Technological Forecasting & Social Change 146 (2019) 339–351

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

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

A deep learning methodology for automatic extraction and discovery of technical intelligence

Jianguo Xu, Lixiang Guo, Jiang Jiang*, Bingfeng Ge, Mengjun Li

College of Systems Engineering, National University of Defense Technology, Changsha, Hunan 410073, China

ARTICLE INFO

Keywords:
Technical intelligence
CRF-BiLSTM
Deep learning
Intelligence monitoring

ABSTRACT

It is imperative and arduous to acquire product and business intelligence of global technical market. In this paper, a deep learning methodology is proposed to automatically extract and discover vital technical information from large-scale news dataset. More specifically, six kinds of technical elements are first defined to provide the concrete syntax information. Next, the CRF-BiLSTM approach is used to automatically extract technical entities, in which a conditional random field (CRF) layer is added on top of bidirectional long short-term memory (BiLSTM) layer. Then, three indicators including timeliness, influence and innovativeness are designed to evaluate the value of intelligence comprehensively. Finally, as a case study, technical news on three military-related websites is utilized to illustrate the efficiency and effectiveness of the foregoing methodology with the result of 80.82 (F-score) in comparison to four other models. In more detail, data on unmanned systems are extracted to summarize the state-of-the-art, and track up-to-the-minute innovations and developments in this field.

1. Introduction

Science and technology (S&T) has emerged from simple to complex over the past century. New initiatives such as Industry 4.0, Reindustrialization, Smart Manufacturing and Super Smart Society 5.0 are proposed in succession, and thus the size of technologies is rapidly expanding. Meanwhile, technologies are no longer independent individuals, but networked with each other in regard to their relationships of coordination and interdependence. In military, technical Intelligence (TECHINT) is intelligence about weapons and equipment used by foreign nations. In a broader social context, TECHINT refers to all technical information to promote the awareness of technical threats and opportunities in global (Ma et al., 2017). To keep or gain competitive edge in the business, it is always necessary for a company to take proactive measures to detect, track and predict emerging technologies or innovative combination of currently-existing technologies (Clauset et al., 2017). The companies should be aware of the products and business information of global technical market. Such intelligence is of great significance for them to expand market, promote new products and develop new areas (Huang et al., 2016).

Typically, the TECHINT process is divided into collection, exploitation and production. There are two basic tasks to exploit TECHINT. On one hand, it tries to answer existing questions with available intelligence. On the other hand, it is more challenging to answer questions by capturing new information rapidly. Moreover, the collected data should be exploited into sorted items rather than raw data as the duty of manager is decision-making but not data analysis. In general, digital publications, reports, news and other various types of resources are available to extract valuable technical information (Henry and Mcinnes, 2017). Especially, many companies have already released news which contains business information which can be tracked by other competitors to exhibit their products and services. Nevertheless, the existing large news data is completely beyond the capability of individuals and lead to information overload which is one of the main reasons to make mistakes.

Basic methods for extracting information of TECHINT contain bibliographic retrieval and bibliometrics, which follow the same routine: calculating the amount and growth speed of publications, creating the technical growth curve and determining the state of technology. More sophisticated method incorporate complex network. In recent years, natural language processing (NLP) with domain-specific ontologies are widely applied in sociology and informatics. However, there are three limitations of the existing methods. First, they suffer from the weak efficiency because of the long duration from writing to publishing. Second, network-based method is effective when processing structured data. So, it is not suitable for valuable text and news data on the Internet. Last, traditional NLP methods should follow the grammar structure of text and only extract general semantics information instead

[*] Corresponding author.
E-mail address: jiangjiangnudt@nudt.edu.cn (J. Jiang).

https://doi.org/10.1016/j.techfore.2019.06.004
Received 27 December 2018; Received in revised form 9 May 2019; Accepted 5 June 2019

of the professional technical information.

To the best of our knowledge, few researches have studied the aforementioned limitations profoundly. We introduce the state-of-the-art of the current NLP-based research and propose a new deep learning methodology for automatic discovery of technical intelligence from news. The main contributions of this paper are summarized as follows:

- We propose an overall methodology to automatically extract technical intelligence by combining text analysis, complex network and deep learning.
- According to the nature of technology, we design a set of tags to describe technology and develop a new dataset from Chinese news.
- We design three indicators including timeliness, influence and innovativeness to measure the value of technical elements and technical news.
- Experimental study on Chinese news is carried out in detail. As a result, numerous advanced products and meaningful news have been discovered.

The rest of this paper is organized as follows. After this brief introduction, we follow with the literature review in Section 2. The overall research process and detailed procedure are described in Section 3. Section 4 illustrates the proposed methodology using three military-related websites. Section 5 and Section 6 present the discussion and conclusions.

2. Literature review

As aforementioned, bibliometrics, complex network and NLP are commonly used for mining valuable technical intelligence. In the bibliometrics method, the growth curve and temporal development of publications have been examined to detect influential research forefronts (Yoon et al., 2014). Numerous bibliometrics indicators have been designed to measure emerging technologies (Rotolo et al., 2015), including radical novelty, relatively fast growth, coherence, prominent impact and uncertainty (Wang, 2017). In addition to papers, topic terms have been used to quantitatively show S&T trends, in which growth curves fit the trend of amount development of selected journal articles. Moreover, keyword-based model was proposed to capture technical flows and emerging patterns(Joung and Kim, 2017). In biomedicine, the Medical Subject Headings (MeSHs) (Leydesdorff et al., 2016) are available as an alternative to keywords and they are updated every year with new terms. So, they are always used to detect emerging topics. However, it is a difficult task even for experts to daily define keywords or key phrase patterns of in new or emerging technology areas. Furthermore, keywords based methods ignore relationships between topics. Association rule (Yu et al., 2016) is another statistical model to determine the likelihood of a relationship existing between terms. Once term co-occurrences are found (Yoon et al., 2011), two statistics are computed for each pair of linked terms, confidence and support. Confidence estimates the percentage of articles containing the linking term that also contains the starting term, and support estimates the count of articles containing both the starting and linking term. Then, a threshold is applied to remove low likelihood relationships.

A collection of conference papers, journal articles, and patents can be viewed as complex networks. Co-occurrences of terms, citing and co-citation can help to construct the technical network (Small et al., 2017). Citation dynamics have already started in 1960s, when Price proposed the cumulative advantage (CA) model to explain the power-law distributions (Fortunato et al., 2018). Many factors have been incorporated into network model, such as the disappearance of nodes, and the aging of the nodes. Pre-existing network structures have strong predictive power to future innovations in innovation network with 1.8 million patents (Acemoglu et al., 2016). In the rapidly growing fields, chronological networks of conference sessions could be used to visualize the scientific and technical streams formed by the session sequences (Furukawa et al., 2015). Unlike homogeneous information networks, heterogeneous information networks (Sebastian, 2017) can encode richer information and capture deeper semantics between various objects. That information can then be used to capture and model the previously unknown associations between research papers. Although complex network is a powerful and sophisticated tool to obtain valuable information, it also takes a certain amount of time to index cited articles in databases after article were published. This time-delay peculiarity of network based methods would be a disadvantage in an investigation of rapidly changing and growing research fields. Most importantly, it is impossible to recognize new concepts that have no correspondence to valid and well-known terms and will fail to discover new information.

Textual analysis, an alternative method for investigating scientific publications, enables researchers to unveil implicit mutual links between publications. Besides direct connections, other invisible and obscure connections can be discovered in textual analysis. In the earlier text-mining technique, trends in term or phrase frequencies in the articles were used to extract the notable research topics expected in relation to emerging technologies in the fields. Technical elements can also serve as the basis for identifying technical opportunities and extracting technical intelligence. However, the uniform definition of technology is absent. Another representing method of technical elements is based on ontology, which is mainly established through the general concepts of the domain. The domain knowledge base developed by ontology has a clear hierarchy structure, rich contents and logical reasoning capability. Latent Dirichlet Allocation (LDA) provides an even more statistically sophisticated technique for the conceptualization and identification of scientific topics (Xie et al., 2013). Furthermore, subject-action-object (SAO) approach(Yang et al., 2017) was designed for technical components retrieval. Function–attribute analysis extracts properties and functions automatically from patents using NLP. Regarding properties and functions as nodes, and co-occurrences as links, an invention property-function network (IPFN) is generated (Yoon et al., 2011). The property and function are mainly described by adjectives and verbs respectively. The engineering design model of reference we are interested in is the so-called Function–Behavior–State (FBS) model (Fantoni et al., 2013). These methods perform well in semantic description. However, due to the ambiguity of semantics and complexity of technology relationships, both of the methods cannot get precise enough results. With the absence of global information, the element captured may not be able to express the global structured semantics.

