Discovery of topic evolution path and semantic relationship based on patent entity representation

Jinzhu Zhang, Yue Liu, Linqi Jiang and Jialu Shi
Department of Information Management, School of Economics and Management, Nanjing University of Science and Technology, Nanjing, China

Received 13 March 2022
Revised 30 May 2022
25 July 2022
Accepted 25 July 2022

Abstract

Purpose – This paper aims to propose a method for better discovering topic evolution path and semantic relationship from the perspective of patent entity extraction and semantic representation. On the one hand, this paper identifies entities that have the same semantics but different expressions for accurate topic evolution path discovery. On the other hand, this paper reveals semantic relationships of topic evolution for better understanding what leads to topic evolution.

Design/methodology/approach – Firstly, a Bi-LSTM-CRF (bidirectional long short-term memory with conditional random field) model is designed for patent entity extraction and a representation learning method is constructed for patent entity representation. Secondly, a method based on knowledge outflow and inflow is proposed for discovering topic evolution path, by identifying and computing semantic common entities among topics. Finally, multiple semantic relationships among patent entities are pre-designed according to a specific domain, and then the semantic relationship among topics is identified through the proportion of different types of semantic relationships belonging to each topic.

Findings – In the field of UAV (unmanned aerial vehicle), this method identifies semantic common entities which have the same semantics but different expressions. In addition, this method better discovers topic evolution paths by comparison with a traditional method. Finally, this method identifies different semantic relationships among topics, which gives a detailed description for understanding and interpretation of topic evolution. These results prove that the proposed method is effective and useful. Simultaneously, this method is a preliminary study and still needs to be further investigated on other datasets using multiple emerging deep learning methods.

Originality/value – This work provides a new perspective for topic evolution analysis by considering semantic representation of patent entities. The authors design a method for discovering topic evolution paths by considering knowledge flow computed by semantic common entities, which can be easily extended to other patent mining-related tasks. This work is the first attempt to reveal semantic relationships among topics for a precise and detailed description of topic evolution.

Keywords Topic evolution path, Semantic relationship, Patent entity extraction, Patent entity representation, Semantic common entity, Knowledge flow

Paper type Research paper

1. Introduction

Topic evolution path describes the evolving process of a topic in a technical field, including development, division, integration, extinction and emerging, which can be used for quick identification of research hotspots, trends and gaps. The result can help researchers understand the history and current situation of the research field, which is essential to scientific and technological innovation (Liu *et al.*, 2020). Based on the construction of topic evolution path, the study of semantic relationship of topic evolution uses structured information contained in the patent text to explore a detailed and comprehensive understanding of relationships among topics. The result can help researchers better understand what leads to topic evolution (Wu *et al.*, 2019).

This work is supported by the National Natural Science Foundation of China (Grant No. 71974095) and the Postgraduate Research and Practice Innovation Program of Jiangsu Province (Grant No. SJCX22_0152). The authors would like to send their sincere appreciation to the anonymous referees for their valuable comments and suggestions.

Aslib Journal of Information Management

2050-3806
DOI 10.1108/AJIM-03-2022-0124

In the discovery of topic evolution path, current studies could be classified into two categories, including patent citation-based and patent text-based methods (Yu *et al.*, 2020). In the first class, researchers often use citation relationships to construct the coupling network, citation network and co-citation network (Chang *et al.*, 2010), then apply multiple network clustering methods to identify topics and compare them within different time windows (Érdi *et al.*, 2013). Thus, the topic evolution path could be constructed via similarity among different topics in different time windows. However, this class of methods cannot discover topic evolution path in time because citations need a long time for accumulation. Moreover, these methods ignore rich text information of a patent in the process of topic evolution path discovery. Patent text-based methods could address this problem from the perspective of patent text mining. Most studies extract keywords from patent text and construct co-word networks for subsequent topic detection and evolution analysis (Kim *et al.*, 2016). In this process, the topic evolution path is discovered through comparisons of common keywords among topics in different time series. However, these common keywords cannot cover the pairs of words that have different expressions but the same meaning in a specific technical field. In addition, a single keyword cannot express the technical meaning effectively because the patent-text semantics is professional and complex.

In semantic relationship discovery of topic evolution, there are few direct studies about this theme, but the theories and methods of semantic relationship discovery among entities could be borrowed and applied. This process can be formulated as a multi-class classification problem, in which each class corresponds to a predefined relation type (Nasar *et al.*, 2021). In these methods, neural network-based entity relationship extraction has been proven to be the most effective method with the best performance. It not only takes into account both local and global context information but also realizes flexible and automatic feature designing (Bhatia *et al.*, 2019). This type of methods has been widely used in knowledge graph, intelligent question answering and neural machine reading comprehension (Chowdhury *et al.*, 2018). However, it is still not used and improved for semantic relationship discovery among topics, which could give more concrete relations among topics for a better understanding of what leads to the topic evolution. In addition, these relationships among entities could not directly be used in topic evolution relationship discovery. It is necessary to predefine clear relationships according to patent analysis in a specific domain.

In summary, this paper mainly addresses the following three problems.

(1) Current studies are more focused on words but less on entities/phrases which contain more semantic and meaningful information. Therefore, this paper intends to form a method from the perspective of patent entity, making the results more accurate and interpretable.

(2) Some patent entities may have the same semantics but different expressions, which has not been considered in topic evolution path discovery. Therefore, this paper designs a method of patent entity extraction and semantic representation to identify synonymous patent entities, making a more accurate discovery of topic evolution path.

(3) Semantic relationships, which could provide more concrete correlations among topics, are not well revealed. Therefore, this paper designs a method to identify semantic relationships among topics based on semantic relationships among entities, for better understanding the factors and reasons of topic evolution.

In summary, this paper proposes a method based on patent entity representation for discovery of topic evolution path and semantic relationship. Firstly, patent entities are automatically extracted by Bi-LSTM-CRF and semantically represented by a representation

learning method. Secondly, topic evolution path is discovered through semantic similarity computation among patent entities, which can find out pairs of entities with the same meaning but different expressions in a special context. Finally, according to the specific task, multiple semantic relationships among topics are clearly designed and automatically discovered through the proposed method.

2. Related works

2.1 Topic evolution path discovery

Topic evolution path in consecutive years is used to analyze the evolution of topics over time, providing up-to-date changes in research trends (Zhang *et al.*, 2017). Current studies can be divided into two categories, i.e. patent citation analysis-based and patent text mining-based methods (Qi *et al.*, 2018).

In terms of patent citation analysis for topic evolution path discovery, Hummon and Dereian (1989) first propose the idea of "main path", which refers to the evolution path formed by the critical connections in the citation network. Pilkington and Meredith (2009) conduct a cluster analysis method based on common citation relationships in a citation network and obtain topic evolution structure in the operation management field. Martinelli (2011) uses citation relationships between patents to connect various research topics and identifies topic evolution paths in the field of telecommunications switching. Lu and Liu (2014) use papers and patents to identify the main path in citation networks respectively and analyze the topic evolution in the field of intellectual property rights. Kim *et al.* (2022) extract multiple longest paths from the multidisciplinary academic field's citation network to trace knowledge transfer and analyze the emergence, authority and topic dynamics of the identified trajectory. Chen *et al.* (2022) propose a semantic MPA (main path analysis) model by leveraging semantic information in two steps of candidate path generation and main path selection, which is capable of discovering more knowledge flows from important sub-fields and improving the topical coherence of candidate paths.

With the rapid development of text processing technology and deep learning, topic evolution path discovery based on patent text analysis has received increasing attention. These methods often use relationships between keywords to construct the co-word network for topic evolution analysis. Coulter *et al.* (1998) cluster topics based on the same keywords in papers or patents and analyze their changes over time to reveal the trends via topic evolution path. Yoon *et al.* (2021) use an LDA (latent Dirichlet allocation) model to obtain topics in different time windows and obtain topic evolution path in ship integrated power system. Wang *et al.* (2021) propose a time-based LDA model to analyze the topic evolution trend from the perspectives of rising, falling and stable and use this model to discover research hotspots and development trends in the field of blockchain. Wu *et al.* (2014) propose a core technology topic identification and evolution analysis method for topic evolution path and core technology discovery using SAO (subject-action-object) structure extraction and network analysis.

In general, most of the related studies have used a variety of complex network analysis methods on papers and patents for topic evolution path discovery, including citation network analysis, co-word network analysis, bibliometric analysis, cluster analysis and visualization analysis. However, these methods are more focused on relationships among words but less on the entities/phrases which have more semantics and meanings. In addition, the topic evolution path discovery via comparisons of common keywords is not accurate enough, because some keywords with different expressions may have the same semantics.

2.2 Patent entity extraction and semantic representation

Patent entity extraction is a type of problem called NER (named entity recognition), and the approaches to this problem are broadly classified into rule-based, unsupervised learning and

feature-based supervised learning methods (Li *et al.*, 2020). At present, this problem is often addressed as a sequence labeling task by jointly using neural network methods and classification methods. Bi-LSTM-CRF is one of the representative state-of-the-art models, which can efficiently use both past and future input features thanks to a bidirectional LSTM component (Huang *et al.*, 2015). It can also use sentence-level tag information thanks to a CRF layer for classification. Chen *et al.* (2020) propose a novel patent information extraction framework, in which two deep learning models, Bi-LSTM-CRF and BiGRU-HAN, are respectively used for entity identification and semantic relation extraction. An *et al.* (2021) apply this framework to extract sequence structures mentioned in patent documents. Wu *et al.* (2019) present a joint model for Chinese clinic named entity recognition, which combines Bi-LSTM-CRF model with self-attention mechanism. In this model, self-attention mechanism can learn long-range dependencies by establishing direct connections between each character.

