International Journal of Knowledge and Language Processing
Volume 9, Number 3, **2018**

www.ijklp.org
KLP International ©2018 ISSN 2191-2734
pp.1–15

A Literature Review on Patent Texts Analysis Techniques

Guanlin Li
School of Software & Microelectronics
Peking University
No.5 Yiheyuan Road Haidian District
Beijing, 100871, China
lgl97@pku.edu.cn

Received Sep 2018; revised Sep 2018

ABSTRACT. *Patent data are expanding explosively nowadays with the advent of new technologies, and it's significant to put forward the method of automatic patent analysis and use appropriate patent analysis techniques to make use of scattered, multi-source and interrelated patent text, in order to improve the efficiency of patent analyzing. Currently there are a lot of techniques being used to process patent intelligence. This literature review focuses on automatic patent text analysis techniques, which use computer to automatically analyze large scale of patent texts and find useful information in them. These techniques are divided into the following categories: semantic analysis based techniques, rule based techniques, machine learning based techniques and patent text clustering techniques.*

Keywords: Patent analysis, text mining, patent intelligence

1. **Introduction.** Patents are important sources of technology information, in which we can find great value of scientific and technological intelligence. At the same time, patents are of high commercial value. Enterprise analyzes patent information, which contains more than 90% of the world's scientific and technological information, to collect competitive intelligence and ensure the success of business decisions.

Patent analysis is a process in which the technology information, economic information and legal information contained in patent documents are dealt with, sorted out and analyzed, to form patent intelligence of high technical and commercial value.[1]

According to *World Intellectual Property Indicator 2017*, in 2016, the number of patent applications filed worldwide exceeded 3 million for the first time, which is an 8.3% YOY increase; China received more patent applications than the European Patent Office, Japan, Korea and the United States combined, ranking first in the world. It can be seen that patent

documents are emerging in large numbers at present, and the traditional manual reading method for patent analysis requires analysts to have a strong professional background, at the same time. The traditional method is also very time-consuming and labor-consuming and has obvious shortcomings in the case of a huge amount of texts. [2] In the era of massive information and big data, it is significant to put forward the method of automatic patent analysis and use appropriate patent analysis technology to make use of scattered, multi-source and interrelated patent text, so as to improve the efficiency of patent analyzing.

2. **A survey of current patent analysis techniques.** At present, many computer-aided technologies have been used to process patent texts. These technologies rely on computer, using algorithms to process patent texts, and extract useful information from large-scale patent texts for analysis. It is helpful to get a general view of related technologies by classifying and reviewing them.

A. Abbas et al. [3] proposed that patent analysis techniques can be divided into two categories: text mining-based patent analysis and visualization-based patent analysis. Patent analysis based on text mining includes natural language processing method, property-function analysis method, rule-based method, semantic-based analysis method and neural network method. Patent analysis based on visualization technology includes patent network method and clustering method. Specifically, the technologies based on natural language processing include: keyword-based analysis method, SAO (subject-action-object) analysis method; rule-based methods include: association rules mining and fuzzy reasoning rules; semantic-based methods include: hierarchical keyword vector, semantic representation and domain ontology-based methods. The methods based on the neural network include the neural network based on back propagation algorithm and the neural network based on Kohonen learning algorithm. Patent analysis based on visualization technology includes: patent network data mining; visualization analysis based on K-means keyword clustering; self-organizing mapping clustering.

Yuen-Hsien Tseng et al.[4] proposed that patent text mining technology mainly includes data pre-processing, indexing, topic clustering and topic map. Specifically, it includes the following technologies: text segmentation based on tags of structured data, text summarization, stop word filtering and stemming extraction, keyword and key phrase extraction, associated word statistics and clustering techniques.

Wang Yuefen et al. [5] proposed that patent analysis includes patent text information pre-processing, patent content information analysis and patent knowledge information processing. The main technologies mentioned in those steps include text segmentation, content feature extraction, feature weight calculation and text clustering.

Qupeng et al. [6] proposed that patent analysis techniques mainly include terminology extraction, clustering, classification, network-based methods, time-based analysis and patent text mining techniques.

This paper summarizes automatic patent texts analysis techniques. These techniques are used to extract useful information after obtaining patent texts. There exist many techniques,

and many techniques are related to each other, so it is very difficult to find distinctive boundaries between different techniques. Based on literature reviews and papers retrieved from academic databases, this paper divides these techniques into the following categories: semantic analysis based techniques, rule based techniques, machine learning based techniques and patent text clustering techniques.

3. **Semantic analysis methods.** Semantic analysis uses computer to analysis large scaled texts and extracts semantic structures such as key words, to analyze the relationship between meanings of texts. Semantic analysis of patent texts mainly includes the following: keyword extraction, semantic analysis method based on SAO structure, patent analysis, ontology based patent semantic analysis, and other semantic representation methods, including hierarchical keyword vector and property-function method.

3.1. **Key word extraction.** When applying statistical analysis to patent texts, the extraction of keywords is a basic step. Manual keyword indexing is time-consuming and labor-consuming and is often dubious. Automatic keyword extraction can greatly improve the efficiency of this process to further analyze patent text.