Deep learning is rapidly advanced in recent years and get state-of-the-art performance within various fields (Habibi et al., 2017). Deep learning can learn complex features by combining simple features learned from data. It takes advantage of large datasets and computationally-efficient training algorithms to outperform other approaches at various machine learning tasks. It is obviously difficult to establish a mathematical model to describe technology. Therefore, it can be treated as a named entity recognition (NER) problem which is a challenging in NLP (Gridach, 2017). There is only a small amount of labeled data for training. Meanwhile, the diversity of words representing named entity makes it difficult for conventional model to identify which words are named entities in a sentence. Neural network has been developing very fast and applied in many fields. In NLP, neural network such as CNN and RNN can automatically capture the semantics feature and better express the sentence (Derczynski et al., 2015). For NER, in recent years, some researchers format it as a sequence labeling problem including location, person, time, songs, movies and brand of products. A typical model, CRF-BiLSTM (Huang et al., 2015), employs bidirectional LSTM to represent the context feature of word in a sentence (Le et al., 2018; Zhang et al., 2016). The CRF, a traditional machine learning method, is then added to catch the dependency among entities. At last, the label sequence is output from the upper layer CRF (Ma, 2016).

The technical intelligence for business requires relatively high quality data without sharply rise of extracting cost. The data sources should be strongly associated with the research domain while collecting Internet information. Technical data sources such as papers, patents, reports, funds and news contain a large quantity of commercial information (Grassano et al., 2016). Papers and patents are resources which are authorized and closely related while funds data is authorized but small in amount. It distributes everywhere on the Internet which makes it hard to be collected (You et al., 2017). Over all, news data have a large quantity and strong efficiency and is easy to capture, which makes it a rather good information source for data analysis in real time (Vossen et al., 2016). For innovative products, most stakeholders would like to propagate their products and services in news as early as possible for attracting attentions and funds. It is helpful to leverage technical news for automatic real-time entity extraction.

3. Proposed methodology

This paper proposes a deep learning based methodology to extract core technical elements and discover valuable technical intelligence on the Internet. The basic idea is to take pre-determined data sources as seeds and discover more unknowns through the CRF-BiLSTM approach. These data sources include not only general news but also reports of specific disciplines. The steps and details of the methodology, as shown in Fig. 1, consist of five components: (1) Six kinds of technical elements are defined to provide the semantic information. (2) Technical news is crawled from the Internet, and linguistic pre-processing is carried out in the order of data cleaning, word embeddings building, and data tagging. (3) The CRF-BiLSTM approach is used to train the structural and semantic information model with tagged dataset and extract new elements with the complete dataset. For the task of technical entities extraction, we model the problem as a sequence labeling one. The BiLSTM can extract the hidden features of a sequence. On the top of BiLSTM, a conditional random field (CRF) layer is added to catch the dependencies among the entity labels. The final outputs are label sequences of sentence. (4) Extracted technical elements and original information are assessed with three indicators for new and high-value intelligence discovery. (5) The data visualization method is used to display the comprehensive situation.

3.1. Description of technology

A set of concepts of technology are usually defined first to provide the formal/well-defined semantics, which allow machines to interpret in an automated manner. Technical elements should be grouped into the six semantically complete interrogatives (i.e., WHO, WHERE, WHAT, WHEN, WHY, and HOW (5W1H)) as data taxonomies at the highest level to ensure consistency (Porter et al., 2015). It is an important characteristic of technology to propel economic growth and long-term well-being. At the same time, technologies are generated and implemented by organizations and individuals to cope with requirements of customers. This is the principal evolutionary rule of technical innovation diffusion. Therefore, technology can be described as method, system, or device (WHAT) out of scientific knowledge (HOW) by concrete performers (WHO) and for practical purposes (WHY). Moreover, location (WHERE) and time (WHEN) are of

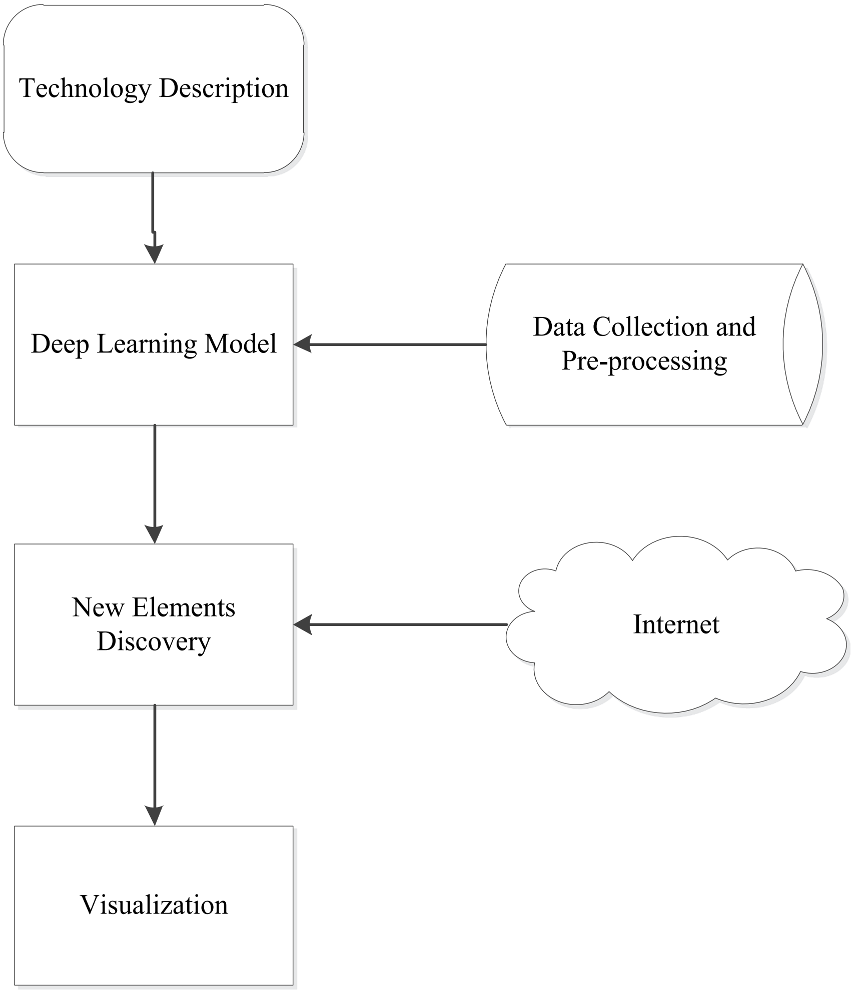


Fig. 1. The process of automatic extraction and discovery

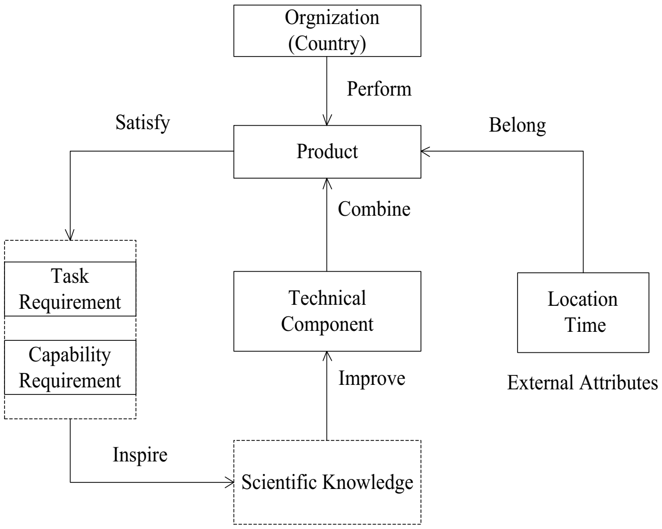


Fig. 2. Description model of technology.

importance to manage and monitor technologies. The Fig. 2 describes elements and relationships among them. Compared with external attributes, product (WHAT) that concretes the nature of technology plays an important role in the business. The companies or other organizations not only satisfy the requirements of customers through products, but also make profits through them. The focus is on the constituent system capabilities and marketing opportunities. Complex products can be decomposed into components or sub-systems. Conversely, requirement also pulls the development of S&T, and finally improves the performance of products. Finally, it is necessary to distinguish between technology and scientific knowledge. Despite not all technologies depend on disruptive advances in science, scientific progress helps to improve performance of products.

Technical elements serve as a template for organizing data in a machine readable format. The data in a described architecture is decomposed into concepts, associations, and attributes. Requirement is the logical expression of the need to transfer information towards nodes. It is always a desired effect under specified standards and conditions through combination of means and ways to perform a set of tasks. Thus, the description of requirement consists of not only what want to be done but also what can be done, which can be generally explained as task requirement and capability requirement. Task requirement can be divided into specific executive activities like operation under high risk and supervision, sensors identifying trees, following soldiers marching, cleansing enemy in buildings and checking improvised explosive devices. Similarly, capability requirement capture the enterprise goals associated with the overall vision, such as controlling several unmanned platforms, exploring in limited space and self-controlling.

