount is computed as the annual weight of the corresponding topic.

The growth rates of topic weight in the recent three years are averaged to reflect a more general development trend, which is defined as follows:

$$\text{Potential of topic } \mathrm{j} = \mathrm{TT}_j = \frac{TW_j^{2018} + TW_j^{2017} + TW_j^{2016}}{TW_j^{2015} + TW_j^{2014} + TW_j^{2013}}, j = (1, 2, \cdots K) \quad (4)$$

where TW_j^T stands for the distribution weight of the j^{th} topic in the year T. In our study, 2018 is set as the latest year for calculation since the DII data of 2019 is incomplete due to the collection time lag.

The two-dimensional evaluation system is developed based on TW and TT to divide the recognized topics into four quadrants. Core and potential technologies are identified with this system. Incorporated with TRM depicted in the following stage, it can help to predict promising evolving paths.

(4) Stage 4 – TRM construction and opportunities discovery

Based on the identified technological topics, the potential linkages between these topics are explored to build TRM and discover promising opportunities. SAO semantic analysis is used to extract evolving linkages between technological topics.

The words for each topic generated in stage 3 are closely related to specific technologies. In this stage, engaging domain experts' opinion, we first select key terms from the list for each technological topic and then use them to retrieve SAO structures through the tag function of TextBlob. Then, they are grouped in terms of the topics S and O belong to. After doing this, we obtain a series of simplified SAO structures for each pair of topics. Then, we explore the potential linkages between topics using SAO semantic analysis that pays much attention to the target topics and their 'actions', which means the relationships. Hence, the specific dynamics between topics are effectively traced.

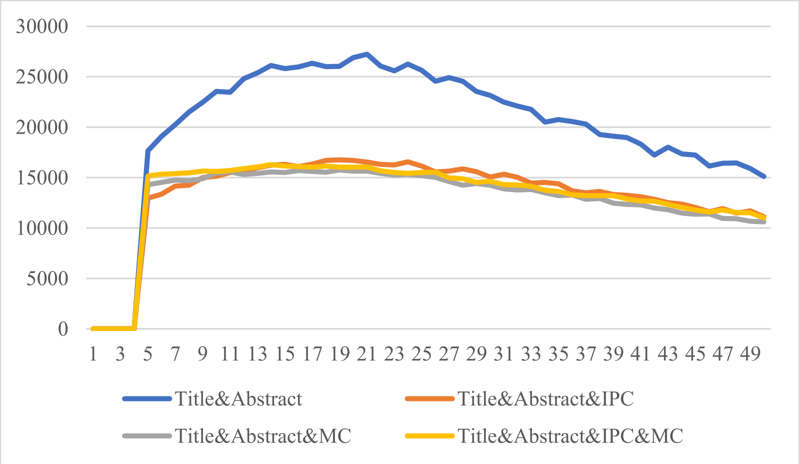


Fig. 4. The average perplexity scores of the four models in group A.

Table 4
Top 10 IPCs and MCs of the training sets in the DSSC case and the comparative case.

The DSSC Case (High coupling) – 827 Records				The Comparative Case (Low coupling) – 1444 Records			
IPCs	Records	MCs	Records	IPCs	Records	MCs	Records
H01M-14/00	**413**	**X15-A02D1**	**494**	H01G-09/20	107	T01-S03	162
H01L-31/04	**373**	**U12-A02A8**	**467**	G05D-01/02	86	X15-A02D1	153
H01G-09/20	**319**	**X16-A04**	**404**	B25J-11/00	70	T01-N01F	143
H01L-31/042	134	**L03-E05B1**	**365**	B25J-09/16	69	D05-H09	137
H01G-09/04	100	**L03-E05B**	**306**	B82Y-30/00	61	L03-E05B1	135
C07F-15/00	65	**X15-A02A**	**269**	B82Y-40/00	60	T01-E05Q	135
C09B-57/00	65	**U12-A02A**	**253**	H04L-29/08	58	T01-M06Q	111
C09B-57/10	56	**A12-E11B**	**248**	G05D-01/00	55	P62-E	108
H01L-31/02	55	U12-A02A3	174	G06K-09/00	54	T01-J07D1	107
H01L-51/46	51	L04-C11C	164	G06N-03/08	53	T01-J07D	106

Table 5
The comparison of Total Precision between cases with high coupling data and low coupling data (%).

	B-1#	B-2#	B-3#	B-4#
The DSSC Case (High coupling)	70.38	70.04	71.06	70.38
The Comparative Case (Low coupling)	70.01	78.57	82.76	**83.10**

Specifically, the SAO structures consisting of only the topics and related verbs are input for SAO semantic analysis. The relationships between topics are categorized into four types: *compose, improve, replace*, and *apply*, according to the experts' opinion. The *compose* refers to the relationship between technology and sub-technology, such as "solar cell contains counter electrode" or "counter electrode composes solar cell." The *improve* refers to one technology that modifies or processes another, such as "a sealing material seals electrolyte." The *replace* refers to a new technology that replaces another, such as "solid electrolyte replaces liquid electrolyte." The *apply* refers to a technology applied to a product or market. Then, a classification model, naïve Bayesian algorithm, is trained to identify the key relationships for each SAO group (the topic pairs). The model is trained on data randomly selected 10% of the SAO strings until the classification accuracy reaches 80%. The trained classification model is used to classify the remaining simplified SAO structures.

Based on the above analysis, the main type of semantic relationship between each pair of topics is determined and then sorted according to the number of their corresponding SAO structures. The significant relationships between topics are selected from the sorted list, incorporating expert judgment. Then, we plot the corresponding topics with their linkages on four layers of TRM, which are the layers of Material & Structure, Component Technology, Product, and Application & Market from the bottom to up (Zhang et al., 2016a). It depicts the evolving paths of the technologies. The core and potential technologies with their promising evolving paths are finally identified to support the exploration of potential opportunities.

4. Case study

4.1. Raw data retrieval & feature extraction

The proposed method is evaluated on DII patents related to dye-sensitized solar cell (DSSC) technology. As a third-generation solar cell solution, DSSC draws much attention from the industry due to its low cost, easy production, environment-friendly raw materials, and relatively high efficiency (Regan and Gratzel, 1991). The familiarities of the DSSC were developed through a series of tech mining analyses. Based on our previous research achievements (Huang et al., 2018; Ma et al., 2014), we constructed the search terms of DSSC, as shown in Table 2. Finally, we retrieved 9,883 DII patents related to DSSC from 1991 to 2019.

Based on the retrieved DII patents, we used the VantagePoint to extract terms from their titles and abstracts and then applied the term clumping process to remove unnecessary data. In total, 7,482 terms were finally obtained. Besides, we obtained 2,554 IPCs and 3,088 DMCs by removing the ones that only appear once.