At present, there are many researches studying keyword extraction at home and abroad. Hundreds of automatic keyword extraction methods are proposed in recent years, in which there are more than 10 mainstream automatic keyword extraction methods,[7] such as linguistic analysis methods, statistical methods and machine learning methods. Statistical methods include frequency-based methods, such as TF-IDF, and word association information, such as mutual information, entropy, PageRank value, etc. Other methods include topic model, mainly LDA [8] model methods and complex network graph methods.

Liu Dacheng et al. [9] proposed a semi-automatic method for extracting technology-effect phrases from Chinese patent abstracts. Technology-effect Clauses (TEC) points out the technologies used in patents and the efficiencies achieved, which can be used for in-depth analysis of patent data in specific fields. To construct technology-effect matrix and TEC, the basic step is to extract technology-effect phrases. Patent abstracts contain the most useful information in patent data, but phrases in abstracts have relatively low word frequency. Therefore, Liu et al. proposed a method for extracting technology-effect phrases based on abstracts without word frequency statistics and with a small amount of manual work. Liu et al. first analyzed the grammatical composition of technology-effect phrases and divided them into two parts: attribute and value, corresponding to two grammatical components, such as adjective+noun. Then attribute corpus and value corpus are constructed. This step requires human participation. Furthermore, domain-dependent corpus are constructed, to reduce human participation. By matching corpus, technology-effect phrases in abstracts are extracted.

Li Junfeng et al. [10] proposed a patent keyword indexing method based on weighted complex graph model. The main methods are as follows: using k-nearest neighbor coupling graph to map patent documents into complex network graph model, then combining the change of average path with average clustering coefficient and influence of current nodes on the mobility of the whole complex graph model to propose the evaluation index of

average connectivity weight. Word position, word span and inverse document frequency of keywords are analyzed, to measure the importance of keywords while combining relevant features of patents. According to the weights of candidate keywords the ranked results are generated, and Top-N results are extracted from ranked results as keywords index.

3.2. **Semantic analysis based on SAO structure.** SAO structure means Subject-Action-Object structure, which is a subject (noun) - predicate (verb) - object (noun) structure. SAO structures can be a problem-solving structure, where AO represents problems or required functionalities and S represents solutions. In addition, S and O can represent components of the system, and A is usually used to describe how these components implement their functions. Therefore, the key words, key technologies and the relationships among their components can be obtained by extracting the SAO structure information from scientific and technology texts. In addition, SAO structure can also represent the inclusion relationship between products or technologies. If verbs are partial verbs such as "have" and "composed of", the components represented by Subject S may contain the components represented by Object O. Therefore, SAO structure information can also be used to describe the relationship between products and technologies.[11] Key words have certain shortcomings in reflecting the key concepts of specific technologies and the structural relationships of components. Compared with keywords, SAO structure can better reflect the key concepts of specific technologies and relationship between them, [12] and can take into account the semantic relationship between keywords.

3.2.1. **Technology analysis contained in patents using SAO structure.** By using the SAO structure to represent the semantics of the patent, information related to the technology can be extracted from the patent based on SAO, thereby the technology in the patent can be represented.

Guo Junfang et al. [11] proposed a new technology pattern recognition method based on SAO semantic mining method: the idea of SAO chain is proposed. On the basis of SAO structure extraction, multiple SAO structures are connected in series or in parallel to form a semantic chain to identify Semantic relationships between key concepts. Based on the SAO chain, the hierarchical relationship of technical keywords can be obtained. Through manual sorting and similar keyword merging, the hierarchical relationship of key components/technologies in the technical field can be obtained, and the technical form structure can be formed.

Sungchul Choi et al. [13] proposed a method for building technology tree based on SAO. A technology tree is a branch diagram that expresses the relationship between components, technologies, or technical capabilities in a particular technology field. Choi proposes a method for generating a technology tree based on the SAO method, so that a large amount of information of different technologies can be provided at the same time. The method first extracts the SAO structure from the patent text, then calculates the SAO similarity to cluster the phrase and AO structure, and then analyzes the phrase and AO type. After analyzing the SAO, the technology tree is constructed based on SAO.

3.2.2. **Patent map based on SAO structure.** According to the semantic representation based on SAO, different kinds of maps can be drawn for patent to discover useful

intelligence.

Duan Qingfeng et al. [14] proposed a method for constructing a patent technology-effect map based on SAO structure. According to the research of Duan, the technology-effect map is a two-dimensional technology space consisting of technology and effect, which quickly and accurately condenses the technology theme and effect theme. Duan also found that the semantic association between the two is the key. The technology and effect terms are embedded in the SAO structure, and the semantic structure of SAO provides a good way to discover the intrinsic relationship of technology-effect. The extraction and condensing of technology words and effect words can be based on SAO. According to their specific semantic types, technology-effect relationships can be extracted. The proposed technology-effect map construction procedure based on SAO structure is as follows: patent data collection - SAO structure extraction and semantic annotation - technology words, effect word extraction - technology, effect theme clustering - technology-effect matrix construction - technology-effect map drawing.