After patent entities are extracted, they can be represented by multiple representation learning methods. Essentially, patent entity representation is very close to word embedding which denotes each word as a multi-dimensional vector. If we treat an entity as a special word, the word embedding method could be directly applied for entity representation. Word2vec proposed by Mikolov (Mikolov *et al.*, 2011) is the first word-based text representation learning method. This method trains the corpus through a neural network model using context information of each word and maps each word to a fixed-dimensional semantic vector. With the development of computing power, deep learning models, such as BERT (Devlin *et al.*, 2018), XLNet (Yang *et al.*, 2019) and RoBERTa (Liu *et al.*, 2019), are constantly improving the effect of word representation.

Based on the above research, the phrases, sentences and even chapters can be denoted as distributed representations with deep learning methods, which are more related to entity representation. Rao *et al.* (2015) extend a word representation learning model to form a phrase representation learning model that extracts phrases related to the identified topics. Zhang and Yu (2020) use word2vec to form word vectors for patent documents and use a linear algorithm to the semantic representations of patent phrases. Haghighian Roudsari *et al.* (2021) perform a pre-training model BERT on patent text for the subsequent classification task. Das *et al.* (2021) present a sentence embedding model trained on a natural language requirements dataset. Yoon *et al.* (2021) use doc2vec to perform representation learning on patent documents and realize technology hotspot prediction in the field of UAV technology.

These methods could provide a good reference for patent entity representation. However, the patent text contains many professional entities which are composed of multiple words or even phrases. Therefore, this paper will design a patent entity representation method by applying the word embedding method.

2.3 Semantic relationship of topic evolution

This paper uses patent entity relationships to recognize semantic relationships of topic evolution, so we will give a review on how entity relationships are designed and extracted.

Multiple types of entity relationships are predefined in different domains and fields. Uzuner *et al.* (2010) pioneer medical entity relation extraction in sexual research, which defines six categories of relationships among medical entities in detail, including present disease–treatment, possible disease–treatment, disease–test, disease–symptom, present symptom–treatment and possible symptom–treatment. In the field of biomedicine, Bachman *et al.* (2018) suggest the following three types of relationships, i.e. drugs and diseases, compounds and proteins and genic interaction. Chen *et al.* (2020) predefine 15 inter-entity relations in the field of thin-film head technology, including the spatial, part-of, operation, generating, in-manner-of, made-of, comparison, measurement, causative relation, formation

and purpose. These studies prove that the relationships among entities should be defined according to the specific domain, and different domains may have different types of relationships.

After the types of entity relations are defined, the relation extraction is formulated as a multi-class classification problem in which each class corresponds to a relation type. These approaches are broadly classified into two types: non-neural methods and neural network methods (Nasar *et al.*, 2021). In non-neural methods, a set of features is generated for each pair of entities and a classifier is then trained to classify any new relation instance. Kambhatla (2004) trains a MEM (maximum entropy model) classifier with 49 types of relations. Zhou *et al.* (2005) study an SVM (support vector machine) classifier incorporating diverse lexical, syntactic and semantic knowledge for relation extraction. Neural network methods could generate features automatically and get increasing attention. Wang *et al.* (2016) incorporate a multi-level attention model with CNN to capture attentions that are specific to relation extraction. Takase *et al.* (2016) employ RNN to extract relationships, considering the meaning of a relational pattern based on the semantic compositionality of constituent words. Liu *et al.* (2018) present an LSTM model on relation extraction which incorporates entity feature, entity position feature and part of speech feature. Zhang *et al.* (2019) first employ capsule networks for the task of relation extraction, using attention and capsule networks on the clustering layer.

The types of entity relationships and the methods of entity relationship extraction provide good references for semantic relationships discovery of topic evolution. However, the semantic relationships in a specific domain are not clear, and the entity relationship discovery methods cannot be directly applied to the semantic relationship discovery of topic evolution.

3. Data and method

The method for discovering topic evolution path and semantic relationship includes four parts, as shown in Figure 1. Firstly, patent entities are automatically extracted by the deep learning model Bi-LSTM-CRF. Secondly, semantic representation of patent entities is learned through a linear transformation of word vectors. Thirdly, topic evolution paths are detected through semantic comparisons among topics identified by a clustering algorithm. Finally, semantic relationships among topics can be determined through the proportion of different types of semantic relationships among entities belonging to each topic.

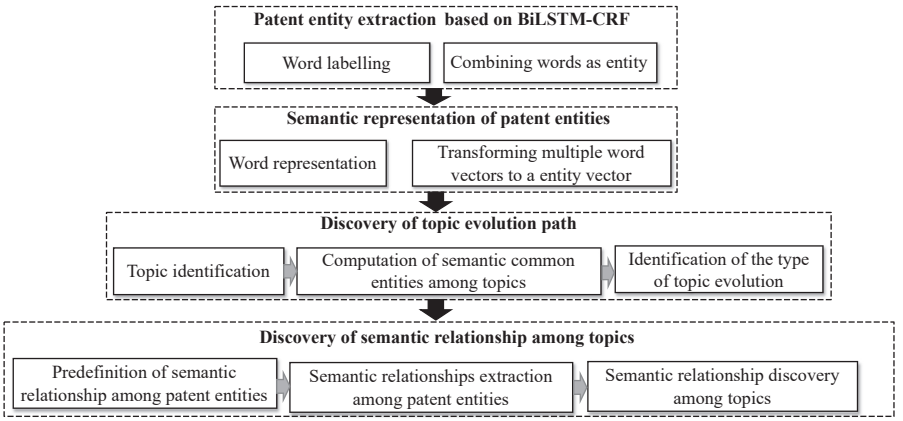

Figure 1.
Framework of the method

3.1 Data description

We collect the research dataset in March 2021 from DII (Derwent Innovations Index) in Thomson Reuters' Web of Knowledge. We gather patent data from 2008 to 2020. The search query for data collection is shown below.

$$\begin{aligned} IP &= B64* \text{ AND } TI \\ &= (((un-manned\ OR\ automatic\ OR\ autonomous\ OR\ remotely\ piloted\ OR\ nonhuman) \\ &\quad AND\ (aircraft\ OR\ \text{“aerial vehicle”}\ OR\ airship*\ OR\ drone\ OR\ plane\ OR\ aircraft \\ &\quad *\ OR\ airplane\ OR\ aerobat*\ OR\ aerostat*))\ OR\ UAV \end{aligned}$$

A total of 4,871 patents are retrieved. After the preprocessing, a total of 4,507 valid patents remain. According to the number of patent applications per year, as shown in Figure 2, it can be seen that a rapid growth stage happens between 2015 and 2017. Therefore, this paper takes the data from 2015 to 2017 in UAV area as the experiment data.

3.2 Patent entity extraction and representation

Due to diverse forms of patent entity expression, the effect of rule-based or feature-based information extraction methods is limited. Therefore, this paper uses the neural network-based method for patent extraction which is regarded as a sequence labeling task in information extraction. Firstly, a small part of patents is randomly selected from the overall dataset for training the model, in which the tags of words used for patent entity combination are manually labeled. Then the sequence labeling model Bi-LSTM-CRF is selected for training because of its good performance in various tasks (Ma and Hovy, 2016). Finally, the words are labeled by this model on the rest of the overall data and combined as the patent entities.

3.2.1 Choosing labels of words for patent entity formation. This paper treats a patent entity as a combination of words sequence. It is generally implemented by cross labeling method which has been often used for NER (named entity recognition), such as person name, place name and agency name. The common labeling sets include {B, I, O}, {B, I, O, S} and {B, I, O, E, S}, in which "B", "I", "O", "E" and "S" denote different positions within an entity and each of them is the abbreviation of word "Begin", "Inside", "Other", "End" and "Single", respectively (Wang and Gu, 2017).

This paper chooses the simplest and most commonly used {B, I, O} as the labels of words for patent entity formation, where "B" represents the beginning word of a patent entity (it can

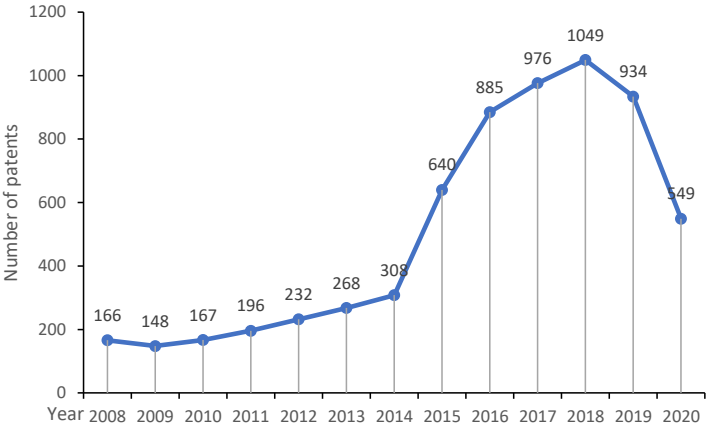

Figure 2.
Distribution of the number of patents applied per year

also denote the ending position of the prior entity), "I" represents the inside words of a patent entity, and "O" represents other words unrelated to the patent entity.

3.2.2 Bi-LSTM-CRF model training on a small dataset. This paper uses a small dataset randomly selected from the overall dataset as the training set, which is manually labeled with {B, I, O}. Then a Bi-LSTM-CRF model is trained on this dataset, which combines the advantages of Bi-LSTM and the CRF. On the one hand, it can consider the impact of long-distance context information on the annotation results. On the other hand, it can capture the global labeling sequence information and realize flexible feature design.

The framework of the model is shown in Figure 3. The first and last layers of the Bi-LSTM-CRF model are the input layer and the output layer, respectively. The input layer receives the segmented words sequence of patent text, and the output layer can predict a corresponding tag for each word in the sequence.

(1) Embedding layer

The embedding layer converts each word into a word vector and transmits it to the next layer. In many previous studies, when using pre-trained word vectors, the convergence speed is faster than using randomly initialized word vectors, and the accuracy and recall are relatively higher. Therefore, this paper selects word2vec, which performs well in multiple tasks, to obtain pre-trained word vectors for the next layer.