4.2. Topic extraction model

By manually screening 2,581 DSSC patents in 2011 and 2012, we obtained 821 records associated with five sub-technologies of DSSC (Table 3) and extracted 392 and 1,520 terms from the title and abstract, respectively. Also, we obtained 177 IPCs and 481 DMCs extracted from the 821 records. We then generate eight Document-Feature Matrixes from the training set according to the proposed feature selection model and input them into the LDA model for analysis. Note that five technologies included in the training set are sub-technologies of DSSC whose coupling degree is high, which brings a big challenge for topic recognition (Yau et al., 2014).

There is a trade-off between comprehensiveness and accuracy: fewer topics make results easier to understand, but more topics lead to higher accuracy (Zhang et al., 2016b). The performance test with different numbers of topics ranged [5, 50], was conducted (Chen et al., 2017; Chen et al., 2019; Yau et al., 2014). Then, we used the average perplexity scores to evaluate the trained model to fit the dataset with a specific choice of K. Specifically, we set parameters $\alpha = 0.5$ and $\beta = 0.1$ for topic modeling and applied 2,000 iterations of Gibbs sampling to infer the needed distributions (Chen et al., 2017). Since the corresponding models in group A and group B share similar data structures, the average perplexities of the four models are only computed in group A. As shown in Fig. 4, the perplexity scores of all models present a consistent inverted U trend, and the lowest score is achieved when K=50, indicating the lowest misrepresentation of the words. Thus, the parameter K is determined to 50 to better capture the topics with easier interpretation.

The TPs of the eight models are computed after LDA topic modeling, as shown in Table 3. Through comparative analysis, we obtained some findings.

1) As shown in Table 3, all the models in group B provide higher TPs than the corresponding models in group A, indicating the significantly improved topic recognition accuracy with terms of the title given double weights. It confirms that the title of the DII patent is more valuable than the abstract when applied to the topic analysis.
2) The comparison of TP values (out of brackets) within the groups indicates that the performances of the #2 and #3 models are not meaningfully different from the corresponding #1 models. In particular, the original accuracy (out of brackets) of the #3 models has barely improved over the corresponding #1 models. This is not consistent with our expectations. Through further analysis, some common classification codes were found, which were shared by many patents in our training dataset because all the patents are related to a specific technology DSSC. This could lead to a problem

Table 6
DSSC Topics with top 5 related terms and IPCs/ DMCs.

No.	Subsystem	Technological Topics	Top Related Terms	Top IPCs/ DMCs
1	Markets & Applications	Building	building, window, house, door, wall, green house	X15-A05, X15-A02B, U12-A02A5, U12-A02A1, H01L-031/05
2		LED/Power supply	LED, controller, storage battery, electric power, power supply	X15-A08, X15-A02X, X16-B01, X26-E01E, X16-G02A
3		Vehicle	Vehicle, roof panel, motor vehicle, electric power, electric vehicle	L03-H05, X15-A05, X21-B04A, X21-A01F, G02F-001/15
4		Portable electronic device	mobile telephone, power supply, calculator, personal computer, mobile phone	W01-C01D3C, L03-H03, L03-H03A, G02-A04A, G02-A04B
5	Products	Organic DSSC & organic electronic device	organic dye-sensitized solar cell, organic thin-film solar cell, organic electronic device, organic integrated circuits, organic laser diodes	E05-R, U11-A15B, E24-A06B, C07D-405/14, E05-E01B
6		Solid state DSSC/ perovskite solar cell	solid state dye sensitized solar cell, solid state, perovskite solar cell, solid state solar cell, hole transport	X15-A02D1, L03-E05B1, U12-A02A8, A12-E11B, A12-W16
7		DSSC module	dye-sensitized solar cell module, solar cell module dye-sensitized solar cell unit, solar cell unit, photoelectric transducer module	U12-A02A5, X15-A05, X21-A01F, X22-F03, E04D-013/18
8		Flexible DSSC	flexible dye-sensitized solar cell, flexible solar cell, excellent flexibility, flexible photoelectrode, high flexibility	U12-A02A4B, U11-C01J8, A11-C04B2, A05-E04E, A11-C04B
9	Component technologies	Liquid electrolyte	liquid electrolyte, electrolyte composition, ionic liquid, electrolyte liquid, ionic liquid electrolyte	A12-M02, L03-E01C3 L03-E08C, H01G-009/022, X16-J01A
10		Gel electrolyte	gel electrolyte, polymer gel electrolyte, quasi-solid electrolyte, electrolyte gel polymer electrolyte	A12-E11B, H01G-009/20, L03-E05B1, A12-W16, L03-B03H
11		Solid electrolyte	solid electrolyte, solid polymer electrolyte, electrolyte, Liquid electrolyte, electrolyte composition	A12-M02, L03-B03H, X16-J02, H01G-009/022, L03-E01C3
12		Flexible/ conductive substrate	transparent conductive layer, flexible substrate, conductive substrate, conductive layer, transparent conductive film	U12-A02A4B, U11-C01J8, A11-C04B2, A05-E04E, A11-C04B
13		Metal-based electrode	Pt electrode, negative electrode, Platinum counter electrode, Platinum electrode, platinum layer	X16-K02, H01G-013/00 X15-A02D1, H01G-009/20, X16-A04
14		Carbon-based electrode	carbon nanotube, counter electrode, carbon atom, carbon counter electrode, carbon electrode	L03-H05, L03-A02G, L03-A02B, C01B-031/02, E05-U05C
15		Light anode	Light anode, anode, photoelectrode, electrode, light electrode	L03-B03H, L03-B03G, L03-B03G2, H01G-009/042, L03-B03G1
16		Porous metal oxide semiconductor	porous semiconductor, metal oxide, metal oxide layer, porous semiconductor layer, oxide semiconductor	X16-D, H01M-014/00, U11-C05F6, U12-A02A4A, L04-C11C
17		Organic dye	organic dye, thiophene, compound, electron acceptor, electron donor	C09B-057/00, E11-F03, E25-E01, E25-B03, E11-A01
18		metal complex dye	metal complex dye, ruthenium complex, ruthenium dye, dye compound, alkali metal	C07F-015/00, E05-M02A, C09B-057/10, E11-F11, U11-A01X
19		Phthalocyanine dye	Phthalocyanine, Porphyrin, phthalocyanine compound, porphyrin derivative, porphyrin compound	E23-B, C09B-047/00, C09B-047/04, E23-A02, C07D-487/22
20	Materials & Structure	Sealing material	sealing material, sealing portion, sealing agent, sealant, sealing element	U12-A02A4E, H01M-002/08, X16-F01A, H01L-031/048, U11-D01B
21		Hole-transport material	hole transport, hole transport layer, p-type semiconductor, hole-transport, P-type organic semiconductor	H01L-051/42, H01L-051/44, H01L-051/46, H01L-031/0256, L04-C06C
22		Conductive glass	conductive layer, conductive material, conductive glass, FTO, fluorine-doped tin oxide conductive glass	U14-H01E, U12-A02A8, U11-A08B, U11-C05B4, U11-A08B2
23		Polymer film	polymer film, PET. plastic substrate, polymer layer, conductive polymer	A05-J12, A09-A03, H01B-013/00, L03-A02E, A10-D06
24		Metal mesh/ foil		J04-E04C, B01J-035/