Wu Feifei et al. [15] proposed a technology application domain identification method based on patent SAO structure. The technology application domain identification method procedure mentioned in the method is: obtaining the USE field (describe the usage-application domain of the group of patent data) - using NLP tool, extracting SAO structure from the USE field - performing semantic similarity calculation on the SAO structure - drawing a patent map to realize the technical application domain identification based on semantic clustering.

Wang XueFeng et al. [16] proposed a R&D partner identification method based on SAO analysis. By extracting the SAO structure in the title and abstract of patent documents, the SAO structure diagrams are drawn, covering the three dimensions of materials, technologies and components, and target. By doing so, Fu proposed mining the institutions with similar research and development objectives and analyzing the possibility of mutual cooperation and possible cooperation directions.

3.2.3. **Patent similarity analysis based on SAO structure.** The SAO structure is a semantic representation of the patent text. Naturally, the similarity between patent texts can be calculated by calculating the similarity between SAOs.

Hyunseok Park et al. [12] proposed a method combining TRIZ and SAO to analyze patent documents to identify potentially useful patents. Park et al. suggested that the rule base of RFJ (reasons for jumps in TRIZ, a set of functions that cause a technology to jump or evolve to the next phase in the trend) can be represented by AO, so the patent can be evaluated by comparing the semantic similarity between rule base and AO component of extracted SAO structure. Park proposes that patent analysis methods using SAO and TRIZ include: patent text collection - analyzing technology life cycle of patent - extracting SAO structure from patent text - analyzing TRIZ trend - calculating semantic similarity between AO of SAO structure and RFJ rule base, thus realize the evaluation of patents. Patent can be classified as patents that may be promising in the future if it is more relevant to TRIZ trends which may have value in the future.

Hyunseok Park et al. [17] also proposed a method for judging patent infringement based

on SAO analysis. Park believed that the method of judging technology similarity based on keywords is not sufficient for patent infringement detection and judgement because keyword vector cannot reflect the structural relationship between key findings of a particular technology and technical components. Therefore, Park proposed a method for judging similarity using SAO structure. Using WordNet to calculate semantic similarity, use MDS to generate patent maps, and finally automatically proposing patent texts that may cause patent infringement through clustering algorithms.

Janghyeok Yoon et al. [18] proposed a method for constructing dynamic patent maps based on SAO structure for patent analysis. This method can be divided into four steps: collecting patent data; extracting SAO structure from patent text; calculating semantic similarity based on SAO structure of patent texts and constructing similarity matrix; visualizing similarity by multidimensional scaling analysis (MDS) and constructing dynamic patent map. The patent map generated can be used to identify technological areas in which patents have not been granted (‘‘patent vacuums’’), areas in which many patents have actively appeared (‘‘technological hot spots’’), R&D overlap of technological competitors, and characteristics of patent clusters. However, the disadvantage is that clustering with MDS and K-means may result in information loss, resulting in incorrect clustering.

Gerken and Moehrle [19] proposed a patent monitoring method for patent novelty discovering based on semantic analysis, in which the method of SAO was adopted to perform analysis from syntactic to semantic. The method is divided into four steps. The first step is to extract the semantic structure, that is, to extract the SAO structure. The second step is to perform language analysis on the domain-related elements. The purpose of this step is to semantically disambiguate domain-related words and filter stop words. The third step is to generate a similarity matrix; the fourth step is to perform similarity matrix calculation on patent texts to detect new patents. The limitation of this method is that new patents are detected by calculate the similarity distance between the patent text and the closest patent, but the relationship between patents may be more complicated and the similarity between patents may be decided by factors which have more weights than the degree of novelty. More so the method of calculating the novelty degree of patents by similarity needs to be improved.

3.3. **Patent semantic analysis based on ontology.** In computer science, ontology is used to describe the concepts and relationships between these concepts in a certain and even wider field, so that these concepts and relationships have a common, clear, and unique definition. [20] The ontology can be used to clearly and standardly illustrate the conceptual model contained in the patent text, so that the patent data is represented in a consistent form and can be processed and analyzed by computer.

3.3.1. **Semantic annotation based on patent ontology** In China, Hu Zhengyin et al. pioneered the research on ontology-based patent semantic annotation. The use of patent ontology not only reflects intelligence about the patent, but also reflects the relationship between intelligence by defining the attribute or property. Semantic Annotation is a common method of extracting meta-information from text data. Ontology-based semantic

annotation is to annotate, in patent texts, elements corresponding to those in ontology. The essence of semantic annotation of patent texts is to map the corresponding patent data into the instance of patent ontology after mapping the traditional patent database structure into patent ontology. [21] Hu Zhengyin et al. studied the application-oriented patent ontology schema construction and the semantic annotation of patent metadata fields and text fields.

Nizar Ghoula et al. [22] proposed an ontology-based semantic annotation technique to analyze patent texts. Based on the hierarchical ontology, appropriate annotation granularity can be selected when generating annotations, knowledge disambiguation and can also be performed, to reduce errors in the annotation process. Ghoula et al. suggested that this annotation process can include: structuring patent documents - mapping patent texts to patent ontology - generating metadata annotations - generating domain-based annotations - merging annotations - generating semantic annotations. MeatAnnot system were used to generate annotations from patent text fields.