(2) Bi-LSTM encoder layer

The main idea of Bi-LSTM encoder layer is to propose a bidirectional training sequence, which propagates from forward and backward through the LSTM network. For the input sequence, it learns not only from left to right but also learns from right to left. Therefore, the complete context information of each word in the input sequence can be fully obtained.

(3) CRF layer

The CRF layer is a classification layer that assigns a tag to each word with its context information. The context information includes both the surrounding words and their corresponding tags. The CRF model uses this information to predict a global optimal tag for the current word.

3.2.3 Patent entity prediction on the overall dataset. After Bi-LSTM-CRF model is trained on a small dataset belonging to the overall dataset, the trained model can be used on the rest of the overall dataset to predict corresponding word labels. Since the word label sequence cannot directly be transformed into the patent entity, this paper establishes simple matching rules to combine words into patent entity.

Three matching rules are designed as follows. The first one is that the prediction result contains both labels "B" and "I". In this case, the words corresponding to label "B" are directly combined with the adjacent words labeled with label "I". The second case is that the prediction result only contains multiple "I", which means the label "B" is missing. In this case, the words corresponding to label "I" are directly combined at first and then judged manually. The third case is that the prediction result is empty or only contains one type of labels. In this case, the result is deleted directly.

For the convenience of understanding, an example of patent entity formation is shown in Figure 4. In this case, the words with label "B" and adjacent "I" are combined as a patent entity "automatic launch control system", and the words with label "O" are not considered because "O" represents other words unrelated to the entity.

3.2.4 Semantic representation of patent entities. Patent entities are mostly long phrases, and word2vec cannot be directly used to form corresponding patent entity vectors. Therefore, this paper firstly obtains the word vectors using the word embedding method word2vec (Mikolov *et al.*, 2011), then combines these vectors as the vector of a patent entity.

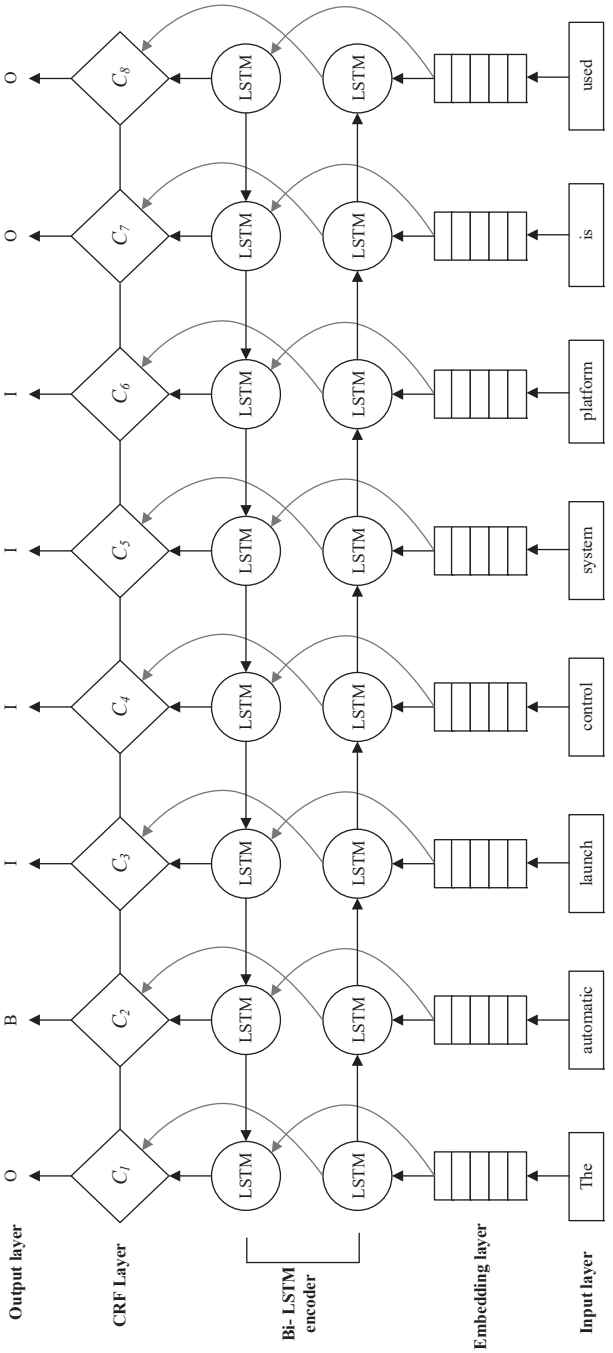

Figure 3.
Bi-LSTM-CRF neural network model for word labeling

Representation of words within a patent entity has been done in the embedding layer of Bi-LSTM-CRF model, so in this section we will introduce a method for combining word vectors into a patent entity vector. According to previous studies, this paper selects the average method for vector combination for its simplicity and good performance (Lei *et al.*, 2016).

The average method first accumulates the word vectors contained in each patent entity, then takes the average value as the vector of a patent entity. The calculation method is shown in Formula (1), where V_p represents the vector of a patent entity, n represents the number of words constituted in this entity, and K_i represents the vector of every word i contained in this entity.

$$V_p = \frac{\sum_{i=1}^{n} K_i}{n} \tag{1}$$

3.3 Discovery of topic evolution path

The method for discovering topic evolution paths includes three parts. Firstly, the topics in different time windows are identified by clustering patent entities. Then semantic common entities between topics are detected using semantic similarities. Finally, the type of topic evolution is identified through knowledge flow computed by semantic common entities.

3.3.1 Topic identification based on K-means. The clustering method is used to group a set of entities in a way that entities in the same cluster are more similar to each other than to entities in other clusters. The similarity is often denoted by the distance between two entities in a multiple-dimensional space, which can be calculated by cosine value of the angle between two vectors. Since the vector of each patent entity has been obtained previously, we just need to select a clustering method for topic identification in this part. Among many clustering algorithms, K-means is a classic method that performs well in many tasks. In addition, it is simple and fast to implement in terms of computational time and complexity and has been widely used in text analysis. Therefore, this paper selects K-Means to gather clusters, in which each cluster is denoted as a topic for subsequent evolution analysis.

Specifically, given a set of observations $(x_1, x_2, \ldots, x_n)$, where each observation is a d-dimensional real vector of a patent entity. K-means clustering aims to partition the n observations into k sets $S = \{S_1, S_2, \ldots, S_k\} (k \leq n)$ so as to minimize the sum of distances between the observations and their respective cluster centroid.

3.3.2 Semantic common entities among topics. Similarity among topics is computed by the number of semantic common entities whose expressions may be different but semantic meanings are almost the same. If two entities have a high semantic similarity, this paper treats them as a pair of semantic common entities.

First of all, a similarity matrix between two topics is formed. Specifically, T_i^t and T_j^{t+1} denote two topics in adjacent time windows t and $t+1$, where T_i^t contains x patent entities and T_j^{t+1} contains y patent entities. Thus an $x * y$ patent entity similarity matrix can be formed between these two topics, in which each element represents the similarity between two entities. This paper takes cosine value of the angle between two vectors as the measurement of similarity.

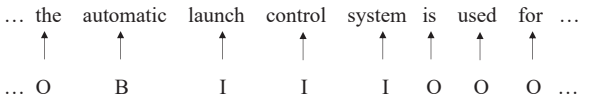

Figure 4.
An example of patent entity annotation and formation

Then, this paper treats a pair of entities as semantic common entities when the similarity reaches a certain threshold. The threshold does not have a fixed value because it is determined by experience and may vary according to different domains or areas. This paper recommends 0.7 or 0.8 as the experience value, but it can be changed with personal experience. The method of how to determine a threshold will be introduced in the experiment because it depends on the specific situation. After this, the number of semantic entities can be computed among topics which could be used for similarity computation among topics.

3.3.3 Identification of topic evolution path based on knowledge flow. If there are some semantic common entities between two topics, it is believed that these topics may have an evolution path. Besides, this paper further identifies which type of evolution has happened based on knowledge inflow and outflow. According to the previous studies, we define five types of topic evolution, including development, division, integration, extinction and emerging (Liu *et al.*, 2020; Song *et al.*, 2014). These types of topic evolution are then used for topic evolution path discovery.

This paper proposes knowledge outflow and inflow to judge the type of topic evolution by using semantic common entities. Knowledge outflow refers to the proportion of patent entities in the previous topic that outflows into the latter topic. The calculation method is shown in Formula (2), where K_{out} represents the knowledge outflow, T_i^t represents the ith topic in time window t, $S(T_i^t \cap T_j^{t+1})$ represents the number of semantic common entities between topic T_i^t and T_j^{t+1} and $N_{T_i^t}$ represents the number of patent entities included in topic T_i^t.

$$K_{out} = \frac{S\left(T_i^t \cap T_j^{t+1}\right)}{N_{T_i^t}} \tag{2}$$

Knowledge inflow refers to the proportion of patent entities in the latter topic that are inflowed from the previous topic. The calculation method is shown in Formula (3), where K_{in} represents the knowledge inflow, $S(T_i^t \cap T_j^{t+1})$ represents the number of semantic common entities between topic T_i^t and T_j^{t+1}, and $N_{T_j^{t+1}}$ represents the number of patent entities included in topic T_j^{t+1}.

$$K_{in} = \frac{S\left(T_i^t \cap T_j^{t+1}\right)}{N_{T_j^{t+1}}} \tag{3}$$

(1) Development evolution: This type of topic evolution means the topic T_i^t in time t develops into topic T_j^{t+1} in time $t+1$. Therefore, most knowledge of topic T_i^t outflows to topic T_j^{t+1} and most knowledge of topic T_j^{t+1} is derived from T_i^t. This paper recommends both knowledge outflow and inflow greater than 0.8 as the experience value for determining this type of evolution, but they may vary according to different domains.

