IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 10, 2018, accepted September 3, 2018, date of publication September 17, 2018, date of current version October 12, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2870203

Prospecting Information Extraction by Text Mining Based on Convolutional Neural Networks–A Case Study of the Lala Copper Deposit, China

LI SHI[1,2], CHEN JIANPING[1,2], AND XIANG JIE[3]

[1] School of Earth Sciences and Resources, China University of Geosciences, Beijing 100083, China
[2] Land Resources Information Development and Research Key Laboratory of Beijing, China University of Geosciences, Beijing 100083, China
[3] MNR Key Laboratory of Metallogeny and Mineral Assessment, Institute of Mineral Resources, CAGS, Beijing 10037, China

Corresponding author: Chen Jianping (3s@cugb.edu.cn)

This work was supported by the Chinese MOST Methods and Models for Quantitative Prediction of Deep Metallogenic Geological Anomalies Project under Grant 2017YFC0601502.

ABSTRACT With geological big data becoming a focus of geoscience research, the vast amount of textual geoscience data provides both opportunities and challenges for data analysis and data mining. In fact, it does not seem possible to meet the demands of the big data age through the traditional manual reading for information extraction and gaining knowledge. In this paper, a workflow is proposed to extract prospecting information by text mining based on convolutional neural networks (CNNs). The aim is to classify the text data and extract the prospecting information automatically. The procedure involves three parts: 1) text data acquisition; 2) text classification based on CNN; and 3) statistics and visualization. First, the large amount of available text data was acquired based on geoscience big data acquisition methodologies. After text preprocessing, the CNN was used to classify the geoscience text data into four categories (geology, geophysics, geochemistry, and remote sensing), with each category consisting of three levels of text scales (word, sentence, and paragraph). Second, the word frequency statistics, co-occurrence matrix statistics, and term frequency–inverse document frequency (TF-IDF) statistics were for words, sentences, and paragraphs, respectively, which aimed to obtain the key nodes and links derived from the content-words. Finally, the deep semantic information of the big data mining of relevant geoscience texts was visualized by word clouds, knowledge graphs (e.g., the chord and bigram graphs), and TF-IDF statistical graphs. The Lala copper deposit in Sichuan province was taken as a test case, for which the prospecting information was extracted successfully by the developed text mining methodologies. This paper provides a strong basis for research into establishing mineral deposits prospecting models based on logical knowledge trees. In addition, it shows the great potential of this method for intelligent information extraction within geoscience big data.

INDEX TERMS Prospecting information, convolution neural networks, text mining, textual geoscience data, visual analysis.

I. INTRODUCTION

With the coming of the big data era, there have been a number of studies on big data applications in different areas such as business, healthcare, security, education, and so on [1]–[3]. The research into geoscience big data is gradually becoming an integral part of the national big data strategy. Geoscience research in the big data era requires us to collect as much geoscience exploration data (both structured data and unstructured technical reports, geological reports and papers) as possible. In recent years, open data initiatives have promoted governmental agencies and scientific organizations to publish data online for reuse [4], [5]. Geoscience literature is a key part of these open data and provides tremendous opportunities for further research. As the amount of geoscience text data increases, it becomes a pressing problem to effectively extract the relevant prospecting information by deeply analyzing, classifying and visualizing the massive amount of geoscience texts.

TABLE 1. Main processes and techniques of text mining.

Process	Techniques
Datasets	Text database, webpage, literature, e-book, etc.
Text Pre-processing	Noise elimination, word segmentation, POS tagging, self-defined stop-words removal, feature representation, feature extraction, etc.
Data Mining	Named entity recognition, word frequency analysis, sentiment analysis, automatic summarization, semantic network, similarity analysis, classification, clustering, association rule, intelligent retrieval, regression, trend analysis, etc.
Visualization	Knowledge and conclusion, graphical interface, command line, word cloud, knowledge graph, etc.

Text mining was proposed by Feldman in 1995 and has been defined as statistical text processing, text information extraction, natural language processing, and intelligent text analysis in different periods. The processes and techniques of text mining are summarized in Tab. 1, including datasets acquisition, text pre-processing, analysis and visualization. In terms of English-language text mining, there have been many achievements in text representation and feature extraction, text mining modelling, and algorithm development, which have been applied to various fields [6]–[10]. However, Chinese language models are quite different from English ones, because there is no space between words in the Chinese language, and it is difficult for computers to identify the meaningful word or phrase. Although there have been some advances in research into Chinese text mining, the processing of geoscience text in Chinese still faces a more difficult situation compared to other languages, as will be summarized below [11]–[13].