Product involves a method, system or device that is manufactured or refined for sale, which includes not only substantive products such as Throwbot XT, Global hawk and First look, but also some technical standards, specifications and methods, like the specification for interoperability of unmanned vehicles. More importantly, service is a special and invisible product. The department of defense (DoD) define solutions specifically in terms of services that can be discovered, subscribed to, and utilized, as appropriate, in executing departmental or joint functions and requirements. The more advanced the product is, the more function customization and after-sale guarantees are emphasized.

A component is a relatively independent part of product and is characterized by its responsibilities, or the interfaces it offers. Components can be decomposed into smaller components or aggregated into larger one. A component can be a collection of classes, a program, a hardware device with embedded functional characteristics or any organized assembly of resources and procedures to accomplish a set of specific functions. It varies a lot, like microphone, composite material, night camera, and laser designator. Technical management processes are applied to allow an orderly progression from one level of development to the next detailed level using controlled baselines. These processes are used for the requirements, products.

Organization refers to an administrative or commercial structure that is managed to meet a requirement or to pursue collective goals. There are various legal organizations, including governments, international organizations, armed forces, universities et al. Furthermore, it includes both as large as countries and as small as research teams. To spot the importance of country, it is listed out as an individual element. Companies are the subject of the technical innovation who implement, promote, and use technologies. Universities and research institutes are the source and knowledge base of the technical innovation. In the military field, armed forces always apply the latest achievements directly.

Both steps and knowledge have self-similarity, which means knowledge can be disintegrated into knowledge units, and can even be further dissolved. Any product and system are the scientific expression of knowledge. Two operations are introduced to simplify the description of technology. On the one hand, we regard the releasing time of the news as the time of technology. On the other hand, we ignore the scientific knowledge as there are too much more details to model. The examples are described in Table 1.

3.2. Data collecting and pre-processing

Once the technical field has been selected, technologies of interest could be collected from the Internet or other database. There are different methods for data collecting according to data types. Papers and patents can be directly downloaded from paper data base like Web of Science (WoS) and patent data base like the United States Patent and Trademark Office (USPTO). Even more, the open source datasets like AMiner also help. For technical reports, they basically need to be collected artificially or crawled from professional sites. However, the data source mentioned above are either hard to collect or poorly efficient. Delightedly, news data and comments data which contain the latest technology developing information are able to make up those drawbacks. Nevertheless, technical news comes out only after professional journalists' edition which guarantees a higher facticity and value density than comment data, since people with different knowledge level may comment technologies without taking responsibility. Also, there are professional technical comments data like MIT's Technology Review. However, it is hard to reach the level of big data. Instead, there are many professional (e.g., UAV bulletin, semiconductor web site) and integrated (e.g., Sina Tech, Google tech news, Tech Daily) technical news websites worldwide. Thus, this article chooses integrated technical news as the primitive data for further analysis.

The element extraction process consists of a few fundamental steps, including data cleaning, word embeddings building, and data tagging. All news is converted into a simple uniform format to ease further processing of data cleaning. In this format, all texts and annotations are stored in one single file. Several points should be taken into consideration in the process of data cleaning. First, news data crawled from the Internet includes a large quantity of advertisements and website links, which should be cleaned out. Second, there is no natural

Table 1
Technical elements list.

Element type	Example	Illustration
Requirement	Autonomous navigation	Including task and capability requirements.
Product	First look, Global Hawk	Including methods, systems devices and services.
Component	Solar cell	It can be regard as a part of product.
Organization	Harvard University	Research subjects
Time	08/09/2012	The releasing time
Country	American, China	A special kind of organization

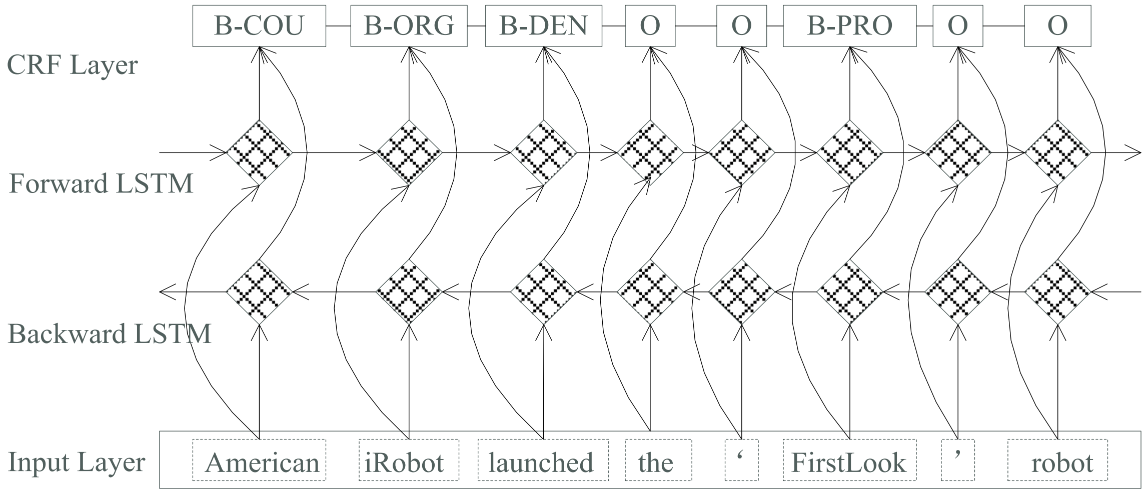


Fig. 3. The process of technical elements extraction.

participle construction for Chinese corpus so that professional text segmentation software is needed.

Word embedding techniques are utilized to capture the semantics of words (and their similarities) based on other words surrounding them. We learn a sparse word representation by sending the pre-trained word embedding through an auto-encoder layer. Genism toolkit[1] was used to train general word embeddings from the text mentioned above. The general word embeddings are trained on an extensive corpus (Le et al., 2018).

This data is converted to standard IOB format where every entity begins with the tag B- followed by I- if it is a multiword named entity. There are six types of tags including requirement, product, component, organization, country, and time. The original sentence begins with BOS and ends with EOS. Meanwhile, the entire sentence ends with NOTHING. For instance, BOS In early October 2009 , Reconnaissance Robot Company of the United States launched a new ReconScout Robot , which is equipped with high-power batteries and has good cross-country mobility . EOS O O B-time I-time O B-organization I-organization I-organization O O B-country I-country O O O B-product I-product O O O O O B-component I-Component O O B-requirement I-requirement I-requirement O O NOTHING.

3.3. Technical elements extraction

While dataset that has been tagged, elements extraction could be viewed as a typically named entity recognition (NER) problem. In this part, we give a technical process to learn and extract entities with CRF-BiLSTM approach which is used in Medicine and Sociology (Hogenboom et al., 2016). Specifically, there is a CRF layer on the top of BiLSTM layer to get the technical tag sequence. Fig. 3 shows the process of technical elements extraction.

Given a sequence of technical elements, $D = (X, Y)$, where $X = (x_1, x_2, \cdots, x_n)$is the actual sequence of a sentence with n words and $Y = (y_1, y_2, \cdots, y_n)$is the corresponding tags. The task is to find the most possible tag sequence t^* for X,

$$\underset{t^*}{\operatorname{argmax}}\, P(t \mid X).\ .$$

The parameters are computed from the conditional probability $P(Y|X)$. We assume the sequential feature of X is expressed with LSTM. The task is decomposed as

$$P(Y \mid X) = \prod_{i=1}^{n} P(y_i \mid g(X)),$$

where function g is a pre-forward LSTM which learns relevant feature by looking at current word and all the previous words. LSTM networks are similar to recurrent neural network (RNN), except that the hidden layer updates are replaced by purpose-built memory cells which are designed to store history information. LSTM memory cell could take advantage of long range relations between technical elements. The cell is distributed as follows.

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f)$$

$$c_t = f_t c_{t-1} + i_t\, tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_i)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o)$$

$$h_t = o_t \tanh(c_t)$$

where σ is sigmoid activation function, and i, f, o, c refer to the input gate, forget gate, output gate and cell vectors. W denotes matrix between two layers. For instance, W_{xo} is the input-output gate matrix.

In reality, the semantics of each word is influenced both the previous words and the following words. The above modeling does not take into account the information from the right context. It is preferred to capture bidirectional information with two opposite learners. We adjust pre-forward LSTM into

$$P(Y \mid X) = \prod_{i=1}^{n} P(y_i \mid g(X), h(X))$$

where h is backward LSTM.