3.3.2. **Patent text analysis techniques based on patent ontology.** Tao Ran et al. [23] proposed an ontology-based patent intelligence discovery system. This system first constructs an ontology for a certain domain, and then performs semantic annotation on the patent text based on the ontology, that is, mapping the patent data to an instance of the ontology. When the patent data is represented by the ontology, the knowledge representation of the patent data is consistent, which provides a guarantee for the inference engine to use external method libraries and the model libraries. Taking the ontology as the core, the patent is analyzed by using a preset algorithm such as technical indicators and reasoning system.

Xu Xin et al. [24] proposed a semantic retrieval and analysis system based on patent ontology. Semantic retrieval is realized by constructing ontology for domain-dependent patents. Different from keyword retrieval, ontology-based semantic retrieval will perform semantic extension on the search term through the ontology layer; the temporary retrieval result obtained by the extension through ontology layer can then acquire other information such as patent agent, patent owner and so on by instance reasoning in ontology. For the results obtained by semantic retrieval, the system can also analyze these results based on the corresponding analysis models and the preset algorithms, to perform patent analysis such as obtaining patent trend information, checking similar patents and detecting important patent detection.

Hu Zhengyin et al. [25] proposed a patent technology mining method based on semantic TRIZ. The study combines SAO structure and ontology to describe the semantic index structure. It constructs semantic TRIZ from three dimensions: concept space, index space and application space; it describes semantic TRIZ index structure from three aspects: the micro layer, SAO basic semantic unit, middle layer, P&S ontology and macro layer technical category. The research framework mentioned in the paper is as follows: Firstly, based on the domain core patent set, the semantic TRIZ concept space is constructed, and the domain core P&S ontology is obtained. Secondly, for the patent data set to be mined, the semantic TRIZ index space is constructed based on the domain core P&S ontology; Last, as for concrete technology mining applications, building specific application space

based on TRIZ patent semantic index, such as technology topic clustering, patent automatic classification, patent technology evolution and so on.

3.4. **Other patent semantic representation methods.**

3.4.1. **Hierarchical keyword vector.** Changyong Lee et al. [26] proposed a method of extracting hierarchical keyword vectors to semantically analyze the scope of patent applications in order to detect possible patent infringement and identify which patent applications need manual detection. The core of the method is to express the dependencies among patent claim elements (as well as unstructured textual information) through hierarchical key vectors, which can solve the problem that traditional methods cannot express the dependencies of semi-structured data or unstructured data. The tree matching algorithm is used to compare the elements of patent right claims. Compared with the method of comparing technical keywords, this method effectively deals with semantic relations.

3.4.2. **Property-function.** Janghyeok Yoon et al. [27] proposed that through grammatical analysis of patent texts, the properties and functions of texts can be obtained to represent the innovative concepts in the patent texts. The property-function method is originally used to analyze other products in different areas to further improve a certain product. Property is a feature of a system that is often described by adjectives; function is a functioning behavior provided by a system, usually described by nouns. [28] Yoon proposed to build a patent network based on property-function extraction, which can describe the relevance of innovative concepts in patent texts. This patent network can analyze the intrinsic association of patents in a large number of new patent texts and conduct quantitative analysis by means of network analysis. Yoon analyzes the syntax dependency tree with NLP method to extract property-function phrases; a matrix is built based on property, function phrases, and their co-occurrence values, and thus calculating similarities to visualize them into networks, and analyzing patent networks.

4. **Patent text analysis based on rules.** Rule-based techniques require expert knowledge to build rule sets through analysis on targeted patent texts, and automatically process and analyze acquired patent texts based on rules.

Meng-Jung Shih et al. [29] proposed a technique to automatically mine changes in patent development trends based on indicator analysis and mining rules. This technique does not require experts to analyze patent charts of different periods. It's based on association rules to discover changes in development trends through computers. The method can be divided into four steps. The first step is to obtain semi-structured HTML data through the Internet; the second step is to structure and clean the data and store them in the database; the third step is to calculate the indicators to determine the patent value. The indicators are: citation indicators, originality, versatility and technology life cycle; the last step is to detect changes in patent development. Shih et al. obtain sets of changes in the development trend by comparing changes in association rule sets of patents in different periods, to assess the extent of the change and generate report.

Qian Zengqi et al. [30] proposed an association rule algorithm based on the

TCM(Traditional Chinese Medicine) patent data set. Through the improved Apriori algorithm, Qian et al. proposed a method for discovering the association rules between prescriptions in TCM patent data.