First, general word segmentation systems such as jieba (Chinese for ‘‘to shutter’’) cannot meet the word segmentation needs of geology. Second, geoscience text data is specific in compilation and description, as well as professional geoscience vocabulary. Traditional Chinese text representation tends to use high-frequency word eigenvalues as the text representation vector, which neither expresses the meaning clearly, nor meets the grammatical changes of geoscience text. In the meantime, it fails to extract prospecting information from geoscience text. Although Wang presented the hybrid corpus which was used to train the rules of Chinese word segmentation, then extracted and visualized key nodes and links derived from geoscience literature, this method was merely applied to one literature [14]. Therefore, there are still many problems in geoscience text mining, especially when dealing with massive amounts of text data, such as text acquisition, text preprocessing, and text classifying and visualization analysis, especially within the context of the Chinese language.

The concept of deep learning, which was proposed by Hinton in 2006, provides a new direction for text data mining [15]–[17]. With the development of deep learning, neural network methods have led to outstanding achievements in the areas of semantic analysis, and the searching and classified models in natural language processing (NLP) [18]–[22]. Convolutional Neural Networks (CNN), as a classification algorithm, has shown good robustness and has been successfully applied to NLP [23], [24]. Bengio *et al.* [25] constructed language models using CNN, thereby mapping word vectors to low-dimensional space. In the meantime, they measured the similarity between words by their distance in the vector space. Kim adopted CNN for sentence classification [26], while Liang *et al.* [27] discussed the feasibility of applying CNN to microblog sentiment analysis. Sun and He [28] used a CNN model to composite feature vectors, which were treated as semantic features, in order to train a multi-label classifier. Subsequently, Feng *et al.* [29] proposed a ranking-based multi-label convolutional neural network model (RM-CNN) to detect multi-label emotions successfully. In addition, some researchers have attempted to extract information from unstructured networks and to visualize text information into a network [30], [31]. Knowledge graphs as semantic networks with directed graph structures have been widely used to improve the search engines of Google, Baidu and Yahoo, and the graphs have also provided new ideas for the visualization of NLP [32]. Knowledge graphs represent a data organization form that expresses semantic relations between entities, concepts and the relationships between them. It is essentially a semantic network [33] that has been widely used in geoscience [34], [35]. Wang *et al.* [14] successfully applied knowledge graphs to visualize the key information extracted from geoscience texts, demonstrating the potential of NLP and knowledge graphs in geoscience research. Morrison *et al.* [36] used network analysis methods to display and explore the knowledge hidden amongst texts dealing with mineral species, localities and observations.

Based on previous research, this paper presents a workflow for extracting prospecting information by text mining based on CNN classification. The workflow introduced in Chapter II mainly consists of three parts: (1) text data acquisition and pre-processing; (2) multi-scale text classification based on CNN; and (3) statistical analysis and visualization.

The technical methods involved in these three parts are explained in detail in Chapter III. Text data acquisition and pre-processing involves obtaining the massive amount of available text data from public and local networks, and forming the datasets after the pre-processing (including data cleaning, format conversion, word segmentation, and the removal of self-defined stop-words). The self-defined stop-words do not include important words such as ‘‘is’’ or ‘‘isn't,’’ which are helpful when expressing semantic information. Multi-scale text classification based on CNN includes constructing word vectors with geoscience semantic and word-order information, and using CNN models to classify the textual geoscience data into four categories (geological data, geophysical exploration data, geochemical exploration data, and remote sensing data). Each type of data contains three text scales: the word-level, sentence-level and paragraph-level. Statistical analysis and visualization refers to the word frequency statistics, co-occurrence matrix statistics, and TF-IDF (term frequency–inverse document frequency) statistics of words, sentences and paragraphs, which aims to extract the content words (ruling out other stopwords) and the links between content words. In addition, word clouds, chords, and bigram graphs were selected to visualize the content-words and the links between prospecting information.

Finally, in Chapter IV, we take the LaLa copper deposit in Sichuan Province as a test case and successfully extract key prospecting information from multivariate, massive and heterogeneous geoscience big data using the presented methodology, which provides directions for the autonomous construction of prospecting models based on the logic tree of geology.