Note that the inputs and outputs of BiLSTM are connected with memory cells. Another model called Conditional Random Fields (CRF) focuses on sequence tagging in sentence level and always produces high accuracy. It is discriminative undirected probabilistic graphical model. We combine a bidirectional LSTM network and a CRF network to form a CRF-BiLSTM network. Then the network uses not only directional information but sentence level information.

3.4. Automatic discovery and visualization

Technical elements are collected from the recognized entities in the last part. As new products are springing up, intelligence needs to be sorted out within a large number of new products or other new elements. They must then be refined to obtain high value intelligence. In general, there are three principles that measure the value of the technical elements as shown in Algorithm 1. The first principle is the timeliness. It may take the form of achieving a new or a changed requirement and the form of introducing new

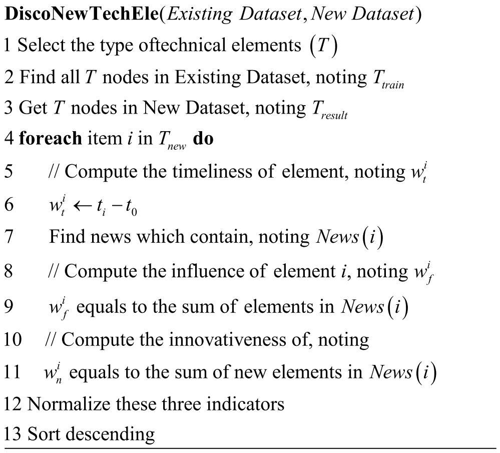
[1] https://radimrehurek.com/gensim/.

products. The sooner you obtain useful information, the earlier you have the initiative to make decisions. New technical element refers to firstly recognized entity in test dataset. These elements are mostly domain-specific and time-sensitive. For example, 'dragon runner' is a small unmanned ground robot. It is neither an athlete nor a game. The new organizations represent potential competitors or partner teammates. New demands represent new market opportunities.

The second is the influence of the technical element. Theoretically, technology exerts a prominent impact on specific domains by changing the composition of components, patterns of interactions among those, and the associated knowledge production processes. Technical elements not only can influence academic circle but the lives of ordinary people. The news has the advantage to understand the views of society. The more frequently it is reported, the more influential the information is. Consequently, it benefits from a wide range of applications.

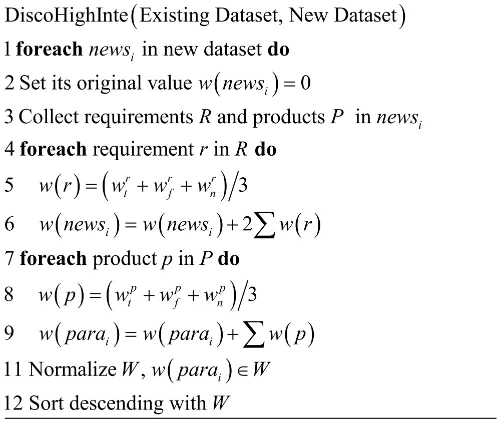
Finally, the innovativeness of the technical element should be taken into account. Innovativeness is not only a characteristic of new product but may also be generated by putting an existing one to a new application. Technical revolution always brings about an immense amount of innovative technologies. Innovative technology may adapt and then emerge as well as potentially invading other domains. Relatively amount of new elements can be exploited to measure the innovativeness. The more new elements it relates to, the greater value it owns.

Algorithm 1. Discover valuable technical elements.



```
DiscoNewTechEle(Existing Dataset, New Dataset)
Select the type oftechnical elements $(T)$
Find all $T$ nodes in Existing Dataset, noting $T_{train}$
Get $T$ nodes in New Dataset, noting $T_{result}$
foreach item $i$ in $T_{new}$ do
    // Compute the timeliness of element, noting $w_t^i$
    $w_t^i \leftarrow t_i - t_0$
    Find news which contain, noting $News(i)$
    // Compute the influence of element $i$, noting $w_f^i$
    $w_f^i$ equals to the sum of elements in $News(i)$
    // Compute the innovativeness of, noting
    $w_n^i$ equals to the sum of new elements in $News(i)$
Normalize these three indicators
Sort descending
```

Algorithm 2. Discover high-value intelligence.



```
DiscoHighInte(Existing Dataset, New Dataset)
foreach $news_i$ in new dataset do
Set its original value $w(news_i) = 0$
Collect requirements $R$ and products $P$ in $news_i$
foreach requirement $r$ in $R$ do
    $w(r) = (w_t^r + w_f^r + w_n^r)/3$
    $w(news_i) = w(news_i) + 2\sum w(r)$
foreach product $p$ in $P$ do
    $w(p) = (w_t^p + w_f^p + w_n^p)/3$
    $w(para_i) = w(para_i) + \sum w(p)$
Normalize $W$, $w(para_i) \in W$
Sort descending with $W$
```

Algorithm 1 calculates the value of each element. The value of each piece of news is weighed by Algorithm 2. As far as specific intelligence is concerned, we should especially focus on requirements and products. Requirement acquisition and analysis is the first step for the development of complex system. The requirement-centered view holds that it is the most valuable when a new product satisfies a new requirement, followed by an old product solving a new requirement; the next is that a new product solves an old requirement; and finally an old product solves an old requirement.

It is easier to understand the technical discovery when it is visualized in two-dimensional space rather than summarized just in tables. To give meaning to these technical elements, they must be represented and displayed graphically. In the following, data visualization and tools are used to display technical intelligence with heterogeneous information. The recognized technical elements are used as nodes, and the relationship of co-occurrence between technical elements in the same sentence is taken as edge. The graphical expression of innovative information can be obtained. These nodes and links are turned into graphical representations. There are lots of tools for creating/deleting nodes and links, displaying information in different formats, and helping to analyze the data. Based on this graph, practical information can be derived. In this paper, we take the Gephi,[2] an open source software for the visual exploration of networks, as visualization tool (Heymann and Grand, 2013)

4. Experiments

4.1. Experimental setup

First, we begin the section in describing the datasets and various

[2] https://gephi.org/.

Table 2
Data sources.

Web	Address	Time	Data size	Fields
Sohu	www.sohu.com	2018.9.6	59,343	Time, title, text
Xilu	www.xilu.com	2018.9.6	34,842	Time, title, text
Sina	www.sina.com.cn	2018.9.6	9062	Time, title, text

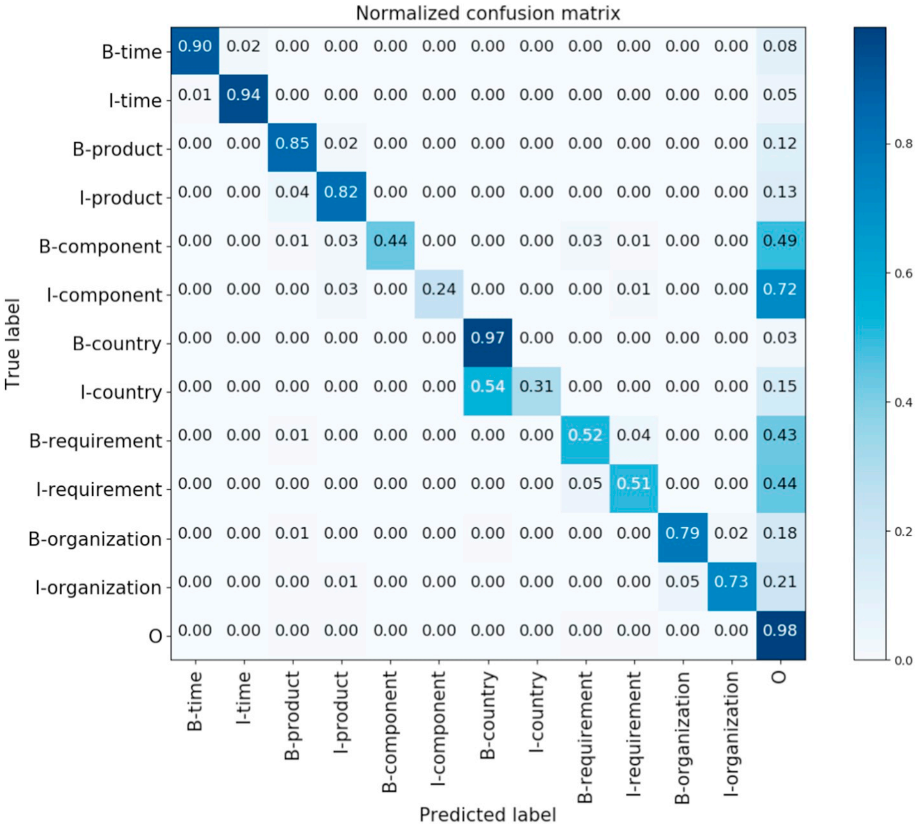
hyper-parameters of our experiments. The hyper-parameters are chosen by doing a random grid search and evaluation on the development set. We adopt the network spider to crawl web pages from three military-related news websites in Table 2. These websites are devoted to report on global political, economic and military, collect rich and authoritative information, and publish constructive analysis and comments on military products. Therefore, it is promising to use these data and information to discover valuable technical elements. Then data have been cleaned by deleting ads, web links and other invalid characters and leaving time, title and text. The Jieba is used for segmentation and python natural language toolkit package was used to grasp data. Finally, there are 106,230 sentences from May 7, 2016 to September 6, 2018 which are divided into two dataset, existing dataset and new dataset. There are 6020 tagged items before September 2017. The remaining 100,210 items are used to extract valuable elements. Five-fold crossover is taken to evaluate the experimental result. That is, the first dataset is divided into five parts, one takes as test set, and the other four as training set. Repeat the experiment 100 times per fold. Take an average of 100 times as the final result. We employ continuous bag of words (CBOW) language model to train the domain-specific word embeddings of 300-D. For the basic LSTM cell, we set the number of units as 100 and the maximum sequence length as 200. During the training stage, the batch size is set as 64, epoch number as 100.