W. D. Yu [31] proposed an IF-THEN rule fuzzy inference system based on patent analysis technology to provide advice for company's technology strategy. Unlike traditional technology management and decision-making tools, the source of the knowledge base of this system is not the domain expert, but the result of analyzing the patent database using patent analysis techniques. Yu uses a computer-aided patent analysis tool to analyze patent data and uses the results as input to the inference system, such as: Patent Quantity (PQ), Revealed Patent Advantage (RPA), Patent Activity (PA), Be Cited Rate (BCR), and Relative Citation Index (RCI). Then, by analyzing the existing development strategy, the IF-THEN rule set is generated, and the reasoning system is constructed by using the Kohonen learning algorithm and the first nearest neighbor heuristic algorithm.

5. **Patent text analysis techniques based on machine learning.**

5.1. **Patent indicator analysis based on machine learning.**

In patent text analysis, machine learning can be used to analyze indicators obtained from patent data.

Zhao Yunhua et al. [32] proposed a patent value evaluation method based on machine learning. The method considers the patent value assessment as an intensity classification problem. First, the patent data are filtered, and the filtered data are quantified, and the appropriate elements are selected from the data as indicators for value evaluation. The intensity of the sorted indicators is classified by decision tree, support vector machine algorithm and neural network algorithm, in order to conduct a patent value assessment. The problem with this approach is that when selecting value assessment indicators, some non-numeric data cannot be converted into numerical values, so they cannot be selected as indicators, and these data may have relatively great impact on the value of patents.

5.2. **Patent auto-classification based on machine learning.**

In the process of patent analysis, the automatic classification of patent texts is mainly carried out by means of machine learning. This process generally includes: preprocessing the patent text, extracting text features, patent text representation, training patent texts, obtaining a classifier, testing the patent text test set using a classifier, and using the obtained results to continuously improve the training to ultimately get a more accurate classifier. [33] When conducting classification research of patent texts, most of the researches focus on transforming existing algorithms or combining single algorithms and applying them to patent texts; the patent classification system mainly refers to IPC. [34] The algorithms for constructing classifiers mainly include NB algorithm, ANN algorithm, Rocchio algorithm, KNN algorithm, SVM algorithm and so on.[33]

5.2.1. **Methods based on naïve Bayes.**

There are two types of classifiers based on Bayesian methods, one is Naive Bayes classifier and the other is Bayesian network classifier. The naïve Bayes classifier is simpler to implement and has a faster training speed, the classification accuracy is high when the independence hypothesis is satisfied; the Bayesian network classifier can consider the relationship between features, but the

implementation is more complicated. [35] The patent text classification research based on naïve Bayes starts early. Guo Yuqiang et al. [35] proposed a naïve Bayes classifier, and introduced the title weight coefficient to improve the weight according to the characteristics of patent text, and realized an automatic classification system for patent texts.

5.2.2. **Methods based on neural network.** The patent text classification based on neural network can map complex nonlinear relationships, and it has high classification accuracy. Amy J.C. Trappey et al. [36] proposed a neural network based on the backward propagation algorithm to classify patent texts. The method first extracts keywords from the patent text and calculates the importance of the keywords according to the word frequency. The similarity of the keywords is calculated by analyzing the co-occurrence value to control the number of keywords. The backpropagation algorithm is used as the classifier algorithm, and the final output is based on the IPC standard.

With the rise of deep learning, deep neural networks have also been applied to the study of patent classification. Grawe et al. [37] proposed an automatic classification method for patent text based on word embedding vectors. The method uses Word2Vec to process patent text data, and uses the LSTM network (long short term memory network) inherited from the recurrent neural network to train the data.

Bing Xia et al. [38] proposed a patent document classification method based on deep learning. The method uses the sparse auto-encoder and deep belief network to conduct feature learning on preprocessed patent data, and then uses softmax regression to classify the text.

Hu Jie et al. [39] proposed a classification model for patents from mechanical fields based on convolutional neural networks and random forest algorithms. Hu Jie et al. proposed that the k-nearest neighbor algorithm needs to compare all the documents in the sample space when predicting new patent document classification labels, and the time complexity is high; and when the training samples are unbalanced, the probability of a document being predicted as a large sample by the classifier increases. The naïve Bayes algorithm treats the features in the patent document independently, but the features in the patent text are closely related. The SVM algorithm needs to train multiple classifiers when performing multi-class and multi-label patent text classification tasks, resulting in greatly increased time cost and computational cost. Therefore, this model proposes a hybrid classification algorithm that combines convolutional neural networks and random forest classifiers to improve the efficiency of automatic classification of patent texts.

5.2.3. **KNN methods.** The KNN algorithm has a wide range of applications in the classification of patent texts. Jae-Ho Kim et al. [40] proposed a patent document classification system based on semantic structure information, and its classification method is KNN algorithm. Kim et al. regards the specific parts of the patent text, such as the claim, purpose, and application field of a patent application as semantic elements, and clusters these elements as the basic features of document classification.

Jiang Chuntao [41] proposed a method for automatically extracting the semantic information of Chinese patent texts using the graph method. The method designs two models of graph structure, one is a text graph model based on keywords, and another is a

text graph model based on dependency tree. The method uses the frequent subgraph mining as the feature to represent the text space vector and constructs the patent text classifier by KNN algorithm.