II. WORKFLOW

In order to extract prospecting information from unstructured geoscience text, an appropriate workflow was defined (Fig. 1). It was divided into three steps: text data acquisition, text classification, and statistics and visualization.

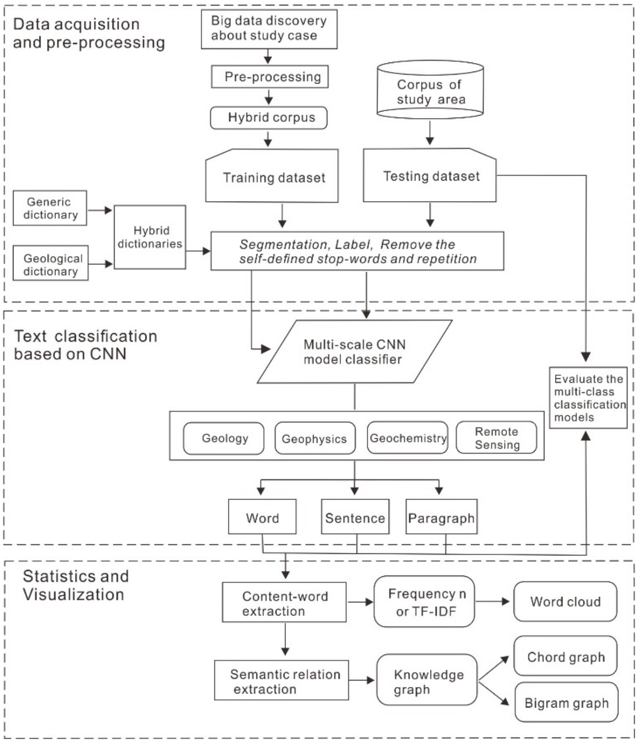

FIGURE 1. The technical workflow followed in this study, showing the three general processes employed (text data acquisition, text classification, and statistics and visualization).

A. DATA ACQUISITION AND PRE-PROCESSING

Big data acquisition is a necessary technical means for realizing data-to-information conversion and an essential step for the intelligent processing of text data. With the characteristics of large volumes, varieties, and velocity of geoscience text, it is necessary to construct logical structure trees and website structure trees according to geological thesaurus and correlational analyses. Only in this way can the data collection work be carried out completely. The hybrid corpus is built by combining a generic corpus and the geological corpus after data cleaning and format conversion, and then forming training datasets. Other necessary actions include word segmentation, tagging, the removal of self-defined stopwords, and repetitive word removal, which are processed on the basis of the hybrid dictionaries which combine the geological and general dictionaries. Similarly, word segmentation is performed on the test dataset of the Lala deposit.

B. TEXT CLASSIFICATION BASED ON CNN

Although CNN has been successfully applied to NLP in previous studies, it has been seldom employed for the classification of textual geoscience data. We selected the CNN classified model based on the TensorFlow to deal with massive geoscience text according to the characteristics of geoscience text and finish the processes of extracting the features, classifying, and mining the prospecting information. The CNN classified models are trained by the training sample sets of different scales, consisting of the preprocessed text data. In addition, the word vectors can also be trained. Then, the test datasets of the study area are classified and labeled into the four categories mentioned above with the three scales. Finally, the macro-averaging of the evaluation metrics (including the precision, recall and F1, explained in the following) of all the categories are applied to evaluate the multi-class classification algorithm.

C. STATISTICS AND VISUALIZATION

The classified and labeled text is counted and the results are visualized, which gives an overview of the acquired prospecting information from the geoscience text. We extract the content-words by frequency and TF-IDF. Then, the results are shown by word clouds, which consist of content-words with their frequencies greater than some threshold. In addition, the words in a sentence are not only controlled by grammar, but also restricted by word collocation. Fluency and meaning of sentences are realized based on words and their co-occurrence. Therefore, we can use the knowledge graph extracted by the co-occurrence of content-words to represent the key information. This can quickly give us a graphical view of a massive amount of text, so that we do not need to read word by word.

The workflow contributes to the convenient processing of geoscience text data, and promotes the extraction and presentation of the prospecting information from large amounts of geoscience text. Thus, it can better serve in mineral research, especially in prospecting model construction.