Table 4
Performance indicators of different methods.

Method	Accuracy	Recall	F-score
CRF	0.7816	0.4940	0.6054
LSTM	0.6859	0.4584	0.5496
Bi-LSTM	0.8721	0.6170	0.7227
CRF-BiLSTM	0.9064	0.7291	0.8082
Joint-RNN	0.7762	0.5798	0.6638

Table 3
Performance of technical elements recognition.

	Requirement	Product	Component	Organization	Country	Time	F-score
1	0.5862	0.8533	0.4313	0.8366	0.9807	0.9344	0.8108
2	0.5625	0.8495	0.5249	0.7968	0.9638	0.9576	0.7925
3	0.5937	0.8577	0.5079	0.7886	0.9651	0.9428	0.8124
4	0.5838	0.8694	0.3103	0.8081	0.9589	0.9590	0.8134
5	0.5805	0.8599	0.6545	0.8034	0.9655	0.8945	0.8120
Avg.	0.5813	0.8579	0.4857	0.8067	0.9668	0.9377	0.8082



	B-time	I-time	B-product	I-product	B-component	I-component	B-country	I-country	B-requirement	I-requirement	B-organization	I-organization	O
B-time	0.90	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08
I-time	0.01	0.94	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05
B-product	0.00	0.00	0.85	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.12
I-product	0.00	0.00	0.04	0.82	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.13
B-component	0.00	0.00	0.01	0.03	0.44	0.00	0.00	0.00	0.03	0.01	0.00	0.00	0.49
I-component	0.00	0.00	0.00	0.03	0.00	0.24	0.00	0.00	0.00	0.01	0.00	0.00	0.72
B-country	0.00	0.00	0.00	0.00	0.00	0.00	0.97	0.00	0.00	0.00	0.00	0.00	0.03
I-country	0.00	0.00	0.00	0.00	0.00	0.00	0.54	0.31	0.00	0.00	0.00	0.00	0.15
B-requirement	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.52	0.04	0.00	0.00	0.43
I-requirement	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.51	0.00	0.00	0.44
B-organization	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.79	0.02	0.18
I-organization	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.73	0.21
O	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.98

Fig. 4. Normalized confusion matrix.

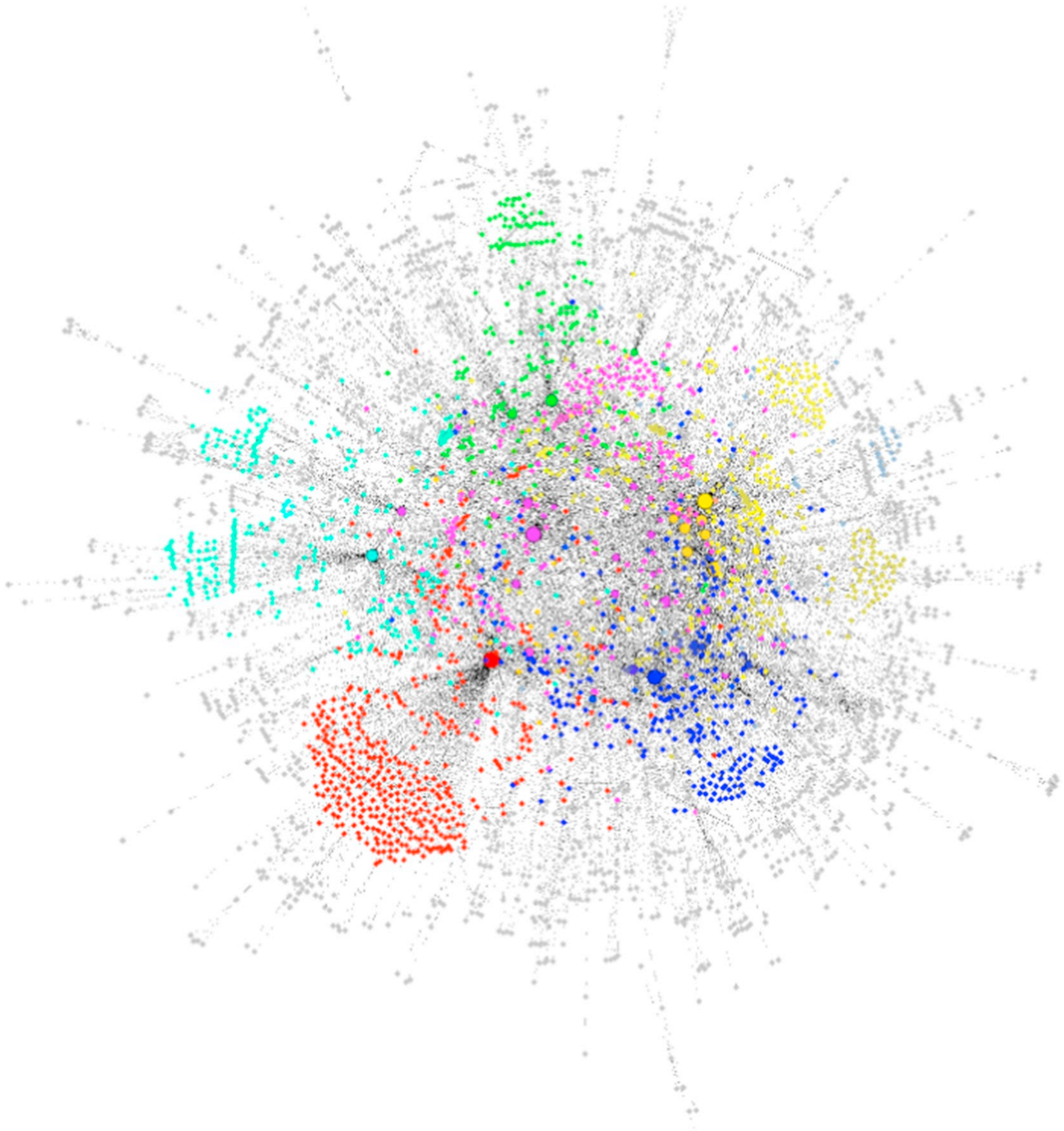


Fig. 5. Overall visualization of the three websites.

Table 5
Top 10 influential technical elements.

Name	Type	Influence	Illustration
Navy	Organization	1	The Chinese navy
J-20	Product	0.89	Five generation fighter
Stealth	Requirement	0.88	Requirement of fighters, UAVs and helicopters.
Air force	Organization	0.82	The Chinese air force
US army	Organization	0.79	
F-22	Product	0.75	The four generation fighter
F-35	Product	0.75	The five generation fighter
J-16	Product	0.74	Three generation fighter
Su-35	Product	0.68	The four generation fighter
interception	Requirement	0.68	Intercepting far, near and near heights

4.2. Model training

In this section we analyze the trained results. Table 3 shows the performance of the CRF-BiLSTM on the development set for the technical entities recognition. It is shown in Table 3 that the total F-score reaches 0.8134, which means that it is capable to identify most of the meaningful entities. The identifying accuracy of country and time is extraordinary high as around 95%, while that of product, organization and requirement decrease successively, which are 0.8694, 0.8366 and 0.5937 respectively. For component, its average accuracy is less than 0.5000 as 0.4857. The expression of nations and time is relatively simple, which leads to a lower identifying difficulty. The identification of components is the hardest and with low value density. The accuracy of product, requirement and organization should be concentrated on. Generally speaking, that is “who designed what to do what.” There are two main reasons which will infect the identification accuracy: One is the amount and quality of data label. The size of dataset used in this article is 6020 which is labeled by graduate students major in Management Science, so that the knowledge of UAV itself may not be very accurate. The other is that the difficulty of identifying is quite high because the expression of Chinese has many meanings and the products are usually named after animals as code, which will lead to a lot of confusions. Also, the representation of requirement and knowledge varies and any knowledge is possible to be involved in the unmanned system.