Liao Liefa et al. [42] proposed a patent text classification method based on LDA model. The LDA model is a full probability generative model. The method uses the LDA model to perform text representation, extracting topics with semantic information to optimize the selection of feature values, in order to improve the efficiency of document classification. The KNN algorithm is used to automatically classify documents.

5.2.4. **SVM methods.** Chih-Hung Wu et al. [43] proposed a patent classification system through SVM based on the new hybrid genetic algorithm (HGA-SVM). HGA-SVM classifies patent documents by recording and learning experts' knowledge and logic. The system has achieved high classification accuracy and versatility.

Lu Baoliang [44] and his team proposed a Parallel Min-Max Modular Support Vector Machine (m3-SVM) for patent text classification. The traditional text classification method is not efficient on a large data scale. Compared to the standard SVM algorithm, the training time of the M3-SVM algorithm is greatly shortened, so it can be used to process large-scale patent data sets. The algorithm also gets higher F-measure.

5.2.5. **Hybrid classification algorithms.** Liu Su Houn et al. [45] proposed a patent classification system that combines multiple classification algorithms. The system combines three classification algorithms (Naïve Bayes, KNN and Rocchio), and the classification results of different classifiers are combined by voting and sampling mechanisms. Results of experiments show that this patent classification system combined with multiple classification algorithms has higher accuracy and is more stable than a single classification method.

Jia Shanshan et al. [46] proposed a patent automatic classification method based on integration of multi-feature and multi-classifier. The classifier extracts from the patent application files the full dictionary TF-IDF feature, the information gain dictionary TF-IDF feature, the paragraph vector feature, the topic model vector feature. It trains Naïve Bayes, support vector machine, AdaBoost classifier respectively to construct the feature-category matrix and integrates with the F1 weight matrix to obtain the final predicted IPC classification number. Results of experiments show that this classification method can improve the classification accuracy of patent texts in specific domains.

6. **Patent analysis based on text clustering.** Text clustering is the process of automatically categorizing document collections. The categories of text clusters are not predetermined but are derived from machine learning related data. The goal of text clustering is to divide the document collection into several classes and ensure the document content similarity in same classes to be as large as possible, and the similarity of the document content in different classes is as small as possible. Research in China on text clustering is conducted relatively late than other countries, and researches mainly introduce methods which originally deal with other languages to process Chinese.[47] The clustering of patent texts selects text features by extracting specific keywords from patent texts and calculates

similarity of texts based on text features to achieve patent text clustering.

6.1. **Methods for clustering patent texts.** Qu Junwei et al. [48] studied the application of self-organizing map (SOM) in patent text clustering. SOM does not require supervision and prior knowledge, it automatically analyzes the intrinsic features of the sample and reveals the similarity of the samples as well as mapping high-dimensional data to low-dimensional, to implement pattern recognition and cluster mapping. The research proposes that SOM-based text processing is a suitable method for clustering large-scale documents. It compares the SOM based patent text clustering with k-means and TwoStep, and concludes that the SOM is better.

Fan Yu et al. [49] proposed a patent clustering technique based on LDA model, and proposed a method combining the potential Dirichlet distribution (LDA) topic model and the OPTICS algorithm. The study uses LDA topic model to transform the representation of patent information from the high-dimensional lexical space to the low-dimensional the topic space, and effectively realizes the dimensionality reduction of patent information. The method then uses the OPTICS algorithm and the k-nearest neighbor to cluster the patent information and analyze it. The number of topics in the specified LDA model is k. By training the model, the distribution of k topics of the document is obtained, and the topic space representation of the document is obtained. OPTICS is a density-based clustering algorithm. The method combines the OPTICS and k-nearest neighbor algorithm to cluster documents.

Tian Dongyang [50] proposed a patent data clustering method based on M3-DGMF, which is a supervised clustering algorithm to solve the clustering performance problem of large-scale data through minimum and maximum modular neural network (M3). Combined with double Gaussian synthesis, the function can be used to effectively trim the scale of the training data and to cluster patent data.

6.2. **Patent text analysis based on clustering.** Chen Xu et al. [51] proposed a patent analysis method based on technology effect matrix. Chen Xu et al. mainly studied the text representation and result visualization in patent text clustering. Unstructured patent texts are annotated based on technology and effect, to construct patent effect matrix, thus conducting patent clustering; Structed data are first roughly divided according to the IPC classification number, and then clustered by the k-means method. The structure of clusters is analyzed to form a multi-level patent map, and experiments are carried out, which show that this method is more efficient and has better clustering effect than the traditional vector space model method, and the visualization of its clustering result is more practical and more extensible.

Woo Jin Lee et al. [52] proposed a method for identifying shale gas development through patent clustering analysis. Methods extract terms in the abstract of related patents and constructs weight matrix, where the matrix dimensionality was reduced by SVD method. Clustering is performed using the maximum expectation algorithm, and each cluster uses keywords to represent the relevant technology field.

Helen Niemann et al. [53] proposed a method for analyzing patent trends over time by clustering. This method uses patent path to analyze changes in patent trends. The patent path can be seen as a distribution of patent clusters over time. The patent path is obtained

by clustering patents and sorting the clusters by period.