(2) Division evolution: This type of topic evolution means the topic T_i^t in time t divides into two or more topics in time $t+1$. Therefore, the knowledge of topic T_i^t outflows to two or more topics. This paper recommends knowledge outflow greater than 0.2 as the experience value for determining this type of evolution, but it can be changed with personal experience.

(3) Integration evolution: This type of topic evolution is the opposite of division evolution. It means two or more topics in time t integrate into one topic in time $t+1$. Therefore, the knowledge of topic T_j^{t+1} comes from two or more topics in time t. This paper recommends knowledge inflow greater than 0.2 as the experience value.

(4) Extinction evolution: This type of topic evolution means the topic T_i^t in time t has few relationships with all topics in time $t+1$. In this case, the total knowledge outflow of a topic is smaller than a threshold, which is set to 0.3 as a recommended value.

(5) Emerging evolution: This type of topic evolution is the opposite of extinction evolution. It means topic T_j^{t+1} in time $t+1$ has few relationships with all topics in time window t. In this case, the total knowledge inflow of a topic is smaller than a threshold, which is set to 0.3 as a recommended value.

3.4 Semantic relationship discovery in topic evolution

The semantic relationship of topic evolution is discovered based on the relationships among entities. Firstly, the relationships among patent entities are predefined. Then the relationships among patent entities are extracted based on a neural network method. Finally, the semantic relationships of topic evolution are identified based on the semantic relationships among patent entities belonging to each topic.

3.4.1 Predefinition of semantic relationship. According to domain characteristics and patent dataset features, this paper predefines five types of relationships among patent entities. The detailed definitions are shown as follows in Table 1.

3.4.2 Semantic relationship extraction among patent entities. Firstly, a small part of relationships are manually labeled on a small dataset as the training set, which is the same as the dataset used in patent entity extraction. Then a neural network model is trained on this dataset. Finally, the relationships between each pair of entities on the entire dataset can be extracted through this model.

(1) Labeling the relationships among patent entities on a small dataset

Before manually labeling the relationship, data preprocessing is necessary. Firstly, word segmentation, lemmatization and case conversion have been done for preparation of the subsequent processing. In addition, duplicate records and records without abstract are removed. Finally, pronouns in abstract are replaced by the original subjects or objects based on regular expression matching.

Then the rapid annotation tool "brat" is used to label the relationship among patent entities (Stenetorp *et al.*, 2012). For example, the patent sentence "the drone is equipped with an ultrasonic reflector and a searchlight" is labeled as a mechanical relationship occurred between "drone" and "ultrasonic reflector" and "searchlight".

Table 1. Predefined relationships among patent entities

Semantic relationship	Detailed description
Mechanical relationship (M)	Refers to containment relationship, position relationship, etc. Among devices or mechanical parts
Efficacy relationship (E)	Refers to effect relationship of among devices or mechanical parts
Function-area relationship (FA)	Refers to the application field or area of some functions
Function-realization relationship (FR)	Refers to some devices or systems that realize certain functions
Control relationship (C)	Refers to a control relationship among devices or mechanical parts

(2) Extraction of the relationships among patent entities on the entire dataset

The above manually labeled dataset was used for model training. This paper chooses open-NRE for training, which is a neural relation extraction toolkit implemented with a convolutional neural network (Han *et al.*, 2019). After the model is trained, it is used on the entire dataset for the extraction of relationships among entities.

3.4.3 Semantic relationship discovery among topics. This paper uses semantic relationships among patent entities to obtain the semantic relationship between two topics. Firstly, a semantic relationship matrix is constructed among entities between each pair of topics, then an indicator is proposed to discover semantic relationships among topics.

(1) Construction of semantic relationship matrix between two topics

The semantic relationship between two topics in the adjacent time window could be determined by semantic relationships among entities belonging to each topic. Specifically, T_i^t and T_j^{t+1} denote two topics in adjacent time window t and $t+1$, where T_i^t contains x patent entities and T_j^{t+1} contains y patent entities. Thus an x by y matrix can be formed to denote semantic relationships between these two topics, in which each element denotes a semantic relationship between two entities.

The $x * y$ matrix is shown in Formula (4), in which the rows denote patent entities $a_1 \ldots a_i \ldots a_x$ included in topic T_i^t, and the columns denote patent entities $a_1 \ldots a_j \ldots a_y$ included in topic T_j^{t+1}, and a_{ij} represents a semantic relationship between a_i in topic T_i^t and a_j in topic T_j^{t+1}.

$$\begin{bmatrix} a_{11}=M & a_{12}=FA & a_{13}=/ & \ldots & a_{1j}=/ & \ldots & a_{1y}=M \\ a_{21}=E & a_{22}=/ & a_{23}=E & \ldots & a_{2j}=M & \ldots & a_{2y}=M \\ a_{31}=FA & a_{32}=C & a_{33}=M & \ldots & a_{3j}=/ & \ldots & a_{3y}=M \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{i1}=E & a_{i2}=/ & a_{i3}=FA & \ldots & a_{ij}=FA & \ldots & a_{iy}=FA \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{x1}=E & a_{x2}=M & a_{x3}=/ & \ldots & a_{xj}=/ & \ldots & a_{xy}=E \end{bmatrix} \tag{4}$$

(2) Semantic relationship discovery of topic evolution

The semantic relationship of topic evolution is determined by the highest proportion of semantic relationships among patent entities in each pair of topics. Therefore, we first compute the proportion of every type of semantic relationships among patent entities and then select the highest one as the semantic relationship of topic evolution.

The proportion of a certain semantic relationship in the above $x * y$ semantic relationship matrix is equal to the number of the certain semantic relationship divided by the number of all semantic relationships. The calculation method is shown in Formula (5), in which P_{SM} represents the proportion of a certain semantic relationship, N_{SM} represents the number of a certain semantic relationship between two topics, and $x * y$ represents the number of all valid semantic relationships.

$$P_{SM} = \frac{N_{SM}}{x * y} \tag{5}$$

After the proportions of five predefined semantic relationships are computed, the semantic relationship between two topics can be determined with the highest one. Thus the semantic relationship of topic evolution is obtained.

4. Experiments and results

The experiment consists of three parts to validate the effectiveness of the method. Firstly, Bi-LSTM-CRF model is trained to extract patent entities on the entire dataset, and each of them is represented as a multi-dimensional vector. Then, topic evolution paths and semantic relationships are discovered in the area of UAV, and the results are visualized and analyzed.

4.1 Patent entity extraction and representation

4.1.1 Patent entity extraction based on Bi-LSTM-CRF. In this experiment, 100 patents are randomly selected from the entire dataset for training the model, in which patent entities are manually labeled by authors using open-source text annotation tool "doccano" (Nakayama *et al.*, 2018). This process consists of two steps: first, the initial entity annotation is done by one person and then the annotation is checked by another person for validation. Finally, about 1,931 patent entities are manually labeled. After this, the labeling results can be exported as text files for subsequent Bi-LSTM-CRF training.

The labeled dataset is then divided into training and testing datasets with a ratio of 8 to 2, and the package "keras" is used to train Bi-LSTM-CRF model. For better performance, this model uses the vectors generated by word2vec as the word vectors input. After multiple training and parameter adjustments, the accuracy of the model exceeds 90% and the loss falls below 5.8 and both of them tend to be stable. At this moment, the model performs best with important parameters setting, as shown in Table 2.

After the model is trained, it is applied to label the words on the entire dataset. These labeled words are then combined as patent entities. Some extracted patent entities are shown in Table 3.

4.1.2 Semantic representation of patent entities. The semantic representation of a patent entity is obtained by combining the word vectors it contains. The main parameters are set as follows: "sg = 1" means the skip-gram model is used in word2vec, which is usually selected for small training data; "size = 100" means the vector dimension is 100, i.e. each word is represented as a point in a 100-dimensional space; "window = 10" means ten words around the current position are used as context information; "hs = 0" means negative sampling is used for training optimization. Some examples of semantic representation of patent entities are shown in Table 4, in which the first column represents the extracted patent entities and the remaining columns of each row together denote the corresponding semantic vector.

Table 2. Important parameters setting in Bi-LSTM-CRF model

Parameter	Value
Batch_size	20
Epoch	100
Learning_rate	0.001
Dropout	0.5

Table 3. Examples of extracted patent entities

ID	Patent entity
1	Image analyzing unit
2	Automatic piloting unit
3	Rotary wing unmanned aerial vehicle
4	Flight stability condition efficiently

4.2 Discovery of topic evolution path

4.2.1 Topic identification based on K-means. The K-Means package in scikit-learn library of Python 3.6.2 is used to cluster patent entities based on cosine similarity. Through multiple clustering attempts and the smooth decline of SSE, the number of clusters in different time windows is obtained. The topics identified in 2015, 2016 and 2017 are shown in Tables 5–7, respectively. Because each topic identified by K-means has a cluster center which is the

Table 4. Examples of semantic representation of patent entities

UAV recovery system	−0.71396101	−0.36148526	0.71284347	0.38837655	0.64201843
Small air/submarine launched UAV	0.16211763	−0.20719976	0.78403493	0.27418809	0.67523550
UAV fuselage	−0.25796422	0.54878825	0.21478591	0.21478591	−0.44911978
Generated multi-rotor UAV	0.02157145	−0.02958778	0.00523921	−0.21977860	0.07056160
UAV operation signals	−0.16479148	−0.06944223	0.06872799	0.49096080	0.21206408
Delivery management service	0.18042174	−0.18189536	0.16898134	0.15731029	0.25604171
UAV flying system	0.10291060	−0.17354462	0.27102665	0.20033685	0.29067087
Controllable clutch assembly	0.01991126	−0.01541813	0.03186072	0.03006795	0.02149010
Tripod head	0.23223495	−0.23116301	0.27135407	0.19398869	0.31752463
Transportable ground station	0.03029315	−0.01738979	0.04095134	0.03399104	0.03372387

Table 5. Number of topics and the representative patent entities in 2015

Topic no	Number of patent entities	Representative patent entities
1	1,701	UAV body, parts and components
2	1,389	UAV communication media and communication functions
3	1,680	UAV function module
4	62	Location of component installation
5	1,862	UAV software system
6	1,708	UAV kinetic energy unit

Table 6. Number of topics and the representative patent entities in 2016

Topic no	Number of patent entities	Representative patent entities
1	833	UAV camera device and picture transmission system
2	2,407	UAV control system
3	2,437	UAV automatic communication
4	2,473	UAV application field
5	23	UAV installation sequence
6	2,325	UAV kinetic energy device automation
7	2,598	UAV transportation software platform
8	64	UAV parts installation position

Table 7. Number of topics and the representative patent entities in 2017

Topic no	Number of patent entities	Representative patent entities
1	3,163	UAV communication application
2	157	UAV camera unit
3	1,771	UAV image transmission system
4	3,211	UAV automatic communication device
5	3,209	UAV control system and its apparatus
6	32	Location of UAV mechanism
7	3,158	UAV navigation, cluster software

arithmetic mean of all the points belonging to this cluster, this paper labels each topic with multiple entities closest to the cluster center following the general practice (Jayabharathy *et al.*, 2011). The number of patent entities in each topic and some representative patent entities are displayed in these tables, which are useful for the subsequent topic evolution interpretation and understanding.