In the normalized confusion matrix in Fig. 4, the horizontal axis represents the predicted label while the vertical axis holds the true label. The elements at the diagonal of this matrix are the recall of the corresponding labels. Take B-time as an example. The amount of words

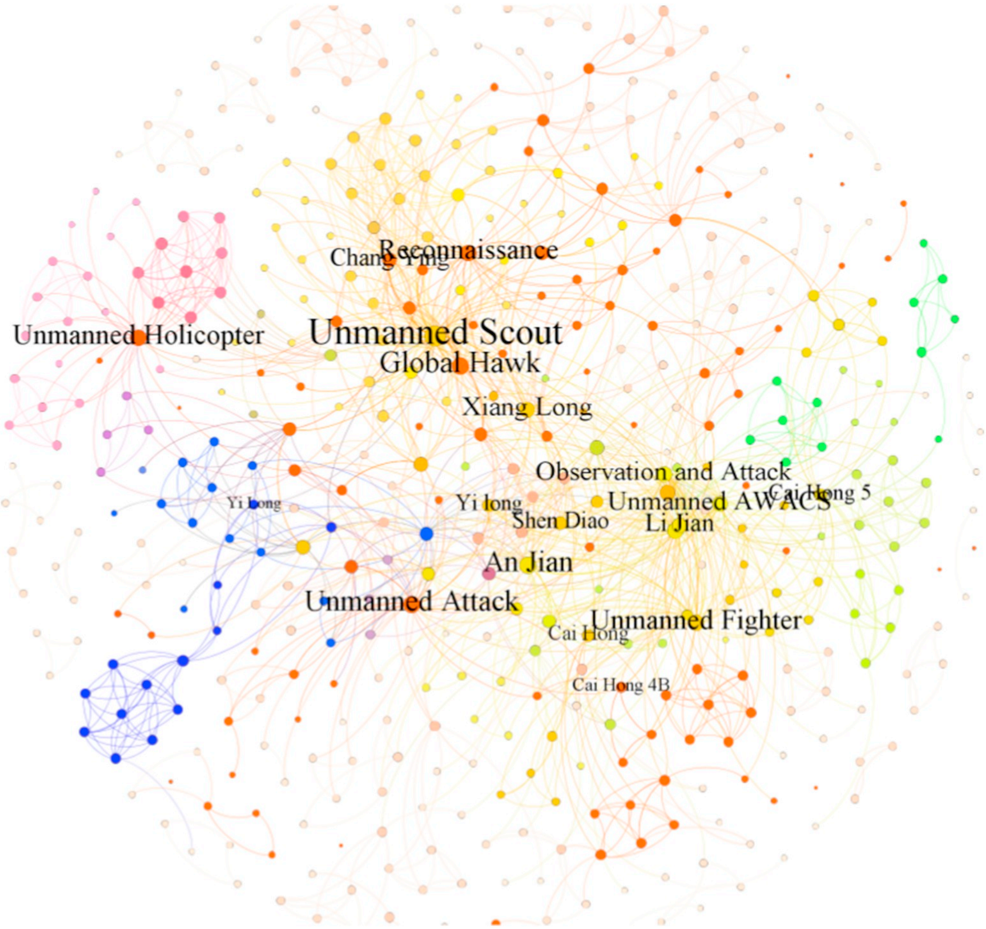


Fig. 6. Visualization effect about unmanned systems.

Table 6
Top 10 valuable products.

Num.	Innovativeness/weight	Influence/weight
1	AnJian/1	Global Hawk/1
2	CaiHong/0.94	AnJian/0.98
3	Global Hawk/0.87	CaiHong/0.96
4	JY-300/0.86	LiJian/0.96
5	LARK-1/0.86	YiLong/0.93
6	MQ-25/0.83	MQ-25/0.91
7	YiLong/0.81	X47-B/0.91
8	ShenDiao/0.81	ShenDiao/0.84
9	BaiLing-1/0.80	LARK-1/0.83
10	LiJian/0.78	XiangLong/0.83

Table 7
Top 10 valuable requirements.

Num.	Requirement	Weight
1	Supersonic flying	1
2	Maritime alert	0.95
3	Autonomous learning ability	0.9
4	To-ground precision strike capability	0.81
5	Autonomous aerial refueling verification	0.80
6	Dynamic task assignment	0.75
7	Support for trekking operations	0.75
8	Formation autonomy transformation	0.72
9	Stealth warning	0.69
10	Eliminate high value targets	0.69

labeled as B-time in the dataset is denoted as 1164. Then there are 0.9* 1164 words being labeled correctly, while 0.02* 1164 words and 0.08* 1164 words are wrongly labeled as I-time and O respectively.

We compare the results with four other models: pure CRF, LSTM, BiLSTM, CRF-BiLSTM, and Joint-RNN. The average performance indicators over all epochs in terms of precision, recall and F-score are provided in Table 4. On average, F-scores between them ranged from 0.5496 to 0.8082. F-score is 0.8082 for the CRF-BiLSTM, Minimum F-score is 54.96% for LSTM. Joint-RNN is better than CRF and LSTM but worse than BiLSTM. An interesting observation is that the Joint method often performs better than baseline algorithms.

4.3. Results of extraction

According to the previous trained model, 5660 technical elements are identified from 100,210 original records which are crawled from the Internet. Specifically, there are 2450 products, 1631 requirements, 1006 organizations and 573 components. All in all, the three websites tend to report on the development of science and technology of the Chinese Navy and Air Force as Fig. 5. Nodes represent six kinds of elements, while edges represent their co-occurrence relationships. The larger the node is, the stronger its influence is. The red nodes represent the relevant entities of the Chinese navy and the blue ones represent technical elements of the Chinese air force. In terms of products, these sites are more concerned about the currently heavily developed unmanned systems, fighters and helicopters respectively visualized with green, yellow and light blue. With respect to requirements, task

Table 8
Top 10 valuable news.

	Time	Source	Content	Illustration (elements)
1	2018-06-15	Souhu	The sixth generation fighter may cooperate with J20.	AnJian UAV, supersonic, super maneuverable, low observable, DSI inlet
2	2018-07-15	Xilu	Japanese destroyers will be equipped with us drones	MQ-8C unmanned helicopter, laser guided, operational management, used on destroyers.
3	2018-06-24	Xilu	China's patch radar is available for KJ-3000.	Patch radar, air early warning mission, KJ-3000
4	2018-07-30	Souhu	China Rainbow UAV refreshes its altitude record in take-off and landing	CH-5, plateau take-off and new special load.
5	2018-06-18	Xilu	A UAV can transport 13 soldiers' food.	Take-off service, clothing service, S.F. Express
6	2018-02-11	Xilu	China announces 4 stealth UAVs	CaiHong-805, XingYing, TianYing, LiJian
7	2018-05-30	Xilu	China's 56 unmanned vessels appear in the South China Sea	Formation maintenance, dynamic task assignment, formation autonomous transformation, cooperative obstacle avoidance and fault tolerant control
8	2018-02-1	Sina	US hypersonic unmanned reconnaissance aircraft is scheduled for trial flight in 2030	Hypersonic unmanned reconnaissance, hypersonic radio reconnaissance and hypersonic unmanned observation
9	2018-3-27	Xilu	Lockheed Martin unveiled concept map of unmanned refueling aircraft	MQ-25, conceptual map, aerial refueling mission.
10	2017-10-05	Souhu	China's "idea control weapon" may be developed successfully	idea control weapon

requirements such as stealth, penetration and interception have attracted wide attention. At the same time, China often compares their technologies with the United States throughout various fields of science and technology visualized with pink nodes. The most influential elements are shown in Table 5. The Chinese navy has developed rapidly in recent years with lots of new products in service. J-20, F-22, F-35, J-16, and Su-35 are widely concerned, and their normalized weights are 0.89, 0.75, 0.75, 0.74, and 0.68, respectively. Stealth demand is not only the need of fighters, but also the demand of stealth development of UAVs and helicopters.

Intelligent unmanned system is widely acknowledged to be one of the most dramatic technical game changers in this era. It is respected to have a disruptive impact on most walks of society and military. To build an intelligent battlefield, all countries are vigorously developing unmanned equipment and experiencing unmanned revolutionary migration, which has led to the reallocation of battlefield division.