(*continued on next page*)

Table 6 (*continued*)

No.	Subsystem	Technological Topics	Top Related Terms	Top IPCs/ DMCs
			platinum, stainless steel, titanium, metal mesh, metallic foil	02, J04-E04, J04-E11, B01J-021/06
25		Metal material	Platinum, Gold, silver, nickel, copper	L04-C10E, U14-H01E1, L03-A02C1, A04-C02E, L04-C10J
26		Carbon material	carbon nanotube, graphene, carbon black, carbon material, graphene oxide	L04-A05, U11-A14, A12-W14, U11-C13, L03-A02G
27		TiO2	titanium oxide, TiO2, titanium dioxide, titanium, tin oxide	E35-K02, E35-H, E35-C02, E35, E31-U01
28		ZnO/SnO2	Zinc, zinc oxide, tin, tin oxide, ZnO	U12-A02A2A, L04-A03D, L04-C10F, U11-C01J3B, L04-C12A
29		Nano structure	nanoparticle(s), nanometer, nanocrystalline, nanowire, nanotubes, nanorods	L04-C16, H01L-031/18, L04-A05, U11-A14, H01L-031/0224
30	Properties	Low cost	low cost, cost-effective manner, simple manner, simple operation, inexpensive manner	X15-A, E11-W, X15-A01, X16-X, X16-E09
31		Durability	excellent durability, excellent stability, high durability, long-term stability, excellent thermal stability	U12-A02A2X, A10-D, L03-E05B1A, A05-J, X15-A02
32		Photoelectric conversion efficiency	high photoelectric conversion efficiency, excellent photoelectric conversion efficiency, excellent photoelectric transfer characteristics, photoelectric conversion efficiency, photoelectric conversion	L04-C18, U11-C05F6, L03-H02, L04-C17, U11-A09
33		Light absorption	electrical energy, light absorption, photoelectric conversion efficiency, light absorbing, light absorption layer	H01L-051/44, H01L-051/42, H01L-031/00, H01L-051/46, U12-A02A1

that these common classification codes may not help to improve the classification accuracy. To address this issue, we removed the common classification codes from the original set, such as IPC H01M-14/00, which is shared by 413 out of 827 patents (Table 4). Six new Document-Feature Matrixes for the #2, #3, and #4 models (#1 model remains unchanged) are generated for topic re-modeling. As shown in Table 3, the updated TP values (in brackets) of the #2, #3, and #4 models are improved over the original ones (out of brackets). The results indicate that the technical classification codes can improve topic recognition and classification accuracy under the removal of commonly shared codes.

Table 7
The importance and potential of the 33 technological topics of DSSC.

Quadrant	Technological Topics	Importance	Potential	Tag
I (High importance & High potential)	Building	247	124.52%	●
	Light anode	413	107.92%	
	LED	422	104.52%	
	Nano structure	431	94.06%	
	Organic DSSC	365	70.79%	
	Light absorption	206	66.27%	
	Solid electrolyte	209	61.17%	
	TiO2	371	60.70%	
	Organic dye	393	57.04%	
II (Low importance & High potential)	Carbon material	175	85.33%	
	Carbon-based electrode	197	78.85%	
	Low cost	195	74.46%	
	Hole-transport material	189	71.89%	
	Durability	179	69.65%	
	Polymer film	152	65.73%	
	Solid DSSC/ Perovskite solar cell	174	64.71%	
III (High importance & Low potential)	Porous metal oxide semiconductor	706	50.18%	⊙
	DSSC module	490	52.79%	
	Portable electronic device	319	42.99%	
	Flexible/conductive substrate	254	32.47%	
	Sealing material	241	47.80%	
	Photoelectric conversion efficiency	211	34.42%	
	Flexible DSSC	210	41.34%	
IV (Low importance & Low potential)	Gel electrolyte	186	56.86%	○
	Vehicle	177	55.47%	
	Metal electrode	122	54.72%	
	Metal material	168	50.28%	
	Porphyrin/ phthalocyanine	182	48.70%	
	Metal mesh/foil	138	40.56%	
	Liquid electrolyte	152	26.88%	
	Metal complex dye	191	33.07%	
	ZnO	177	32.93%	
	Conductive glass	116	28.12%	

3) TP values of the #2 and #3 models are compared, including the original and updated TPs. The #2 model shows higher accuracy than the #3 model in the original results, while lower accuracy in the updated results. The in-depth analysis on the IPCs and DMCs of the training set of DSSC explained these conflicting results. More common DMCs were found than common IPCs shared by the five sub-technologies of DSSC. Specifically, our training set only has three IPCs (H01M-14/00, H01L-31/04, H01G-09/20) but eight DMCs (X15-A02D1, U12-A02A8, X16-A04, L03-E05B1, L03-E05B, X15-A02A, U12-A02A, A12-E11B), which are shared by more than 30% patents (Table 4). With the removal of these common classification codes (the bold ones in Table 3), the reversed comparison results were obtained. Also, inspired by the possibility that DMC is a better feature selection for topics identification than IPC because it is more clear, detailed, and standardized (Ma and Porter, 2015), we conducted another test. We set up another training set including patents related to DSSC, gold nanoparticle, self-driving vehicle, quantum computer, artificial intelligence, and the internet of things, as shown in Table 5, to conduct a comparative analysis. Since the six technologies included in the new training set belong to different technical categories, their patents are less coupled and share fewer