Gabjo Kim et al. [54] predict potential technology by clustering patents and analyze clusters. Firstly, the obtained patent text data is clustered by cooperative patent classification (CPC) and K-means methods to form technology clusters; these clusters are defined by the description of CPC; by analyzing forward citations, triadic patents and independent claims of technology clusters, the method predicts potential technologies.

7. **Conclusions.** As information technology develops, the patent data has increased substantially, and the automatic analysis of patent texts has been of great significance in detecting the trend of technology development, planning technology development strategies and discovering the patent values. Studies on natural language processing and text mining technology have been carried out earlier in other countries than in China, and they have been applied to the analysis of patent texts earlier; in China, research has also been carried out in related fields recent years, mainly focusing on the application of existing text mining techniques or improvement of these techniques to suit them for patent texts process. How to apply a variety of technologies properly or improve existing technologies, so that they can better analyze patent texts in specific domains, meet particular needs or perform high-efficiency analysis on large-scale texts, is worth considering in the case of application.

REFERENCES

[1] Fang Wei, Zhang Wei, and Xiao. J., Patent Intelligence Analysis Methods and Applied Research, Library and Information Knowledge no.4, pp.64-69, 2007.

[2] Hu Pei, et al., A review of patented subject analysis based on text mining, no.12, pp.88-92, 2013.

[3] Abbas A., L. Zhang, and S.U.J.W.P.I. Khan, A literature review on the state-of-the-art in patent analysis, vol.37, no.4, pp.3-13, 2014.

[4] Tseng Y.H., et al., Text mining techniques for patent analysis, vol.43, no.5, pp.1216-1247, 2007.

[5] Wang Yufen, Xu Dandan, Patent Information Content Mining and Experimental Research, Data Analysis and Knowledge Discovery vol.24, no.12, pp.59-65, 2008.

[6] Qu Peng, et al., Domestic and Foreign Patent Mining Research (2005-2014), no.20, pp.131-137, 2014.

[7] Zhao Jingsheng, et al., Overview of Automatic Keyword Extraction Research, vol.28, no.9, pp.2431-2449, 2017.

[8] Blei D.M., A.Y. Ng, and M.I.J.J.o.M.L.R. Jordan, Latent Dirichlet Allocation, vol.3, pp.993-1022, 2012.

[9] Liu D., et al. Technology effect phrase extraction in Chinese patent abstracts, Asia-Pacific Web Conference. Springer, 2014.

[10] Li Junfeng, Lu Xueqiang, and Zhou. J., Patent Keyword Indexing Research on Weighted Complex Graph Model, Modern Library and Information Technology vol.31, no.3, pp.26-32, 2015.

[11] Guo Junfang, et al., A new type of technical pattern recognition method based on SAO semantic mining method, vol.34, no.1, pp.13-21, 2016.

[12] Park H., J.J. Ree, and K.J.E.S.w.A. Kim, Identification of promising patents for technology transfers using TRIZ evolution trends, vol.40, no.2, pp.736-743, 2013.

[13] Choi S., et al., An SAO-based text mining approach to building a technology tree for technology planning, vol.39, no.13, pp.11443-11455, 2012.
[14] Duan Qingfeng and Jiang J., Research on the construction of patent technology efficacy map based on SAO structure, Modern Intelligence vol.37, no.6, pp.48-54, 2017.
[15] Wu Feifei, Li Qian, and Huang. J., Research on identification methods of technology application fields based on patented SAO structure, scientific research management vol.35, no.6, pp.1-7, 2014.
[16] Wang Xuefeng, et al., R&D Partner Identification Research Based on SAO Analysis, vol.36, no.10, pp.19-27, 2015.
[17] Park H.J.S., Identifying patent infringement using SAO based semantic technological similarities, vol.90, no.2, pp.515-529, 2012.
[18] Yoon J., H. Park, and K.J.S. Kim, Identifying strategies for R&D planning using dynamic patent maps: SAO-based content analysis, vol.94, no.1, pp.313-331, 2013.
[19] Gerken, J.M.J.S., A new instrument for technology monitoring: novelty in patents measured by semantic patent analysis, vol.91, no.3, pp.645-670, 2012.
[20] Du Xiaoyong, Li Man, and Wang. J., Review of Ontology Learning, Journal of Software vol.17, no.9, pp.1837-1847, 2006.
[21] Hu Zhengyin, Fang Wei, and Xiao Guohua, Application Research of Ontology-based Semantic Annotation Technology in Patent Analysis, 2010.
[22] Ghoula, N., K. Khelif, and R. Dieng-Kuntz. Supporting Patent Mining by using Ontology-based Semantic Annotations, Ieee/wic/acm International Conference on Web Intelligence. 2007.
[23] Book and Information, Tao.J., Research on Patent Intelligence Discovery System Based on Ontology, no.4, pp.23-27, 2007.
[24] Xu Xin, et al. Design and Implementation of a Semantic Retrieval Analysis System for Patent Ontology, vol.58, no.9, pp.96-104, 2014.
[25] Hu Zhengyin, Fang Wei, and Zhang. J., Research on patent technology mining based on semantic TRIZ, Chengdu Literature and Information Center, Chinese Academy of Sciences, 2014.
[26] Lee C., et al., How to assess patent infringement risks: a semantic patent claim analysis using dependency relationships, vol.25, no.1, pp.23-38, 2013.
[27] Yoon, J. and KJESwA Kim, An analysis of property–function based patent networks for strategic R&D planning in fast-moving industries: The case of silicon-based thin film solar cells, vol.39, no.9, pp.7709-7717, 2012.
[28] Engineering, S.D.J.P., Directed variation of properties for new or improved function product DNA – A base for connect and develop, vol.9, pp.646-652, 2011.
[29] Shih M.J., D.R. Liu, and M.L.J.E.S.w.A. Hsu, Discovering competitive intelligence by mining changes in patent trends, vol.37, no.4, pp.2882-2890, 2010.
[30] Qian Zengqi, Xin Yan, and Ju J., Association Rule Discovery Algorithm Based on Traditional Chinese Medicine Patent Data Set, Computer Application Research vol.24, no.7, pp.61-63, 2007.
[31] Yu W.D. and S.S.J.A.i.C. Lo, Patent analysis-based fuzzy inference system for technological strategy planning, vol.18, no.6, pp.770-776, 2009.
[32] Zhao Yunhua, et al., Research on patent value evaluation methods based on machine learning, vol.31, no.12, pp.15-18, 2013.
[33] Liu Hongguang, Ma Shuanggang, and Liu.J., A review of patent text classification algorithms based on