To go a step further, the situation of unmanned systems is analyzed in detail. As a result, a total of 1505 unmanned technical elements and 809 unmanned systems are identified. It involves all operations such as reconnaissance, surveillance, command, decision and attack. It can be roughly divided into unmanned reconnaissance systems, unmanned combat systems, unmanned attack systems, unmanned AWACS and unmanned helicopters, as shown in Fig. 6.

The typical unmanned combat aircrafts such as the AnJian, the CaiHong and the Global Hawk still occupy the main layout of the news. YiLong, YunYing, ShenDiao, BaiLing and other unmanned equipment emerge in endlessly. It also involves unmanned spacecraft, unmanned submersible, unmanned warships and unmanned target drone. A total of 472 items were identified, including stealth, signal monitoring, autonomous recovery, joint operations. The China has successful in rapidly developing and fielding unmanned systems. Numerous breakthroughs have been made in unmanned reconnaissance and attack and explored the role of intelligent decision making within a joint operations.

According to the Algorithm 1, we get the innovativeness and influence rank of technical products respectively in Table 6. The AnJian UAV manufactured by the Shenyang Aircraft Design and Research Institute has supersonic, super-high maneuverability, low detectability and performs best in innovativeness. It is emphasized that the fuselage of the AnJian is quite different from the mainstream aircrafts. Meanwhile, the CaiHong UAV has been developing rapidly in recent years, and has continuously improved the launch of new models, covering multiple applications. The US Global Hawk is outstanding in terms of voyage, cruising time and flight altitude. As a result, it behaves well in both innovativeness and influence. At the same time, China has developed many kinds of unmanned products in the fields of strategic reconnaissance, early warning and so on. In terms of influence, the U. S. Global Hawk UAV ranks first, not only benefiting from its high technology and product stability, but also the actual combat which has been widely used. A number of UAVs of China follows, including the AnJian and CaiHong, which were widely reported and have a profound impact. On 30 August 2018, the U.S. Navy formally confirmed that Boeing's MQ-25 unmanned aerial vehicle program won the U.S. Navy's future carrier-based unmanned aerial vehicles. Its influence exceeds the original X-47B, which can greatly expand the scope of operation of manned / unmanned carrier aircraft. Although Da Jiang UAV occupies a large share of the civil market, it is seldom mentioned in military websites, and its influence is only 0.16.

Next, we analyze the requirements of the unmanned systems in the same way where the weight comes from the sum of innovativeness and influence in Table 7. Overall, 472 requirements were extracted, of which 190 were capacity requirements and 282 task requirements. Capacity requirements are mainly focused on supersonic, autonomous learning, precision strike, stealth, and high altitude long endurance. The task requirements mainly pay attention to early warning, reconnaissance, task allocation, formation transformation and air

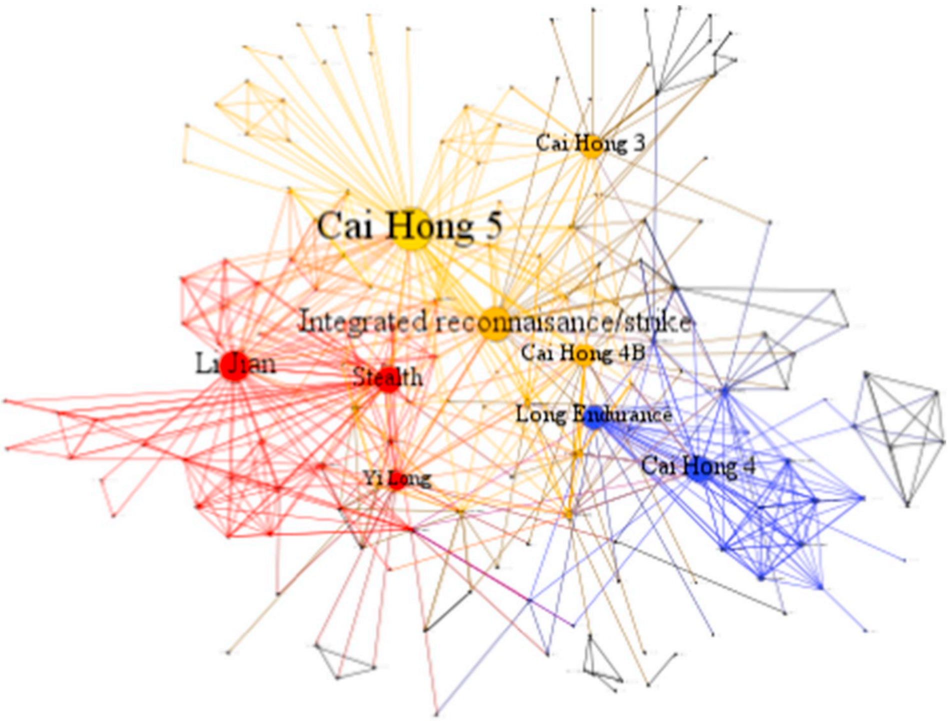


Fig. 7. Visualization result of CaiHong UAVs.

refueling. It can be seen that, in addition to the traditional reconnaissance and early warning, there are also emerged intelligent features such as autonomous transformation, autonomous distribution, dynamic decision-making, and supersonic high altitude and long flight.

In addition to gaining important market requirements and new products, we extract and track other technical elements. The discovery of new organizations can not only provide new cooperative information for field research, but also identify important competitors. Influential companies include the Boeing, the Israel Aviation Industry Corporation, the EADS Group, the Shenyang Institute of Automation and the China Aerospace Long March International Trade Co., Ltd.

Finally, we discover the most valuable news according to Algorithm 2 as shown in Table 8. The result involves many new types of UAVs all over the world, which can effectively contain all kinds of technical elements. On the one hand, the newer products and requirements are, the higher the ranking is. On the other hand, the greater the innovativeness and influence is, the greater the weight is. At the same time, S.F. Express provides battlefield take-out services, shipment services, and other new services with UAVs. Despite the Global Hawk has performed well in terms of innovativeness and influence, there have been few separate reports, and that is, most of the cases have been contrasted with Chinese products. It should be emphasized again that the above results may be different from the actual development of UAVs. The results are only responsible for data sources used in this article.

5. Discussion

The case in this paper employs the data of three websites which belongs to the same field and can supplement data integrity each other. But the actual situation may be more complicated, and analysts need to pay attention to dozens or even hundreds of data sources. Despite news has the advantages of timeliness and high value, it shows a mediocre difference. The elements extracted are mainly concentrated in the US army, the Chinese Navy, and the Chinese Air Force. Journalists always report hot topics which lead to bias to unpopular fields. Furthermore, they often exaggerate the actual performance of the products of their countries. That is to say, some products may only be initially in the laboratory, news reports directly. Finally, it is difficult to cover the basic theory of technology due to the company's confidentiality. To track basics of specific element, it is recommended to further analyze the related literatures.

The core of the proposed methodology is deep learning, so it has the 'black box' feature and encapsulates the specific recognition process. It takes advantage of practicality and simplicity. As a result, performers only need to master basic terminology and annotation methods (Huang et al., 2015). At the same time, this method is also a sort of NLP method, so it works with not only news but also the traditional scientific literature, micro-blog, and technical reviews. In other words, it is available where text exists. In real life, companies hardly publish their core products in the form of papers. Research papers are more focused on specific scientific algorithms However, it is possible to discover public information from news. To sum up, the method proposed in this paper is effective to track and manage technical intelligence.

The results of the experiment perform well, and our methodology can effectively identify majority of new technical products and other technical elements. However, there are still many details to be considered. First, a product or other element may have multiple expressions such as MQ-9, also known as the reaper, the reaper drone or the reaper unmanned attack aircraft. Second, attention should be paid to punctuation and code. CH 4 and CH 5 are different, while SH 4 and SH-4 belong to the same product. Third, inclusion relations between elements should be analyzed, especially in the dimension of requirement. We extracted one requirement as surveillance, reconnaissance, and the other as intelligence, surveillance, and reconnaissance. The above discussions help to improve the quality of intelligence. Fourth, when visualizing, the Fig. 5 covers the maximal connected subgraph but not all nodes. There are still many isolated nodes and smaller connected blocks. They are meaningful for mining emerging topics. In previous studies, this problem was not obvious (Ritter et al., 2011). Theoretically there is no entirely independent individual in sociology. However, in science and technology, especially in military science and technology, there are some information islands called 'new concepts' and 'killer mace'.