As shown in the above tables, topics in 2015 are more focused on UAV hardware and functions, topics in 2016 mention the image transmission system and the communication system, and topics in 2017 are more focused on automatic communication and control and platform-based software.

4.2.2 Identification of the type of topic evolution based on knowledge flow. This paper determines the type of topic evolution via knowledge outflow and knowledge inflow. If different types of topic evolution are identified, i.e. development, division, integration, extinction and emerging, then the topic evolution paths can be formed. This paper chooses two cases for detailed description, i.e. topic evolution from 2015 to 2016 and topic evolution from 2016 to 2017. In the experiment, the threshold used to judge semantic common entities is set to 0.7. The reasons for this threshold setting are as follows: on the one hand, the randomly selected pairs of patent entities with a threshold greater than 0.7 almost have the same semantics; on the other hand, the randomly selected pairs of patent entities with a threshold less than 0.6 have obviously different semantics.

(1) Topic evolution from 2015 to 2016

Knowledge outflow and inflow of topics from 2015 to 2016 are shown in Table 8, in which T_i represents the ith topic in different years. For example, "T_1(1,701)" means the 1st topic in 2015 contains 1,701 patent entities and "T_1(833)" means the 1st topic in 2016 contains 833 patent entities. In addition, each element in Table 8 represents the knowledge outflow/inflow between two topics, in which the italicized number denotes the knowledge inflow. For example, the first element "5%/*2%*" means 5% of T_1^{2015} outflows to T_1^{2016} and 2% of T_1^{2016} is inflowed from T_1^{2015}.

In Table 8, each row means the knowledge of a topic in 2015 outflows to other topics in 2016. Taking the first row as an example, it means that 5%, 14 and 32% of T_1 in 2015 outflows to T_1, T_2 and T_3 in 2016, respectively. Therefore, we can find a lot of division evolutions, e.g. T_1^{2015} divides into T_2, T_3, T_6 and T_7 in 2016 with an outflow threshold greater than 14%, and T_2^{2015} divides into T_1, T_2 and T_3 in 2016 with an outflow threshold greater than 20%. These results are consistent with the actual situations. For example, T_2^{2015} divides into three topics in 2016 reveals the actual evolution of UAV communication functions: UAV communication was starting to receive attention in 2015 and subdivided into three specific devices or systems in 2016, i.e. UAV camera device, picture transmission system and UAV control system.

Similarly, each column means the knowledge of a topic in 2016 is inflowed from multiple topics in 2015. Taking the first column as an example, it means that 2%, 66 and 7% of T_1 in 2016 are derived from T_1, T_2 and T_4 in 2015, respectively. Therefore, we can find a lot of evolution types of integration, e.g. T_2, T_3 and T_5 in 2015 are integrated as T_2^{2016} with knowledge inflow greater than 20%, and T_1 and T_2 in 2015 are integrated as T_3^{2016} with knowledge inflow greater than 14%. The effectiveness of these results can be proved from the description of the topics, three topics in 2015, i.e. T_2 (UAV communication media and communication functions), T_3 (UAV function module) and T_5 (UAV software system) are truly part of T_2^{2016} (UAV control system).

Both considering the knowledge outflow and inflow, we can find other types of topic evolution. For example, 100% of T_4^{2015} outflows to T_8^{2016} and 96% of T_8^{2016} is inflowed from T_4^{2015}, which means there is a development evolution between them. It can be proved from the

description of the topics, in which T_4^{2015} mentions "location of component installation" and T_8^{2016} mentions "UAV parts installation position". Both of them are different expressions of "UAV installation position".

In addition, the total proportion of knowledge inflow of a topic could judge whether it is an emerging topic. For example, the total proportion of knowledge inflow of T_7^{2016} is 40%, which means most of this topic's knowledge is new and could be regarded as an emerging topic. This result is in line with the development of UAV: as the maturity of UAV hardware and functional modules, "UAV transportation software platform" was clearly regarded as the next hotspot and frontier of UAV in 2016.

(2) Topic evolution from 2015 to 2016 discovered by traditional method

The traditional method used for topic evolution is often based on common keywords. Therefore, this paper takes common entities among topics as the traditional method for comparison. This method takes the number of common entities for computation of knowledge outflow and inflow and identifies the topic evolution in the same way as the proposed method. The results are shown in Table 9, in which each element represents the knowledge outflow/inflow between two topics and the italicized number denotes the knowledge inflow.

It can be clearly seen that the knowledge outflow and knowledge inflow computed with the traditional method is lower than the proposed method. It leads to that the topic evolution is hardly identified in Table 9, e.g. it is hard to judge the evolution type of T_1, T_3, T_5 and T_6 in 2015, and it is also hard to judge where the topics T_2, T_3, T_4, T_6 and T_7 in 2016 are derived from. The reason is that the traditional method cannot identify the patent entities which have the same meaning but different expressions. Therefore, the type of topic evolution identified from the traditional method is not accurate enough.

In order to clearly interpret what leads to the difference of knowledge flow between the two methods, this paper displays some patent entities with different expressions but the same semantics. As shown in Table 10, pairs of entities with the high semantic similarity between T_2^{2015} (UAV communication medium and function) and T_1^{2016} (camera device and image transmission system) are selected and shown.

Table 8. Knowledge outflow and inflow from 2015 to 2016

2016 / 2015	T_1(833)	T_2(2,407)	T_3(2,437)	T_4(2,473)	T_5(23)	T_6(2,325)	T_7(2,598)	T_8(64)
T_1(1,701)	5 %/*2%*	14 %/*6%*	32 %/*22%*	12 %/*10%*	0 %/*0%*	20 %/*15%*	16 %/*8%*	0 %/*0%*
T_2(1,389)	40 %/*66%*	21 %/*12%*	24 %/*14%*	7 %/*4%*	5 %/*3%*	10 %/*6%*	1 %/*0%*	6 %/*2%*
T_3(1,680)	4 %/*0%*	24 %/*17%*	18 %/*11%*	20 %/*14%*	0 %/*0%*	14 %/*6%*	19 %/*5%*	0 %/*0%*
T_4(62)	100 %/*7%*	11 %/*8%*	12 %/*7%*	0 %/*0%*	37 %/*100%*	1 %/*0%*	0 %/*0%*	100 %/*96%*
T_5(1,862)	3 %/*0%*	23 %/*17%*	18 %/*2%*	13 %/*7%*	0 %/*0%*	17 %/*10%*	24 %/*17%*	0 %/*0%*
T_6(1,708)	5 %/*2%*	14 %/*2%*	8 %/*4%*	*34%/23%*	*0%/0%*	23 %/*16%*	16 %/*10%*	0 %/*0%*

Table 9. Knowledge outflow and inflow from 2015 to 2016 by traditional method

2016 / 2015	T_1(833)	T_2(2,407)	T_3(2,437)	T_4(2,473)	T_5(23)	T_6(2,325)	T_7(2,598)	T_8(64)
T_1(1,701)	3 %/*6%*	2 %/*2%*	3 %/*2%*	2 %/*1%*	0 %/*0%*	2 %/*2%*	1 %/*1%*	0 %/*0%*
T_2(1,389)	14 %/*23%*	4 %/*2%*	4 %/*2%*	2 %/*1%*	0 %/*0%*	3 %/*2%*	2 %/*1%*	1 %/*6%*
T_3(1,680)	3 %/*6%*	2 %/*2%*	2 %/*2%*	2 %/*1%*	0 %/*0%*	2 %/*1%*	2 %/*1%*	0 %/*0%*
T_4(62)	19 %/*1%*	1 %/*0%*	0 %/*0%*	0 %/*0%*	37 %/*99%*	0 %/*0%*	0 %/*0%*	41 %/*39%*
T_5(1,862)	3 %/*7%*	2 %/*1%*	2 %/*2%*	2 %/*1%*	0 %/*0%*	3 %/*2%*	2 %/*1%*	0 %/*0%*
T_6(1,708)	3 %/*6%*	3 %/*2%*	3 %/*2%*	2 %/*1%*	0 %/*0%*	2 %/*1%*	2 %/*1%*	0 %/*0%*

From Table 10, it can be found that the meaning of each pair of entities is almost the same, which should be considered in topic evolution analysis. For example, "3D", "UAV" and "GPS" are the abbreviations of "three dimensional", "unmanned flight vehicle" and "global positioning system", respectively. In addition, "four-rotor" has the same meaning with "quadrotor", and "digital inflight message" has the same semantics with "communication devices". These pairs of entities could be helpful and useful for accurate topic evolution identification.