III. TEXT MINING METHODS

A. DATA ACQUISITION AND PRE-PROCESSING

We acquire the corpus from local area networks (LANs) and public domain networks (PDNs), following the workflow shown in Fig. 2. Data searching and filtering from the LANs

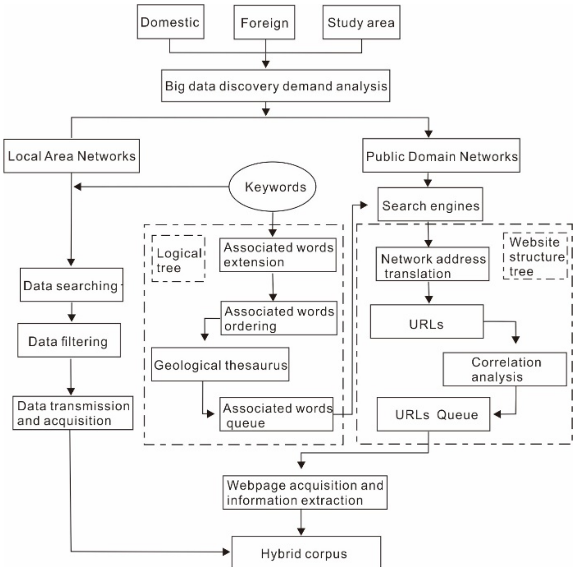

FIGURE 2. The workflow of the data acquisition process.

are based on the C# platform, which is used to develop the software EVERYTHING, and realized using the MySQL relationship database. The data is acquired by both P2P online transmission and FTP offline transmission. For the PDN data, we propose the bi-iteration method of Chinese keywords and URLs on the basis of the geologic thesaurus and URLs correlation analysis. We establish the logical tree based on the expansion according to the geological dictionary, so that we obtain many new keywords. Then, the seed URLs are formed by searching for these keywords using search engines such as Baidu and Google. In this process, new URLs are discovered after correlation analysis and URLs with strong correlations are continuously added to the website structure tree. The two branches are mutually iterative to form comprehensive search encirclement in two directions.

Text mining in Chinese requires some pre-processing work, including data cleaning, format conversion, and word segmentation. Wherein, (1) data cleaning is a process of removing tags and erroneous or duplicate URLs of the XML-formatted text data obtained through big data discovery. (2) Format conversion means converting different formats (e.g., PDF, CAJ, etc.) of geoscience documents to TXT through the C# platform. In the pre-filtering process, the documents are divided into different levels according to their relevance to the study area and weighted to form the hybrid corpus. (3) The methods of Chinese word segmentation were classified as dictionary-based, statistically-based, and hybrid. In this study, we combine the accurate segmentation carried out on the basis of the combination of the labeled geological dictionary with the general dictionary.

B. TEXT CLASSIFICATION BASED ON CNN

Text classification is the most critical part of text mining. We select CNN to classify and label massive amounts of unstructured geoscience text, based on the following considerations.

First, CNN reduces the number of network parameters and the complexity of the network itself. It is closer to biological neural networks. CNN can directly process original geological text, thereby avoiding the complex feature extraction and data reconstruction used in traditional algorithms, hence simplifying the pre-processing work.

Second, CNN has outstanding feature extraction capabilities. Much research has proven that CNN is effective when applied to this kind of Chinese text, with advantages in identifying and processing text data that has diverse expressions. The analysis of existing geoscience text data, general literature and reports are characterized by limited sentence length, rigorous language, and compact structures. They express meanings independently and have no complex or abstract literary expressions. Therefore, CNN is useful in identifying and extracting local features of such geoscience text data.

Third, CNN has a high level of computational efficiency. CNN, as one of the most widely used models in the field of deep learning, takes linear convolution as the core operation, which significantly increases the parallelism of the required computations. In the meantime, it uses hardware such as GPU (Graphics Processing Unit) to accelerate the processing, thereby greatly improving the computational efficiency. Above all, CNN has obvious advantages in classifying geoscience text data.

1) CNN CLASSIFICATION MODEL

In this study, the stop-words list is redefined to retain the original semantic information as far as possible, based on the comprehensive analysis of the structural and semantic features of geological big data text. We select the CNN model based on the open source TensorFlow[1] software library for analysis and for training the word vectors of the geological content-words by using massive geoscience text sample sets. The CNN model is a variant of a neural network that usually includes an input layer, convolutional layer, pooling layer, and fully connected layer. After a word vector lookup table W^e is established, the CNN-supervised classification model is trained by training sets. This paper takes the geoscience text data for the Lala deposit as the test sets to be classified, the process of which is divided into three cases. The output results are the classification results of the different scales (words, sentences and paragraphs) and the results include geology, geophysics, geochemistry, and remote sensing.