We did not carry out more detailed analysis on the timeliness of high value nodes and news, largely because that all data used in test dataset are latest reported in 2017 and 2018. It is also recommended to divide the time into months or days. We evaluate influence and innovativeness with frequency based algorithm. The increase of data can eliminate some random errors in the big data environment. However, it prefers general concepts. For instance, the frequency of attack UAV must be greater than that of Reaper attack UAV. The Fig. 7 shows UAVs of CaiHong series after eliminating the general concept. Nodes are technical elements related to the CaiHongs, while edges represent their co-occurrence relationships. They mainly complete the integrated Mission of reconnaissance and strike under high altitude and long endurance. There are CaiHong-3, CaiHong-4, CaiHong-4B, CaiHong-5 as shown by yellow and blue nodes. They are mainly compared with Li Jian, Yi Long as shown by red nodes. Consequently, we can continuously weed out the general concepts to get more professional intelligence.

A variety of deep learning algorithms can extract entities from text. Our solution is the best one. Putthividhya described patents with unified social tag, such as movie, song, and book. It may be useful for service oriented enterprises, but not suitable for technology oriented ones (Putthividhya and Hu, 2011). Majumder obtained product titles from online and concentrated around the attribute 'Brand' (Majumder et al., 2018). Moreover, traditional natural language processing methods based on semantic structure show low-tech information, which is unacceptable in large data environments. Emerging technical products do not conform to established knowledge. They subvert the pre-existing technical requirement and challenge established managerial practice.

6. Conclusions

In this paper, the deep learning methodology is discussed for automatic extraction and discovery of technical intelligence. The major innovation is to identify the core technical elements which can be illustrated across the physics, social and cognitive spaces. The description of technology is deliberated at the front end on the idea of 5W1H, and evaluation of intelligence value is added to the back end. The three news webs we examined enabled us to grasp the trend of military industry in China, especially the unmanned products. Individuals or companies can apply our methodology not only to track existing scientific and technical information, but also to rapidly discover new information around the world. This will greatly improve the efficiency of technical activities. In addition to technology-intensive enterprises, the methodology can be used for strategic planning of science and technology by the council for science and technology.

Although the proposed solution is useful for informed decision making, it can still be strengthened as follows. First, the extracted information is still highly influenced by the choices of experts, such as the field of technology, the selection of data sources and word segmentation granularity. Second, the discussion on intelligence value is relatively poor. In addition to the three indicators proposed in this paper, technical foundations can also be used to evaluate the value of intelligence. Finally, it is badly in need of strengthening the study of intelligence differences to prevent identification results from focusing on the same product and the same country. The ongoing and future work of the authors is aimed at annotating data circularly that is critical to improve the accuracy of the solution with the technical elements identified in the previous stages.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grants 71690233 and 71671186. The authors would like to thank the anonymous referees for furnishing comments and constructive suggestions that improved the quality of this paper.

References

Acemoglu, D., et al., 2016. Innovation network. Proc. Natl. Acad. Sci. U. S. A. 113 (41), 11483.

Clauset, A., et al., 2017. Data-driven predictions in the science of science. Science 355 (6324), 477.

Derczynski, L., et al., 2015. Analysis of named entity recognition and linking for tweets. Inf. Process. Manage. 51 (2), 32–49.

Fantoni, G., et al., 2013. Automatic extraction of function–behaviour–state information from patents. Adv. Eng. Inform. 27 (3), 317–334.

Fortunato, S., et al., 2018. Science of science. Science 359 (6379), eaao0185.

Furukawa, T., et al., 2015. Identifying the evolutionary process of emerging technologies: a chronological network analysis of World Wide Web conference sessions. Technol. Forecast. Soc. Chang. 91, 280–294.

Grassano, N., et al., 2016. Funding data from publication acknowledgements: coverage, uses and limitations. J ASSOC INF SCI TECH 68 (4), 999–1017.

Gridach, M., 2017. Character-level neural network for biomedical named entity recognition. J. Biomed. Inform. 70, 85–91.

Habibi, M., et al., 2017. Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics 33 (14), 37–48.

Henry, S., Mcinnes, B.T., 2017. Literature based discovery: models, methods, and trends. J. Biomed. Inform. 74, 20–32.

Heymann, S., Grand, B.L., 2013. Visual analysis of complex networks for business intelligence with Gephi. information visualisation. In: The 17th International Conference Information Visualisation. 15. pp. 307–312.

Hogenboom, F., et al., 2016. A survey of event extraction methods from text for decision support systems. Decis. Support. Syst. 85 (C), 12–22.

Huang, Z., et al., 2015. Bidirectional LSTM-CRF models for sequence tagging. Computer Science (arXiv:1803.11284v1).

Huang, Y., et al., 2016. Big data and business: tech mining to capture business interests and activities around big data. IEEE International Conferences on Big Data and Cloud Computing 145–150.

Joung, J., Kim, K., 2017. Monitoring emerging technologies for technology planning using technical keyword based analysis from patent data. Technol. Forecast. Soc. Chang. 114, 281–292.

Le, H.Q., et al., 2018. D3NER: biomedical named entity recognition using CRF-biLSTM improved with fine-tuned embeddings of various linguistic information. Bioinformatics 34 (20), 3539–3546.

Leydesdorff, L., et al., 2016. Cited references and Medical Subject Headings (MeSH) as two different knowledge representations: clustering and mappings at the paper level. Scientometrics 109 (3), 2077–2091.

Ma, X., 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 64–74.

Ma, T., et al., 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. Chang. 116, 162–180.

Majumder, B.P., et al., 2018. Deep recurrent neural networks for product attribute extraction in eCommerce. arXiv (1803.11284v1).

Porter, A.L., et al., 2015. MetaData: BigData research evolving across disciplines, players, and topics. In: IEEE International Congress on Big Data, pp. 262–267.

Putthividhya, D., Hu, J., 2011. Bootstrapped named entity recognition for product attribute extraction. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 1557–1567.

Ritter, A., et al., 2011. Named entity recognition in tweets: an experimental study. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 1524–1534.

Rotolo, D., et al., 2015. What is an emerging technology? Res. Policy 44 (10), 1827–1843.

Sebastian, Y., 2017. Literature-based discovery by learning heterogeneous bibliographic information networks. ACM SIGIR Forum 51 (1), 75–76.

Small, H., et al., 2017. Discovering discoveries: identifying biomedical discoveries using citation contexts. J. Inf. Secur. 11 (1), 46–62.

Vossen, P., et al., 2016. NewsReader: using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news. Knowledge-Based Syst 110, 60–85.

Wang, Q., 2017. A bibliometric model for identifying emerging research topics. J ASSOC INF SCI TECH 69 (2), 290–304.

Xie, W., et al., 2013. TopicSketch: real-time bursty topic detection from Twitter. In: 2013 IEEE International Conference on Data Mining, pp. 837–846.

Yang, C., et al., 2017. Requirement-oriented core technological components' identification based on SAO analysis. Scientometrics 112 (3), 1–20.

Yoon, J., et al., 2011. Invention property-function network analysis of patents: a case of silicon-based thin film solar cells. Scientometrics 86 (3), 687–703.

Yoon, J., et al., 2014. Tracing evolving trends in printed electronics using patent information. J. Nanopart. Res. 16 (7), 2471.

You, H., et al., 2017. Development trend forecasting for coherent light generator technology based on patent citation network analysis. Scientometrics 111 (1), 297–315.

Yu, C.-H., et al., 2016. Quantum algorithm for association rules mining. Phys. Rev. A 94 (4), 042311.

Zhang, Y., et al., 2016. Event recognition based on deep learning in Chinese texts. PLoS One 11 (8), e0160147.

Jianguo Xu received the degree of Master of Technology Economics and Management in Management from the National University of Defense Technology (NUDT), Changsha, P. R. China, in 2016. He is pursuing the Ph.D. degree of Management Science and

Engineering in the College of Systems Engineering at NUDT. His main research interests include Deep Learning, Technological Forecasting, and Science of Science.

Lixiang Guo received the B.Eng. degree in systems engineering and the M.Eng. degree in management science and engineering from the National University of Defense Technology (NUDT), Changsha, China, in 2014 and 2016, respectively, where he is currently pursuing the Ph.D. degree in the College of Systems Engineering. His research interests include information extraction, data mining, and text analytics.

Jiang Jiang is a visiting scholar at Harvard University, Cambridge, MA, USA. He received the Ph.D. degree in management science and engineering from the College of Information System and Management, National University of Defense Technology, Changsha, China. He is an Associate Professor of the College of Systems Engineering, National University of Defense Technology. His research interests include evidential reasoning, data mining, and big data.

Bingfeng Ge is currently an Associate Professor of the College of Systems Engineering at the National University of Defense Technology. He was a visiting scholar with the Conflict Analysis Group, Department of Systems Design Engineering, University of Waterloo, Waterloo, ON, Canada. His research interests include system-of-systems architecting and engineering management, portfolio decision analysis and conflict resolution.

Mengjun Li is a Professor of the College of Systems Engineering at the National University of Defense Technology. He was a senior visiting scholar at University of Birmingham, UK. His research interests include development of methodology for technology and innovation management.