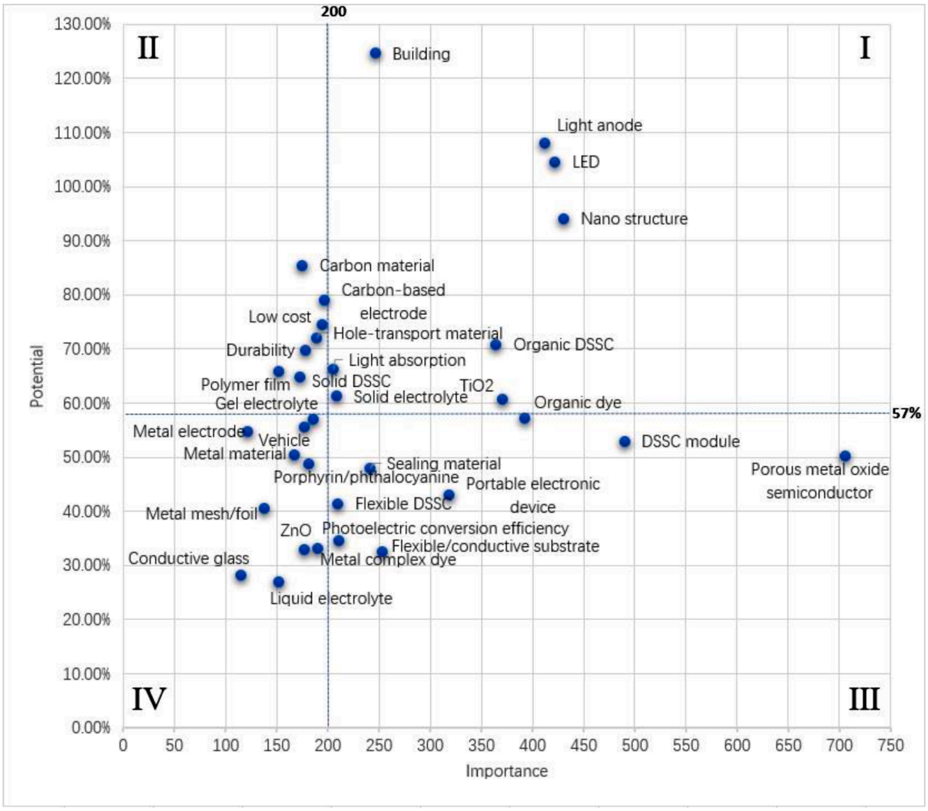


Fig. 5. The scatter diagram of the 33 technological topics of DSSC.

classification codes[1]. The comparison between the #2 and #3 models in the new data (Table 5) is consistent with the comparison of the updated TPs in the DSSC case, confirming that DMC (contained in B-3# model) has a better performance than IPC (contained in B-#2 model) on topic recognition and classification.

4) Further comparison of the two cases shown in Table 5 shows that the accuracy of the models that contain IPCs or/and DMCs (the B-#2, B-#3, and B-#4 models) in the new case is improved more significantly than that in the DSSC case. For instance, in the DSSC case, the accuracy of the model B-#3 was improved by 0.68% compared to the model B-#1, while in the new case, the accuracy improvement was up to 12.75%. To comprehend this difference, we further analyze the characteristics of the datasets in the two cases in terms of the IPCs and DMCs of each topic. As a result, it was found that the new case dataset is low-coupled. Its topics share fewer technical classification codes because they belong to different technical categories. In contrast, the topics in the DSSC dataset are highly coupled. This indicates that technical classification codes have better topic recognition performance when less coupled with each other, i.e., fewer shared classification codes.
5) The comparison of model #4 and the corresponding models #2 and #3 gives that the models involving both features of IPC and DMC perform better than only involving one of them. It indicates that DII patent data is high-quality ST&I data to identify technological topics since it provides a specific feature – DMC, providing clearer and more detailed technical classifications.

The experimental analyses show that the B-#4 model, involving double-weighted title terms, abstract terms, IPCs, and MCs, performs best on topic clustering. Accordingly, the B-#4 model is finally selected for the patent topic recognition model used in the next step.

4.3. Core and potential technologies identification for DSSC

We employed the best model (B-#4 model) to extract technological topics from the set of 9,883 DSSC patents that we retrieved at stage 1. Also, we set the number of topics to 50 according to the average perplexity. The following topics were removed with the engagement of experts: 3 topics that have little to do with DSSC - lithium battery, silicon solar cell, and fluorescence dye, 2 general topics - electrode and electrolyte, and 2 hard-to-define topics. Accordingly, 43 topics were obtained. Then, we combined 19 topics into 9 technological topics to make all the topics are within the appropriate technological scope and subdivision level. Finally, we obtained 33 technological topics. Determining the subsystem of DSSC is an essential part of exploring the TRM. Thus, we divided subsystems based on expert judgment and then classify the 33 technological topics into the 5 subsystems: Markets & Applications, Products, Component technologies, Materials & Structures, and Properties of DSSC. Table 6 lists the 33 DSSC topics with the top 5 related terms and IPCs/DMCs. We identify 4 markets & applications, 4 DSSC products, 11 component technologies, 10 DSSC materials & structures, and 4 properties of DSSC.

The core and potential technological topics were identified from the 33 technological topics of DSSC. The TW and TT values were computed for the 33 technological topics from two aspects of importance and potential, as shown in Table 7. Then, the two-dimensional evaluation system was established with the TW and TT values, where 33 technological topics were divided into four quadrants, as shown in Fig. 5. Through this process, we identified nine core and potential technological topics (in Quadrants I of Table 3 and Fig. 5). Among these topics, 'Building' and 'LED' are the most important and potential applications of DSSC. The DSSC module can adhere to an indoor/outside wall, installed in the window or greenhouse to absorb the lights and convert it

[1] From Table 6, we can see that all the top IPCs or DMCs are shared by lower percent of patents in the comparative case, indicating that few common IPCs or DMCs exist and the dataset is low coupled.

Table 8
Important semantic relationships with examples between topics.