(3) Topic evolution from 2016 to 2017

Knowledge outflow and inflow of topics from 2016 to 2017 are shown in Table 11, in which T_i represents the ith topic in different years. Each element in Table 11 denotes knowledge outflow and inflow respectively, and the italicized one denotes the knowledge inflow.

In Table 11, each row means the knowledge of a topic in 2016 outflows to other topics in 2017. Thus we can find a lot of division evolutions, e.g. T_1^{2016} is mainly divided into T_2 and T_3 in 2017, and T_2^{2016} is mainly divided into T_1, T_4 and T_5 in 2017. The effectiveness of these results can be proved from the description of the topics, T_1^{2016} (UAV camera device and picture transmission system) is obviously divided into two topics in 2017, i.e. T_2 (UAV camera unit) and T_3 (UAV image transmission system).

Similarly, each column means the knowledge of a topic in 2017 is inflowed from multiple topics in 2016. Thus we can find a lot of integration evolutions, e.g. T_2, T_3, T_4 and T_6 in 2016 are integrated as T_1^{2017}, and T_3 and T_6 in 2016 are integrated as T_4^{2017}. These results are consistent with the actual situation. For example, four topics in 2016 are integrated into T_1^{2017} reveals the actual evolution of UAV applications: various components of UAV were applied in different fields in 2016 and integrated into one complete UAV application system in 2017.

Both considering the knowledge outflow and inflow, we can find other types of topic evolution from 2016 to 2017. For example, 64% of T_1^{2016} outflows to T_2^{2017} and 100% of

Entities of T_2 in 2015	Entities of T_1 in 2016	Semantic similarity
Unmanned flight vehicle	UAV	0.98
Three dimensional	3D	0.96
Aerial photography	Camera assembly	0.91
Unmanned flight vehicle	Unmanned aerial vehicles	0.9
Global positioning system	GPS	0.9
Four-rotor	Quadrotor	0.84
Digital inflight message	Communication devices	0.8

Table 10. Pairs of entities have the same semantics but different expressions between T_2^{2015} and T_1^{2016}

2017 2016	T_1(3,163)	T_2(157)	T_3(1,771)	T_4(3,211)	T_5(3,209)	T_6(32)	T_7(3,158)
T_1(833)	9 %/*6%*	64 %/*100%*	89 %/*42%*	8 %/*0%*	2 %/*0%*	9 %/*1%*	2 %/*1%*
T_2(2,407)	30 %/*22%*	0 %/*0%*	13 %/*6%*	19 %/*9%*	24 %/*18%*	0 %/*0%*	14 %/*9%*
T_3(2,437)	22 %/*22%*	0 %/*0%*	18 %/*11%*	38 %/*28%*	12 %/*11%*	0 %/*0%*	11 %/*8%*
T_4(2,473)	29 %/*22%*	0 %/*0%*	6 %/*4%*	17 %/*12%*	21 %/*16%*	0 %/*0%*	27 %/*21%*
T_5(23)	0 %/*0%*	13 %/*7%*	64 %/*1%*	0 %/*0%*	0 %/*0%*	93 %/*91%*	0 %/*0%*
T_6(2,325)	21 %/*15%*	0 %/*0%*	8 %/*3%*	27 %/*19%*	29 %/*21%*	0 %/*0%*	15 %/*5%*
T_7(2,598)	11 %/*9%*	0 %/*0%*	2 %/*1%*	14 %/*9%*	29 %/*21%*	0 %/*0%*	43 %/*35%*
T_8(64)	4 %/*2%*	73 %/*30%*	65 %/*1%*	3 %/*1%*	0 %/*0%*	0 %/*0%*	0 %/*0%*

Table 11. Knowledge outflow and knowledge inflow from 2016 to 2017

T_2^{2017} is inflowed from T_1^{2016}, which means there is a development evolution between them. It can also be treated as an extinction because parts of knowledge is extinct when flows to the next year. This is consistent with the actual situation, because "UAV camera device" and "picture transmission system" have gradually matured in 2016. Similarly, there is a development evolution between T_5^{2016} and T_6^{2017}. It can be proved from the description of the topics, in which T_5^{2016} mentions "UAV installation sequence" and T_6^{2017} mentions "Location of UAV mechanism". Both of them are expressions of the location of installations.

(4) Visualization of topic evolution path from 2015 to 2017

After the topic evolution paths are discovered, this paper visualizes them for a better understanding of the whole process. As shown in Figure 5, the blue line represents integration/division evolution, the yellow line represents extinction evolution, the green line represents development evolution, and the width of the line represents the proportion of knowledge flow.

It can be seen more intuitively from Figure 5 that the knowledge outflow and inflow among multiple topics in different time series. From knowledge flow, we can easily find the topic evolution path from 2015 to 2017. For example, only a small part of knowledge in T_7^{2016} (UAV transportation software platform) is derived from topic T_5^{2015} (UAV-software system), which means T_7^{2016} is an emerging topic. Then, the topic T_7^{2016} divides into two topics T_5 and T_7 in 2017 as topics evolving.

We can find other topic evolution paths via knowledge flow. For example, two topics in 2015, i.e. T_2 "UAV communication media and communication functions" and T_1 "UAV body, parts and components", are integrated as T_3^{2016} called "UAV automatic communication", and then T_3^{2016} divides into two topics T_4 and T_1 in 2017. In addition, other compound paths can be found, e.g. T_4^{2015} develops into T_8^{2016}, and then T_8^{2016} divides into T_2 and T_3 in 2017, T_2 and T_4 in 2015 are integrated into T_1^{2016}, and then T_1^{2016} becomes extinct or develops into T_2^{2017}, etc.

4.3 Semantic relationship discovery of topic evolution

4.3.1 Semantic relationship extraction among patent entities. After data preprocessing, the semantic relationships are manually labeled on 100 randomly selected patents, which is the same as patent entity extraction. The relation is annotated using the annotation tool "brat"

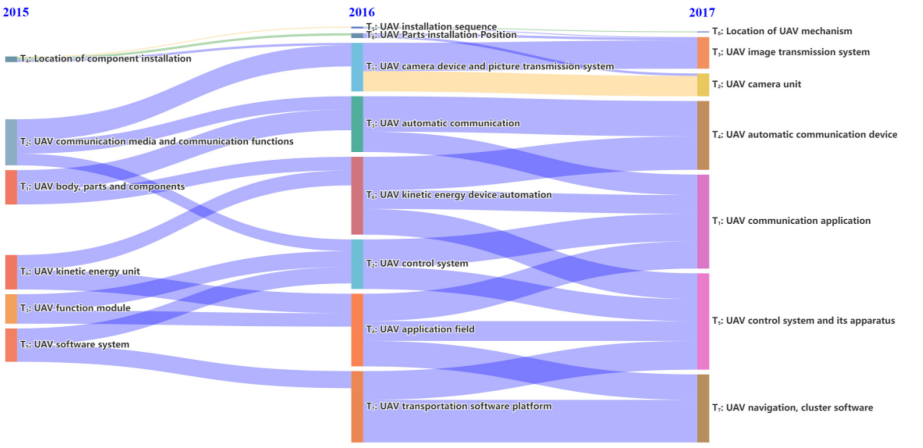

Figure 5.
Visualization of topic evolution path from 2015 to 2017

(Stenetorp *et al.*, 2012). This process consists of two steps: first, the initial relation annotation is done by one person, then the annotation is checked by another person for validation. Finally, about 1,287 entity relations are manually labeled. After this, the labeled results can be exported as text files for subsequent Bi-LSTM-CRF training.

The labeled dataset is then divided into training and testing datasets with a ratio of 8 to 2, and a neural network model open-NRE is used for training the model. After 100 training iterations, the accuracy of the model on the training set exceeds 97% and becomes stable. At this moment, the model performs best with the following important parameter settings, i.e. "batch size = 1", "learning rate = 1e−1", "weight decay = 1e−5", "maximum sentence length = 40", and "max number of training epochs = 100".

Then semantic relationship between each pair of entities on the entire dataset can be extracted through the trained model. In summary, a total of 1,287 semantic relationships are extracted, in which "Mechanical-relationship" accounts for 68.09%, "Efficiency-relationship" accounts for 9.49%, "Function-Area relationship" accounts for 5.24%, "Function-Realization relationship" accounts for 10.19%, and "Control-relationship" accounts for 6.99%.

After the model is trained, it is applied to predict the relationship between two patent entities on the entire dataset. Some predicted semantic relationships between two entities are shown in Table 12.

4.3.2 Semantic relationship discovery among topics. The proportions of five predefined semantic relationships among entities belonging to each pair of topics are computed. Then we find that the highest proportion is always the mechanical relationship, which is consistent with the training dataset. The reason may be that the area of UAV is related to a mechanical device, in which the mechanical relationship dominates the semantic relationships among patent entities. Therefore, this paper uses the difference of proportions between the manually labeled dataset and the identified topics to determine the type of semantic relationship.

(1) Semantic relationship between 2015 and 2016

The result of semantic relationships among topics between 2015 and 2016 is shown in Table 13. It is useful to interpret what leads to the topic evolution. For example,

ID	A pair of patent entities		Semantic relationship
1	Rotary wing unmanned aerial vehicle	Image transmission	Function realization
2	Unmanned aircraft	Fuselage	Mechanical
3	Drone for observing scene	Sensor	Control
4	Sleeve	Small volume	Efficacy
5	Trigger circuit board	Charging process	Function area

Table 12. Examples of extracted semantic relationships between patent entities

2016 2015	T_1(833)	T_2(2,407)	T_3(2,437)	T_4(2,473)	T_5(23)	T_6(2,325)	T_7(2,598)	T_8(64)
T_1(1,701)			E			E		
T_2(1,389)	M	E	E					
T_3(1,680)		E		FA				
T_4(62)	M				M			M
T_5(1,862)		E					E	
T_6(1,708)				FA		FA		

Table 13. Semantic relationships among topics between 2015 and 2016

T_3 (UAV functional module) and T_6 (UAV kinetic energy device) in 2015 are integrated into topic T_4^{2016} (UAV application field), the reason may be that there are both "Function-Area" relationships among them. It means the functions in T_3 and T_6 in 2015 have been applied in multiple other areas and applications in 2016.