machine learning, Library and Information Research no.3, pp.79-86, 2016.
[34] Qu Peng and Wang.J., Research on the basic issues of patent text classification, Modern Library and Information Technology vol.29, no.3, pp.38-44, 2013.
[35] Guo Yuqiang, Wen Jun, and Wen J., Patent Classification Based on Bayesian Model, Computer Engineering and Design vol.26, no.8, pp.1986-1987, 2005.
[36] Trappey A.J.C., et al., Development of a patent document classification and search platform using a back-propagation network, vol.31, no.4, pp.755-765, 2006.
[37] Grawe M.F., C.A. Martins, and A.G. Bonfante. Automated Patent Classification Using Word Embedding, IEEE International Conference on Machine Learning and Applications. 2018.
[38] Xia B., L. Baoan, and X. Lv. Research on patent document classification based on deep learning, International Conference on Artificial Intelligence and Industrial Engineering. 2016.
[39] Jie H.U., et al., A Patent Classification Model Based on Convolutional Neural Networks and Rand Forest, 2018.
[40] Kim J.H., K.S.J.I.P. Choi, and Management, Patent document categorization based on semantic structural information, vol.43, no.5, pp.1200-1215, 2007.
[41] Jiang J., using the graphic method to automatically extract the semantic information of Chinese patent texts, Library and Information Work vol.59, no.21, pp.115-122, 2015.
[42] Liao Lifa, Le Fugang, and Zhu J., Application of LDA Model in Patent Text Classification, Modern Intelligence vol.37, no.3, pp.35-39, 2017.
[43] Wu, C.H., K. Yun, and T.J.A.S.C.J. Huang, Patent classification system using a new hybrid genetic algorithm support vector machine, vol.10, no.4, pp.1164-1177, 2010.
[44] Ye, Z.F., B.L. Lu, and C. Hui, Patent Classification Using Parallel Min-Max Modular Support Vector Machine. 157-167. 2008.
[45] Liu, S.H., et al. Patent Classification Using Hybrid Classifier Systems, Advanced Materials Research. Trans Tech Publ, 2011.
[46] Jia Shanshan, et al., Patent Automated Classification Research Based on Multi-feature Multi-Classifier Integration, vol.1, no.8, pp.76-84, 2017.
[47] Cao J., A review of text clustering research, intelligence exploration no.1, pp.131-134, 2016.
[48] Qu Junwei, Qiao Xiaodong, and Gui J., Application Research of Self-Organizing Mapping in Patent Text Clustering, Digital Library Forum no.9, pp.13-19, 2010.
[49] Fan Yu, Fu Hongguang, and Wen J., Patent Information Clustering Technology Based on LDA Model, Computer Application vol.33, no.s1, pp.87-89, 2013.
[50] Tian J., Research on patent data clustering method based on M^3-DGMF, Computer Application and Software vol.30, no.3, pp.297-303, 2013.
[51] Chen Xu, et al., Patent Cluster Analysis Based on Technical Power Matrix, vol.35, no.3, pp.526-531, 2014.
[52] Lee W.J. and S.Y.J.E.P. S., Patent analysis to identify shale gas development in China and the United States, vol.74, no.74, pp.111-115, 2014.
[53] Niemann H., et al., Use of a new patent text-mining and visualization method for identifying patenting patterns over time: Concept, method and test application, vol.115, pp.210-220, 2016.
[54] Kim, G., J.J.T.F. Bae, and S. Change, A novel approach to forecast promising technology through patent analysis, vol.117, 2016.