No.	Topic1	Topic 2	Example of SAOs	Relationship
1	Gel electrolyte	Liquid electrolyte	CN 1624837 **quasi solid electrolyte**is set to replace the **liquid electrolyte**	Replace
2	Gel electrolyte	hole transport material	WO 2012011642 **quasi-solid polymer electrolyte**comprises a **hole transporting material**	Compose
3	Solid electrolyte	Liquid electrolyte	CN 1645632 **solid electrolyte**is assembled to replace the **liquid electrolyte**	Replace
4	Solid electrolyte	hole transport material	KR 2019035572 **solid electrolyte**containing mixture of **p-type organic semiconductor**	Compose
5	Sealing material	Liquid/gel/solid electrolyte	JP 2014165061 **sealing material**seals the **electrolyte solution**	Improve
6	Solid state DSSC/ perovskite solar cell	Gel electrolyte	US 2006174936 **quasi-solid state**is obtained using the water based **electrolyte gel**	Apply
7	Solid state DSSC/ perovskite solar cell	Solid electrolyte	KR 2019035572 **solid state dye-sensitized solar cell**comprises **solid electrolyte**	Apply
8	Flexible DSSC	Polymer film	CN 101728082 **flexible dye-sensitive solar battery**comprises electrically-conductive **polymer layer**	Apply
9	Flexible DSSC	Flexible/ conductive substrate	TW 201409724 **Flexible dye-sensitized solar cell**comprises first **flexible substrate**	Apply
10	Flexible/conductive substrate	Polymer film	EP 1605479 **flexible conducting substrate**composed of a flexible **transparent polymer**	Compose
11	Flexible/ conductive substrate	Metal mesh/foil	EP 1605479 **flexible conducting substrate**composed of a **metal**	Compose
12	Metal mesh/foil	Conductive glass	TW 424609 **metal mesh**to replace transparent **conductive glass**	Replace
13	Metal mesh/foil	Metal material (Pt,Au,Ni,Stainless)	JP 2006286534 **metal mesh**are made of **platinum**	Compose
14	Metal -based			
electrode	Metal material	JP 2006324111 **counter electrode**comprises a **metallic foil**	Compose	
15	Carbon-based electrode	Carbon material	US 2013255763 **Carbon electrode**comprises a **graphene**	Compose
16	Carbon-based electrode	Metal -based electrode	CN 102290251 **graphene film**canreplace **platinum counter electrode**	Replace
17	Carbon material	TiO2	JP 2013118127 **carbon nanotubes**are coated with the **titanium oxide**	Improve
18	Carbon material	ZnO/SnO2	CN 102779650 **Carbon particles**are provided with **zinc oxide**	Improve
19	Nano structure	TiO2	WO 2012016160 **nanostructured film**is made of **titanium oxide**	Compose
20	Nano structure	ZnO/SnO2	CN 102891191**nano-wire**is made of **zinc oxide**	Compose
21	Porous metal oxide semiconductor	Nano structure	CN 101692411 **nano-particle film** attached to **porous semiconductor electrode**	Improve
22	Porous metal oxide semiconductor	TiO2	WO 2014181792 **semiconductor layer**is made of **titanium oxide**	Compose
23	Porous metal oxide semiconductor	ZnO/SnO2	JP 2008150694 **porous semiconductor layer**made of **zinc oxide**	Compose
24	Light anode	Porous metal oxide semiconductor	JP 2014053263 **photoelectrode**equipped with a **semiconductor layer**	Compose
25	Light anode	DSSC module	JP 2011204789 **photoelectrode**used for **dye-sensitized solar cell**	Apply
26	Organic dye	Organic DSSC	S ss CN 102838881 **Organic dye**for preparing **organic dye-sensitized solar cell**	Apply
27	DSSC products	portable electronic device	CN 103237102 **Dye-sensitized solar cell**used for **mobile phone**	Apply
28	DSSC products	vehicle	WO 2011053068 **Dye-sensitized solar cell**used for **vehicles**	Apply
29	DSSC products	building	KR 2013006088 **Dye-sensitized solar cell module**adhered to **indoor wall**	Apply
30	DSSC products	LED	KR 1290858 **dye sensitive solar cell module**is connected to an **LED module**	Apply

*We only list some important and meaningful semantic relationships that are shown in the following DSSC TRM, in order to make the TRM look clear.

into electric energy so as to assist the construction of an energy-saving building (Reale et al., 2014). Also, through connecting with LED, the DSSC can supply power to LED lamps at night time using converted and stored electricity during daytime, ensuring higher efficiency of LED. The efficient LED is broadly applied to road lighting at night. We also found that ’organic DSSC’ has more significance and potential compared to the others, and its component technology of ’organic dye’ draws increasing attention. Moreover, ’light anode’ and ’solid electrolyte’ are the other two core and potential component technologies of DSSC, while the material ’TiO2′ and basic structure ’Nano structure’ are considered as the main solutions to manufacture semiconductor of DSSC. Another seven sub-technology fields were identified with relatively high potential values shown in Quadrant II, reflecting the future development opportunities of DSSC to a certain degree.

4.4. *TRM and opportunity discovery of DSSC*

The relationships among DSSC topics are explored in this methodology step. The SAO structures between topics are extracted, and then the deep semantic analysis is conducted to evaluate the potential relationships between topics. We first obtained 302 groups of SAOs which contain two topics, a total of 28,815 SAO strings. Then 2,882 (10%) SAO strings were randomly selected as the training dataset for the semantic relationship classifier. The accuracies of four classifiers, the naïve Bayesian (82%), decision tree (77%), maximum entropy algorithm (56%), and Logistic regression (80%), were compared. The highest accuracy classifier, the naïve Bayesian model, was selected to determine the relationships between these groups and classify them as *compose, improve, replace*, or *apply*. Synthesizing the expert’s opinion, we select the core evolutionary relationships among the 33 topics from a ranked list. Table 8 summarizes several important semantic relationships with examples.

According to the listed relationships of topics (Table 7), we build the DSSC TRM and tag the topics with *high importance, potential*, or both (Fig. 6).

With careful consideration of the analysis results and experts’

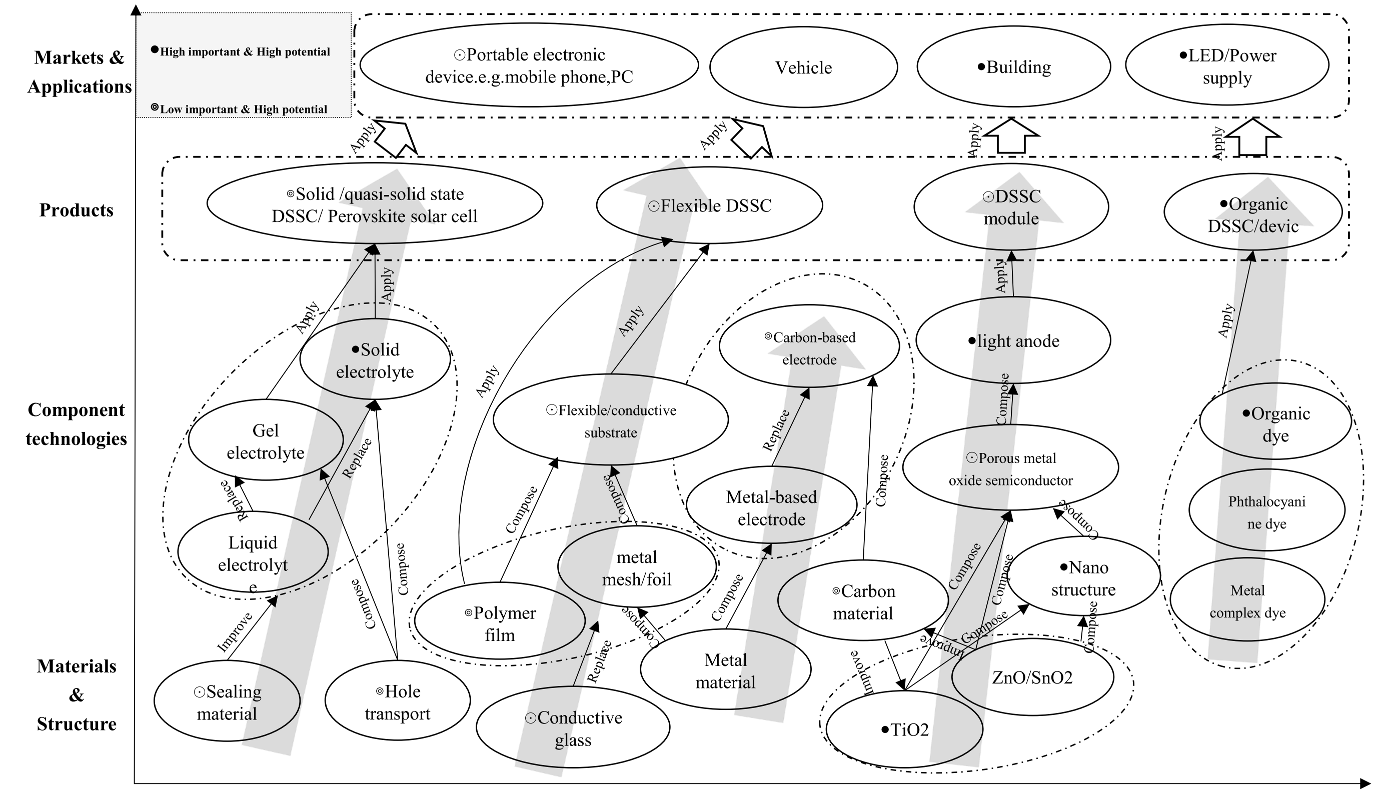


Fig. 6. The technology roadmap of DSSC.