In addition, T_2^{2015} (UAV communication media and communication functions) is divided into T_1 (UAV camera device and picture transmission system), T_2 (UAV control system) and T_3 (UAV automatic communication) in 2016. The reason may be that the semantic relationship between T_2^{2015} and T_1^{2016} is "Mechanical", and the semantic relationships between T_2^{2015} and T_2, T_3 in 2016 are both "Efficacy". The result also shows that T_4^{2015} develops to T_8^{2016} is driven by mechanical semantic relationship.

(2) Semantic relationship between 2016 and 2017

The result of semantic relationships among topics between 2016 and 2017 is shown in Table 14. It is useful to interpret what causes the topic evolution in this time series. For example, T_3 (UAV automatic communication) and T_6 (UAV kinetic energy device automation) in 2016 are integrated into topic T_4^{2017} (UAV automatic communication device), the reason may be that there are both efficacy relationships among them. In addition, T_1^{2016} (UAV camera device and picture transmission) develops into T_2^{2017}(UAV camera unit) because the semantic relationship between them is "mechanical". It means that one topic is part of or containment of devices mentioned in the other topic.

(3) Visualization of semantic relationship among topics between 2015 and 2017

After the semantic relationships among topics are discovered, this paper visualizes them together with the topic evolution path for a better understanding of the whole process. As shown in Figure 6, the semantic relationships among topics are labeled on the line.

It can be seen more intuitively how topics evolve and what leads to topic evolution. Combined with knowledge flow, we can easily find the topic evolution path and semantic relationships among topics from 2015 to 2017. For example, T_7^{2016} comes from topic T_5^{2015} with an efficacy relationship and divides into T_5 and T_7 in 2017 with efficacy and mechanical relationship, respectively. In addition, T_1 and T_2 in 2015 are integrated as T_3^{2016} with efficacy relationship, and T_3^{2016} divides into T_1 and T_4 in 2017 with efficacy and mechanical relationship, respectively.

5. Conclusion

This paper aims to propose a method for better discovering topic evolution path and semantic relationship from the perspective of patent entity extraction and semantic representation.

Table 14.
The result table of the semantic relationship between the topics in 2016–2017

2017 2016	T_1(3,163)	T_2(157)	T_3(1,771)	T_4(3,211)	T_5(3,209)	T_6(32)	T_7(3,158)
T_1(833)		M	E				
T_2(2,407)	M				E		
T_3(2,437)	M			E			
T_4(2,473)	M				E		E
T_5(23)			FA			M	
T_6(2,325)	E			E	M		
T_7(2,598)					E		M
T_8(64)		M	M				

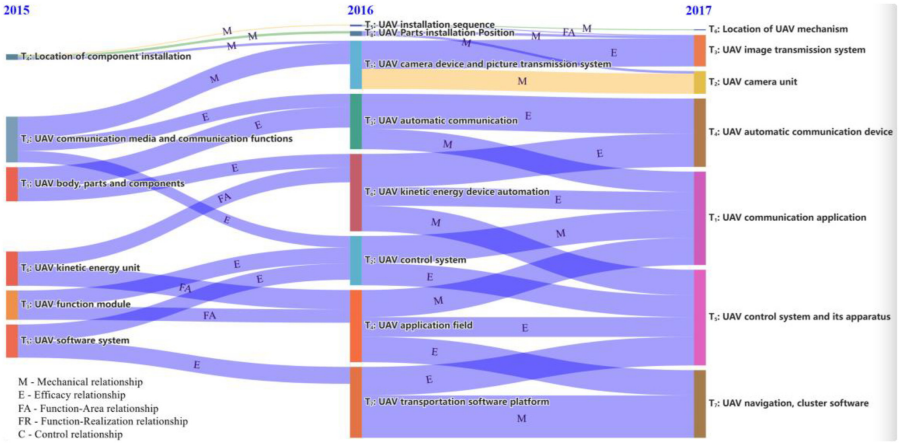

Figure 6.
Visualization of semantic relationships between 2015 and 2017

This method provides a research idea that combining entity extraction and representation with topic evolution analysis. It could give a more precise and detailed description of relationships among topics. In addition, the designed method for patent entity extraction and representation could be extended to other applications and areas. Moreover, the knowledge flow computation based on semantic common entities could not only be used for topic evolution path discovery, but also for emerging technology detection and radical innovation identification. Furthermore, a relationship extraction method is borrowed for semantic relationship discovery among topics, for better detection, interpretation and understanding of the topic evolution.

According to the empirical study, the results have proved the effectiveness of the proposed method. Compared with the traditional method for topic evolution path discovery, this method performs better by providing a more accurate and comprehensive result. It can help relevant researchers quickly understand technology paths and changes in a specific field. In addition, this paper further extracts semantic relationships automatically, which gives a clearer and more detailed correlation among topics. It can provide suggestions to explain what drives topic evolution.

This method provides a new perspective for topic evolution path discovery. However, it is a preliminary study and still needs to be further improved in the following directions. Firstly, the semantic relationships among patent entities in this paper are predefined for mechanical devices or electronic devices, which could be easily extended to related domains with minor modifications, but hard to be adapted in unrelated areas directly, such as medical, sports and computer science. In addition, it also needs to label the new dataset manually in different fields for training the model. Therefore, the model's feasibility is limited when applied to unrelated technical fields. Secondly, this paper only identifies synonymous entities for topic evolution analysis by using state-of-the-art deep learning methods, but polysemous entities in different time series have not been considered. Therefore, future related studies should consider the effect of polysemous entities on topic evolution path discovery and choose the latest and best alternative deep learning methods. Finally, the proposed method only discovers topic evolution paths and relationships on patents, while in real life, besides patent data, there are a large amount of other rich technical information or innovative data, e.g. technology reports and scientific literature. Therefore, it is necessary to design different entity extraction and representation learning methods on different data types for topic evolution analysis.

References

An, X., Li, J., Xu, S., Chen, L. and Sun, W. (2021), "An improved patent similarity measurement based on entities and semantic relations", *Journal of Informetrics*, Vol. 15 No. 2, 101135, doi: 10.1016/j.joi.2021.101135.

Bachman, J.A., Gyori, B.M. and Sorger, P.K. (2018), "FamPlex: a resource for entity recognition and relationship resolution of human protein families and complexes in biomedical text mining", *BMC Bioinformatics*, Vol. 19 No. 1, pp. 248-262, doi: 10.1186/s12859-018-2211-5.

Bhatia, P., Celikkaya, B., Khalilia, M. and Senthivel, S. (2019), "Comprehend medical: a named entity recognition and relationship extraction web service", in Arif Wani, M., Khoshgoftaar, T.M., Wang, D., Wang, H. and Seliya, N. (Eds), *2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, Boca Raton, FL, USA, Piscataway, NJ, December 16-19, 2019, pp. 1844-1851, doi: 10.1109/ICMLA.2019.00297.

Chang, P.L., Wu, C.C. and Leu, H.J. (2010), "Using patent analyses to monitor the technological trends in an emerging field of technology: a case of carbon nanotube field emission display", *Scientometrics*, Vol. 82 No. 1, pp. 5-19, doi: 10.1007/s11192-009-0033-y.

Chen, L., Xu, S., Zhu, L., Zhang, J., Lei, X. and Yang, G. (2020), "A deep learning based method for extracting semantic information from patent documents", *Scientometrics*, Vol. 125 No. 1, pp. 289-312, doi: 10.1007/s11192-020-03634-y.

Chen, L., Xu, S., Zhu, L., Zhang, J., Xu, H. and Yang, G. (2022), "A semantic main path analysis method to identify multiple developmental trajectories", *Journal of Informetrics*, Vol. 16 No. 2, 101281, doi: 10.1016/j.joi.2022.101281.

Chowdhury, S., Dong, X., Qian, L., Li, X., Guan, Y., Yang, J. and Yu, Q. (2018), "A multitask bi-directional RNN model for named entity recognition on Chinese electronic medical records", *BMC Bioinformatics*, Vol. 19 No. 17, pp. 499-509, doi: 10.1186/s12859-018-2467-9.

Coulter, N., Monarch, I. and Konda, S. (1998), "Software engineering as seen through its research literature: a study in co-word analysis", *Journal of the American Society for Information Science*, Vol. 49 No. 13, pp. 1206-1223, doi: 10.1002/(SICI)1097-4571(1998)49:133.0.CO;2-F.

Das, S., Deb, N., Cortesi, A. and Chaki, N. (2021), "Sentence embedding models for similarity detection of software requirements", *SN Computer Science*, Vol. 2 No. 2, pp. 1-11, doi: 10.1007/s42979-020-00427-1.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018), "Bert: pre-training of deep bidirectional transformers for language understanding", *arXiv preprint arXiv:1810.04805*, doi: 10.48550/arXiv.1810.04805.

Érdi, P., Makovi, K., Somogyvári, Z., Strandburg, K., Tobochnik, J., Volf, P. and Zalányi, L. (2013), "Prediction of emerging technologies based on analysis of the US patent citation network", *Scientometrics*, Vol. 95 No. 1, pp. 225-242, doi: 10.1007/s11192-012-0796-4.

Haghighian Roudsari, A., Afshar, J., Lee, W. and Lee, S. (2021), "PatentNet: multi-label classification of patent documents using deep learning based language understanding", *Scientometrics*, Vol. 127 No. 1, pp. 207-231, doi: 10.1007/s11192-021-04179-4.

Han, X., Gao, T., Yao, Y., Ye, D., Liu, Z. and Sun, M. (2019), "OpenNRE: an open and extensible toolkit for neural relation extraction", in Padó, S. and Huang, R. (Eds), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Hong Kong, China, Stroudsburg, PA, System Demonstrations, pp. 169-174, available at: https://aclanthology.org/D19-3029.

Huang, Z., Xu, W. and Yu, K. (2015), "Bidirectional LSTM-CRF models for sequence tagging", *arXiv preprint arXiv:1508.01991*, doi: 10.48550/arXiv.1508.01991.

Hummon, N.P. and Dereian, P. (1989), "Connectivity in a citation network: the development of DNA theory", *North-Holland*, Vol. 11 No. 1, pp. 39-63, doi: 10.1016/0378-8733(89)90017-8.

Jayabharathy, J., Kanmani, S. and Parveen, A. (2011), "Document clustering and topic discovery based on semantic similarity in scientific literature", *2011 IEEE 3rd International Conference on Communication Software and Networks*, IEEE, Piscataway, NJ, Xi'an, China, 27-29 May 2011, pp. 425-429, doi: 10.1109/ICCSN.2011.6014600.

Kambhatla, N. (2004), "Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction", *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, Association for Computational Linguistics, Barcelona, Spain, Stroudsburg, PA, July, 2004, p. 22-es, doi: 10.3115/1219044.1219066.

Kim, M., Park, Y. and Yoon, J. (2016), "Generating patent development maps for technology monitoring using semantic patent-topic analysis", *Computers and Industrial Engineering*, Vol. 98, pp. 289-299, doi: 10.1016/j.cie.2016.06.006.

Kim, E.H., Jeong, Y.K., Kim, Y. and Song, M. (2022), "Exploring scientific trajectories of a large-scale dataset using topic-integrated path extraction", *Journal of Informetrics*, Vol. 16 No. 1, 101242, doi: 10.1016/j.joi.2021.101242.

Lei, Z., Wang, G. and Zou, X. (2016), "A study of Chinese document representation and classification with Word2vec", *International Symposium on Computational Intelligence and Design*, IEEE, Hangzhou, China, NY, 10-11 December 2016, Vol. 1, pp. 298-302, doi: 10.1109/ISCID.2016.1075.

Li, J., Sun, A., Han, J. and Li, C. (2020), "A survey on deep learning for named entity recognition", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 34 No. 1, pp. 50-70.

Liu, J., Ren, H., Wu, M., Wang, J. and Kim, H.J. (2018), "Multiple relations extraction among multiple entities in unstructured text", *Soft Computing*, Vol. 22 No. 13, pp. 4295-4305, doi: 10.1007/s00500-017-2852-8.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. (2019), "Roberta: a robustly optimized bert pretraining approach", arXiv preprint, Vol. arXiv: 1907.11692, doi: 10.48550/arXiv.1907.11692.

Liu, H., Chen, Z., Tang, J., Zhou, Y. and Liu, S. (2020), "Mapping the technology evolution path: a novel model for dynamic topic detection and tracking", *Scientometrics*, Vol. 125 No. 3, pp. 2043-2090, doi: 10.1007/s11192-020-03700-5.

Lu, L. and Liu, J.S. (2014), "A survey of intellectual property rights literature from 1971 to 2012: the main path analysis", in Kocaoglu, D., Anderson, T., Daim, T., Kozanoglu, D., Niwa, K. and Perman, G. (Eds), *Portland International Conference on Management of Engineering and Technology, 27-31, July, 2014*, IEEE, Kanazawa, Japan, New York, NY, Infrastructure and Service Integration, pp. 1274-1280.

Ma, X. and Hovy, E. (2016), "End-to-end sequence labeling via bi-directional lstm-cnns-crf", *arXiv preprint arXiv:1603.01354*, doi: 10.48550/arXiv.1603.01354.

Martinelli, A. (2011), "An emerging paradigm or just another trajectory? Understanding the nature of technological changes using engineering heuristics in the telecommunications switching industry", *Research Policy*, Vol. 41 No. 2, pp. 414-429, doi: 10.1016/j.respol.2011.10.012.

Mikolov, T., Kombrink, S., Deoras, A., Burget, L. and Cernocky, J.H. (2011), "RNNLM - recurrent neural network language modeling toolkit", *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Hilton Waikoloa Village Resort, Big Island*, IEEE Signal Processing Society, Hawaii, US, Piscataway, NJ, pp. 196-201.

Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y. and Liang, X. (2018), "Doccano: text annotation tool for human", available at: https://github.com/doccano/doccano.

Nasar, Z., Jaffry, S.W. and Malik, M.K. (2021), "Named entity recognition and relation extraction: state-of-the-art", *ACM Computing Surveys (CSUR)*, Vol. 54 No. 1, pp. 1-39, doi: 10.1145/3445965.

Pilkington, A. and Meredith, J. (2009), "The evolution of the intellectual structure of operations management-1980-2006: a citation/co-citation analysis", *Journal of Operations Management*, Vol. 27 No. 3, pp. 185-202, doi: 10.1016/j.jom.2008.08.001.

Qi, Y., Zhu, N., Zhai, Y. and Ding, Y. (2018), "The mutually beneficial relationship of patents and scientific literature: topic evolution in nanoscience", *Scientometrics*, Vol. 115 No. 2, pp. 893-911, doi: 10.1007/s11192-018-2693-y.

Rao, Q., Wang, P. and Zhang, G. (2015), "Text feature analysis on SAO structure extraction from Chinese patent literature", *Acta Scientiarum Naturalium Universitatis Pekinensis*, Vol. 51 No. 2, pp. 349-356.

Song, M., Heo, G.E. and Kim, S.Y. (2014), "Analyzing topic evolution in bioinformatics: investigation of dynamics of the field with conference data in DBLP", *Scientometrics*, Vol. 101 No. 1, pp. 397-428, doi: 10.1007/s11192-014-1246-2.

Stenetorp, P., Pyysalo, S. and Topić, G. (2012), "Brat rapid annotation tool", available at: https://brat.nlplab.org/.

Takase, S., Okazaki, N. and Inui, K. (2016), "Modeling semantic compositionality of relational patterns", *Engineering Applications of Artificial Intelligence*, Vol. 50, pp. 256-264, doi: 10.1016/j.engappai.2016.01.027.

Uzuner, O., Mailoa, J., Ryan, R. and Sibanda, T. (2010), "Semantic relations for problem-oriented medical records", *Artificial Intelligence in Medicine*, Vol. 50 No. 2, pp. 63-73, doi: 10.1016/j.artmed.2010.05.006.

Wang, X. and Gu, Y. (2017), "Cross-label suppression: a discriminative and fast dictionary learning with group regularization", *IEEE Transactions on Image Processing*, Vol. 26 No. 8, pp. 3859-3873, doi: 10.1109/TIP.2017.2703101.

Wang, L., Cao, Z., De Melo, G. and Liu, Z. (2016), "Relation classification via multi-level attention cnns", in Erk, K. and Smith, N. (Eds), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 7-12 August, 2016*, Association for Computational Linguistics, Berlin, Germany, Stroudsburg, PA, pp. 1298-1307.

Wang, J., Fan, Y., Zhang, H. and Feng, L. (2021), "Technology hotspot tracking: topic discovery and evolution of China's blockchain patents based on a dynamic LDA model", *Symmetry*, Vol. 13 No. 3, p. 415, doi: 10.3390/sym13030415.

Wu, F., Qian, L. and Huang, L. (2014), "The method of identifying the application field of technology based on the SAO structure of patents", *Science Research Management*, Vol. 35 No. 6, pp. 1-7.

Wu, Q., Kuang, Y., Hong, Q. and She, Y. (2019), "Frontier knowledge discovery and visualization in cancer field based on KOS and LDA", *Scientometrics*, Vol. 118 No. 3, pp. 979-1010, doi: 10.1007/s11192-018-2989-y.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R. and Le, Q.V. (2019), "Xlnet: generalized autoregressive pretraining for language understanding", in Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. and Garnett, R. (Eds), *Neural Information Processing Systems*, Vancouver Convention Center, Vancouver Canada, Neural Information Processing Systems (NIPS), La Jolla, CA, Vol. 32, pp. 3088-3099.

Yoon, B., Kim, S., Kim, S. and Seol, H. (2021), "Doc2vec-based link prediction approach using SAO structures: application to patent network", *Scientometrics*, pp. 1-30, doi: 10.1007/s11192-021-04187-4.

Yu, D., Xu, Z. and Wang, X. (2020), "Bibliometric analysis of support vector machines research trend: a case study in China", *International Journal of Machine Learning and Cybernetics*, Vol. 11 No. 3, pp. 715-728, doi: 10.1007/s13042-019-01028-y.

Zhang, J. and Yu, W. (2020), "Early detection of technology opportunity based on analogy design and phrase semantic representation", *Scientometrics*, Vol. 125 No. 1, pp. 551-576, doi: 10.1007/s11192-020-03641-z.

Zhang, Y., Zhang, G., Zhu, D. and Lu, J. (2017), "Scientific evolutionary pathways: identifying and visualizing relationships for scientific topics", *Journal of the Association for Information Science and Technology*, Vol. 68 No. 8, pp. 1925-1939, doi: 10.1002/asi.23814.

Zhang, X., Li, P., Jia, W. and Zhao, H. (2019), "Multi-labeled relation extraction with attentive capsule network", *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI, Palo Alto, CA, Hawaii, USA, January 27-February 1, 2019, Vol. 33, pp. 7484-7491, doi: 10.1609/aaai.v33i01.33017484.

Zhou, G., Su, J., Zhang, J. and Zhang, M. (2005), "Exploring various knowledge in relation extraction", *Proceedings of the 43rd annual meeting of the association for computational linguistics (acl'05)*, Association for Computational Linguistics, Stroudsburg, PA, Ann Arbor, pp. 427-434.

Corresponding author
Jinzhu Zhang can be contacted at: zhangjinzhu@njust.edu.cn