opinion, five core technology evolving paths were determined, which correspond to its five sub-technology fields: electrolyte, substrate, counter electrode, light anode, and sensitive dye. Further, the potential opportunities are predicted by exploring the dominant or weak technologies along these paths. Our findings are summarized as follows:

(1) TiO2 is the most promising material for producing porous metal oxide semiconductors from the perspective of basic material and structure, which is a core component of the light anode. TiO2 has the advantages of being relatively cheap, abundant, and non-toxic (Muhammad Norhaffis et al., 2019). Compared to other materials, such as ZnO and SnO2, it demonstrates good photovoltaic performance (Jose et al., 2009). The TRM also demonstrates two promising development directions of TiO2: i) manufacturing TiO2 semiconductors with high surface area by modifying nano-structure of TiO2, and ii) producing carbon materials doped TiO2 that has better electrical conductivity and a higher degree of porous (Kang et al., 2007). The carbon materials, such as carbon, graphene, carbon black, and graphite, are also essential materials to produce carbon-based counter electrode, which has the advantage of low cost compared to the metal-based counter electrode (e.g., Pt counter electrode) (Gao et al., 2018). It could be another promising development direction of carbon material. In addition, hole transport material and polymer are promising materials for manufacturing gel or solid electrolyte and flexible substrate to improve the flexibility and solid-state stability of DSSC (Chen et al., 2010; Gao et al., 2018).

(2) The solid electrolyte, TiO2-based light anode, and organic dye are the dominant technologies in their corresponding sub-technology fields of DSSC from the perspective of component technology. The solid electrolyte has the highest safety and stability compared to liquid and gel electrolyte, which are essential for the industrialization and commercialization of DSSC (Cao et al., 2007). The light anode technique is the most crucial determinant of photoelectric conversion efficiency of DSSC, and its evolving path indicates that the TiO2-based light anode has the most promising potential. Besides, organic dye, as the core of developing organic DSSC, remains an important and hot topic.

(3) The organic DSSC is a hot and potential DSSC product from the perspective of the product, and its core sub-technology, organic dye, has drawn much attention due to high molar extinction coefficient, low cost, easy modification of molecular and convenient synthesis (Krishnan and Senthilkumar, 2018). Despite lower efficiency than a metal complex dye in photoelectric conversion, its growth trend shows its potential. Thus, discovering natural dyes with low cost but good performance and synthesizing new dyes by modifying molecular structure are both potential opportunities to advance the efficiency of organic DSSC (Krishnan and Senthilkumar, 2018). In addition, solid-state DSSC is another promising product of DSSC. Although it has received less attention than organic DSSC and flexible DSSC, it has attracted increasing attention in recent years. Its evolving path of ’hole transport material – solid electrolyte – solid-state DSSC/perovskite solar cell’ (Fig. 6) indicates the promising potential of the hole transport electrolyte in advancing solid-state DSSC and wide applicability to produce perovskite solar cell (Naoyuki et al., 2018), which is a new and popular product of solar cell in recent years.

(4) Building and LED are the most concerned and promising application fields from the perspective of market or application. DSSC has the advantages of colorful, flexible, and transparent, making it well integrated into architectural design. It can provide energy for building without affecting beauty (Reale et al., 2014). Besides, DSSC can generate electricity by indoor light as well as sunlight, which is a good option for the auxiliary power supply of electronic products, especially for LED lighting devices. Further, it can be applied to portable electronic devices, such as mobile phones, computers, and handheld calculators.

5. Conclusion

This paper proposes a hybrid methodology based on topic modeling, semantic SAO analysis, machine learning, and expert judgments to extract meaningful technical intelligence from DII patents and identify technological topics and potential opportunities. The proposed method is validated on the case analysis for DSSC in terms of its potential to contribute technical intelligence for R&D management.

The main contributions of this study are two-fold. First, we obtain an optimized topic extraction model with high accuracy, considering both the patent classification code and term location. The validation results answer the two questions initially raised: 1) higher weight in terms of title improves the accuracy of patent topic recognition, and 2) patent classification codes contribute to topic recognition when integrated with patent terms. Through further analysis, we find that the fewer classification codes shared between topics, the more pronounced the accuracy improvement of topic recognition. Therefore, removing the commonly shared codes by different topics increases the recognition accuracy for the highly-coupled technological topics. Further, DMC has a better performance contribution on topic recognition and classification than IPC because of clear and more detailed technical classifications of the DMC. Moreover, combining both IPC and DMC provides better accuracy than the sole use of one of them. These findings enrich our knowledge about patent data and provide effective ways to improve the topic recognition accuracy. On this basis, the optimized model is constructed and then applied to extract topics from DSSC patents in our empirical study.

Second, we propose a hybrid method that combines topic modeling, SAO technique, and machine learning to explore semantic relationships among technological topics. The proposed hybrid method advances the semantic analysis and topic relationship recognition, supplementing the shortcomings of the existing SAO analysis.

Our study still has some limitations that should be improved further. First, the DII patents are not real-time data due to a time lag. Therefore, the latest technical information could not be obtained through patent analysis to identify and predict technology. Second, only the influence of patent classification code and term position on topic recognition were analyzed, limiting the topic extraction model in patent data. Third, the weight ratio between the title and abstract was empirically determined to 2:1 without influence analysis of different weight ratios. Fourth, due to the complexity of SAO structures, it is challenging to identify semantic relationships from them even with the help of experts, resulting in the limited accuracy of the machine learning model. Further, the relationships between technologies are more complicated in practical applications, limiting the ability to in-depth explore technology evolution paths. Accordingly, future works will include the following four directions. 1) To explore valuable future-oriented or real-time data and integrate them into our analysis framework to improve the prediction ability. 2) To extend our research to other databases (e.g., SCI/SSCI, BkCI, Medline) and import more tech-related features (e.g., Tech focus of DII and the research field of SCI/SSCI) to our feature selection model for multiple data sources. 3) To analyze the influence of different weight ratios between the title and abstract on the recognition accuracy. 4) To propose a more detailed and systematic relationship classification for deep semantic analysis to enhance the exploration ability of the evolving paths of technology.

Author statement

None.

Declaration of Competing Interest

None.

Acknowledgments

This work was supported by the National Foundation for Young Scholars [grant number 71804016, 71704139]; the Science Foundation of Ministry of Education of China [grant numbers 18YJC870014]; the Beijing Municipal Commission of Education [grant number SM201910037005] and the National Science Foundation of Shaanxi Province Award [grant number 2019JQ-661].

References

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

Blei, D.M., Ng, A.Y., Jordan, M.I., Lafferty, J., 2003. Latent Dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.

Cao, H., Jia, H., 2013. Tibetan text classification based on the feature of position weight. International Conference on Asian Language Processing.

Cao, Y., Saygili, Y., Ummadisingu, A., Teuscher, J., Luo, J., Pellet, N., et al., 2007. 11% efficiency solid-state dye-sensitized solar cells with copper(ii/i) hole transport materials. Nat. Commun. 8 (8), 15390, 15390.

Chaomei, C., 2006. CiteSpace II: detecting and visualizing emerging trends and transient patterns in scientific literature. J. Am. Soc. Inf. Sci. Technol. https://doi.org/10.1002/asi.20317.

Chen, H., Guangquan, Z., Donghua, Z., Jie, L., 2017. Topic-based technological forecasting based on patent data: a case study of Australian patents from 2000 to 2014. Technol. Forecast. Soc. Change 119, 39–52.

Chen, H., Wang, X., Pan, S., Xiong, F., 2019. Identify topic relations in scientific literature using topic modeling. IEEE Trans. Eng. Manage.

Chen, H., Zhang, Y., Zhang, G., Lu, J., Zhu, D., 2015. Modeling technological topic changes in patent claims. 2015 Portland International Conference on Management of Engineering & Technology IEEE.

Chen, L., an, W., Zhang, J., Zhou, X., Zhang, X., Lin, Y., 2010. Fabrication of high performance pt counter electrodes on conductive plastic substrate for flexible dye-sensitized solar cells. Electrochim. Acta 55, 3721–3726.

Chiu, T.F., Hong, C.F., 2015. Recognizing and evaluating the technology opportunities via clustering method and google scholar. Intell. Inf. Database Syst., 159-169..

Erzurumlu, S.S., Pachamanova, D.A., 2020. Topic modeling and technology forecasting for assessing the commercial viability of healthcare innovations. Technol. Forecast. Soc. Change 156.

Gao, C., Wang, H., Han, Q., Hu, Z., Wu, M., 2018. High-efficiency magnetic carbon spheres counter electrode for dye-sensitized solar cell. Electrochim. Acta 264 (20), 312–318. https://doi.org/10.1016/j.electacta.2018.01.134.

Glenisson, P., Glanzel, W., Janssens, F., De Moor, B., 2005. Combining full text and bibliometric information in mapping scientific disciplines. Inf. Process. Manag. An Int. J. 41, 1548–1572.

Griffiths, T.L., Steyvers, M., 2004. Finding scientific topics. PNAS 101, 5228–5235.

Guo, Y., Ma, T., Porter, A.L., Huang, L., 2012. Text mining of information resources to inform forecasting innovation pathways. Technol. Anal. Strat. Manag. 24, 843–861.

Hofmann, T., 1999. Probabilistic latent semantic indexing. International Acm Sigir Conference on Research & Development in Information Retrieval ACM.

Huang, A., Lehavy, R., Zang, A., Zheng, R., 2018. Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach. Social Science Electronic Publishing.

Jeong, B., Yoon, J., Lee, J.M., 2019. Social media mining for product planning: a product opportunity mining approach based on topic modeling and sentiment analysis. Int. J. Inf. Manage. 48, 280–290.

Jing Ma, aa., Natalie, F.A.bB., Alan, L.P.cC.dD., Donghua Zhu, ee., Dorothy Farrell, bb., 2019. Identifying translational indicators and technology opportunities for nanomedical research using tech mining: the case of gold nanostructures. Technol. Forecast. Soc. Change 146, 767–775.

Jose, R., Thavasi, V., Ramakrishna, S., 2009. Metal oxides for dye-sensitized solar cells. J. Am. Ceram. Soc. 92, 289–301.

Kang, S.H., Kim, J.Y., Kim, Y.K., Sung, Y.E., 2007. Effects of the incorporation of carbon powder into nanostructured tio2 film for dye-sensitized solar cell. J. Photochem. Photobiol. A Chem. 182, 234–241.

Krishnan, S., Senthilkumar, K., 2018. Theoretical probe on modified organic dyes for high-performance dye-sensitized solar cell. Curr. Appl Phys. 18 (9), 1071–1079.

Kuhn, T.S., 1962. The structure of scientific revolutions. Phys. Today 16, 69.

Lee, C., Kang, B., Shin, J., 2015. Novelty-focused patent mapping for technology opportunity analysis. Technol. Forecast. Soc. Change 90, 355–365.

Lee, J., Kim, C., Shin, J., 2017. Technology opportunity discovery to R&D planning: key technological performance analysis. Technol. Forecast. Soc. Change 119, 53–63.

Lee, P.C., Su, H.N., 2011. Quantitative mapping of scientific research—the case of electrical conducting polymer nanocomposite. Technol. Forecast. Soc. Change 78, 132–151.

Li, Y., Ma, Y., 2008. The research of weight calculation method of text feature words based on latent semantic index. J. Comput. Appl. 6, 102–104. +108.

Lin, LiuL., Lin, TangT., Wen, DongD., Shaowen, YaoY., Wei, ZhouZ., 2016. An overview of topic modeling and its current applications in bioinformatics. SpringerPlus 5. Article number: 1608.

Ma, J., Porter, A.L., 2015. Analyzing patent topical information to identify technology pathways and potential opportunities. Scientometrics 102, 811–827.

Ma, T., Alan, L., Porter, Ying, G., Jud, R., Chen, X., Lidan, G., 2014. A technology opportunities analysis model: applied to dye-sensitised solar cells for China. Technol. Anal. Strat. Manag. 26, 87–104.

Ma, T., Zhang, Y., Huang, L., Shang, L., Kangrui, W., Yu, H., Zhu, D., 2017. Text mining to gain technical intelligence for acquired target selection: a case study for China's computer numerical control machine tools industry. Technol. Forecast. Soc. Change 116, 162–180.

Muhammad Norhaffis, M., Suhaidi, S., Mohd Haniff, W., Yusran, S., 2019. Preparation of TiO2 compact layer by heat treatment of electrospun TiO2 composite for dye-sensitized solar cells. Thin Solid Films 693, 137–699.

Naoyuki, S., Shota, F., Hidetaka, S., Hiroyuki, K., Seigo, I., 2018. Influence of transparent conductive oxide layer on the inverted perovskite solar cell using pedot: pss for hole transport layer. Mater. Res. Bull. 106, 433–438.

Olsson, O., 2005. Technological opportunity and growth. J. Econ. Growth 10, 35–57.

Pavlinek, M., Podgorelec, V., 2017. Text classification method based on Self-Training and LDA topic models. Expert Syst. Appl. 80, 83–93.

Phaal, R., Farrukh, C.J.P., Probert, D.R., 2004. Technology roadmapping–a planning framework for evolution and revolution. Technol. Forecast. Soc. Change 71, 5–26.

Porter, A.L., Detampel, M.J., 1995. Technology opportunities analysis. Technol. Forecast. Soc. Change 49, 237–255.

Reale, A., Cinà, L., Malatesta, A., DeMarco, R., Brown, T.M., DiCarlo, A., 2014. Estimation of energy production of dye-sensitized solar cell modules for building-integrated photovoltaic applications. Energy Technol. 2, 531–541.

Regan, B.O., Gratzel, M., 1991. A low cost, high efficiency solar cell based on dye sensitized colloidal TiO2 films. Nature 353, 737–740.

Shibata, N., Kajikawa, Y., Takeda, Y.eE.aA., 2011. Detecting emerging research fronts in regenerative medicine by the citation network analysis of scientific publications. Technol. Forecast. Soc. Change 78, 274–282.

Teufel, S., Siddharthan, A., Batchelor, C., 2009. Towards discipline-independent argumentative zoning.

Waltman, L., Eck, N.J.V., Noyons, E.C.M., 2010. A unified approach to mapping and clustering of bibliometric networks. J. Inf. 4, 629–635.

Wang, X., Ma, P., Huang, Y., Guo, J., Zhu, D., Porter, A.L., Wang, Z., 2017. Combining SAO semantic analysis and morphology analysis to identify technology opportunities. Scientometrics 111, 3–24.

Xin, L., Wang, J., Huang, L., Jiang, L., Jian, L., 2008. Empirical research on the technology opportunities analysis based on morphology analysis and conjoint analysis. Foresight 12, 66–76.

Xing, W., Croft, W.B., 2007. Investigating retrieval performance with manually-built topic models.

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

Yoon, B., Magee, C.L., 2018. Exploring technology opportunities by visualizing patent information based on generative topographic mapping and link prediction. Technol. Forecast. Soc. Change 132, 105–117.

Yoon, B., Park, Y., 2005. A systematic approach for identifying technology opportunities: keyword-based morphology analysis. Technol. Forecast. Soc. Change 72, 145–160.

Yoon, B., Park, Y., 2007. Development of new technology forecasting algorithm: hybrid approach for morphology analysis and conjoint analysis of patent information. IEEE Trans. Eng. Manage. 54, 588–599.

Zhang, Y., Chen, H., Lu, J., Zhang, G., 2017. Detecting and predicting the topic change of Knowledge-based Systems: a topic-based bibliometric analysis from 1991 to 2016. Knowl.-Based Syst. 133 (1), 255–268.

Zhang, Y., Robinson, D.K.R., Porter, A.L., Zhu, D., Zhang, G., Lu, J., 2016a. Technology roadmapping for competitive technical intelligence. Technol. Forecast. Soc. Change 110, 175–186.

Zhang, Y., Zhang, G., Chen, H., Porter, A.L., Zhu, D., Lu, J., 2016b. Topic analysis and forecasting for science, technology and innovation: methodology with a case study focusing on big data research. Technol. Forecast. Soc. Change 105 (C), 179–191.

Zhang, Y., Zhou, X., Porter, A.L., Gomila, J.M.V., Yan, A., 2014a. Triple Helix innovation in China's dye-sensitized solar cell industry: hybrid methods with semantic TRIZ and technology roadmapping. Scientometrics 99, 55–75.

Zhang, Y., Zhou, X., Porter, A.L., Vicente Gomila, J.M., 2014b. How to combine term clumping and technology roadmapping for newly emerging science & technology competitive intelligence: "problem & solution" pattern based semantic TRIZ tool and case study. Scientometrics 101, 1375–1389.

Zhou, X., Guo, Y., et. al., 2020. Identifying and assessing innovation pathways for emerging technology: hybrid approach based on text mining and altmetrics. IEEE-Trans. Eng. Manag.

Zhou, X., Huang, L., Porter, A., Vicentegomila, J.M., Phillips, F., 2019. Tracing the system transformations and innovation pathways of an emerging technology: solid lipid nanoparticles. Technol. Forecast. Soc. Change 146, 785–794.

Zhou, X., Porter, A., Robinson, D.K.R., Guo, Y., 2013. Analyzing research publication patterns to gauge future innovation pathways for nano-enabled drug delivery. Technology management in the It-driven services. IEEE. Technology Management in the IT-Driven Services (PICMET), 2013 Proceedings of PICMET '13. IEEE..

Zhu, D., Porter, A.L., 2002. Automated extraction and visualization of information for technological intelligence and forecasting. Technol. Forecast. Soc. Change 69, 495–506.

Tingting Ma, Ph.D, associate professor of School of Logistics, Beijing Wuzi University. She takes charge of the projects sponsored by the Chinese National Science Foundation (Award #71804016) and the Science Foundation of Ministry of Education of China(Award #18YJC870014). Her research interests focus on technological opportunity analysis and Tech-driven M & A. Email: matingtingmay@163.com

Xiao Zhou, Ph.D, associate professor of School of Economics and Management, Xidian University. She takes charge of Chinese National Science Foundation (Award #71704139) and the National Science Foundation of Shaanxi Province Award (Award #2019JQ-661).. Her research interests focus on technology innovation pathway and Tech-driven M & A. Email: belinda1214@126.com

Jia Liu, received the B.S. degree in Information management and information system from Bei Hang University, Beijing, China in 2006, and the Ph.D. degree in Management Science and Engineering from Beijing Institute of Technology, Beijing, China in 2011. Since 2019, she has been in School of Economics & Management, Communication University of China, her current research focuses on knowledge discovery and data mining. Email: liujia3891@163.com

Zhenkai Lou, received the Ph.D degree in management from Beijing Institute of Technoloy. His current research includes innovation management and supply chain management. In recent two years, he has published seven papers indexed by sci. Email: louzk@ahut.edu.cn

Zhaoting Hua, received the B.S. degree from Weifang University, Shandong, China. Since 2018, she has been a master student studying in School of Logistics, Beijing Wuzi University. Her current research focuses on technological forecasting and text mining.

Ruitao Wang, received the B.S. degree from Linyi University, Shandong, China. Since 2018, he has been a master student studying in School of Logistics, Beijing Wuzi University. His current research focuses on web mining and topic analysis.