Fig. 3 shows a single convolutional layer architecture as an example, which was proposed by Kim [26] in 2014 with different widths of convolution windows to perform the CNN model training of sentence-level training dataset, while performing equally well with the paragraph-level one.

Sentence- and paragraph-level geological text sample sets are used as the initial input, and their feature information is passed to the multi-layer model for extraction.

[1] https://www.tensorflow.org/

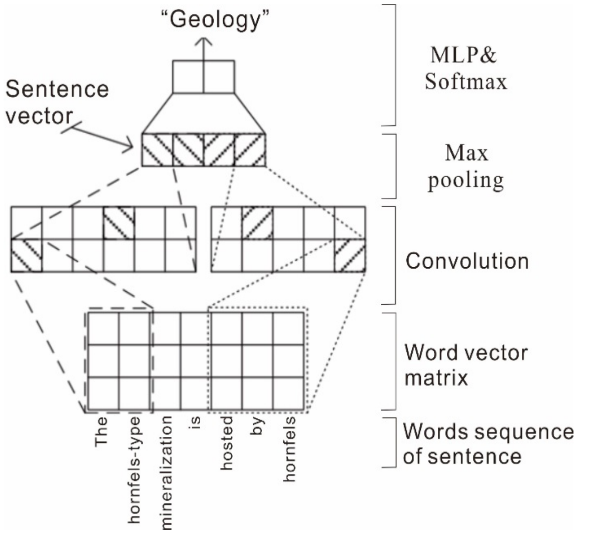

FIGURE 3. The structure of CNN model used in this work.

Important features of the text are extracted through different convolutions in different layers, the deeper the layer, the more abstract the extracted features would be. Each convolution layer (feature extraction layer) of the CNN is connected to a pooling layer (computation layer) that is used to search local averages (or maxima), which ensures the model has the appropriate sample fault tolerance and feature identification capability. The training process of the classification model based on paragraph-level sample set can be regarded as a training process of many long sentences that is roughly the same process as above.

The basic theories of text classification based on CNN may be summarized as follows. For text classification, the neurons in the CNN are usually arranged in two dimensions: width and height. The sizes of the input layer, convolutional layer, pooling layer, and output layer are therefore width × height. For example, when a sentence has 70 words and each word is represented by a 128-dimensional vector, the size of the input layer is 70 × 128.

Suppose that the number of words in the dictionary is N, the word vector dimension of a single word is d, and then the word vectors are found by lookup in a word-embedding matrix $W^e \in \mathbb{R}^{d\times N}$. If the geological sentence $S \in T$ is expressed as a matrix $S = (W_1, W_2, \cdots, W_k)$, where k is the maximum sequence length in the dataset (if the length is less than k, then the sentence is padded with zeroes), then W_j of the j^{th} word is represented by:

$$W_j = (\mathrm{w}_1, \mathrm{w}_2, \mathrm{w}_3, \ldots, \mathrm{w}_\mathrm{d}) \tag{1}$$

where d is the vector dimensionality (e.g., $d = 128$, to use the example above), and w_i is the random value of the i^{th} dimensions, $w_i \in (-0.1, 0.1)$, where $i \in (1, d)$.

To calculate the convolutional layer, we first concatenate all the words in sentence S as a long vector. Therefore, every sentence is represented by:

$$R^s = W_1 \oplus W_2 \oplus \cdots \oplus W_k \tag{2}$$

where $\oplus$ is the concatenation operator and $R^s \in \mathbb{R}^{d\times k}$ is the word vector matrix shown in Fig. 3.

Second, the convolution operation is performed on the sentence vector matrix R^s, assuming that the filter is $W^c \in \mathbb{R}^{d\times k_{win}}$, which is applied to produce a feature map. If k_{win} is the width of the convolution window, then we obtain:

$$\begin{aligned} r^f_{ij} &= \mathrm{g}(W^c_i \oplus Z_j + b^f_i) \\ Z_j &= r_{j-k_{win}+1} \oplus \cdots \oplus r_j \end{aligned} \tag{3}$$

where r^f_{ij} is the processing result of the i^{th} filter W^c_i at the word W_j, Z_j is the word vector matrix in window k_{win}, g (x) is the nonlinear transformation function, $\oplus$ is the inner product of the matrix, and b^f_i is the bias value of the i^{th} filter.

As mentioned above, we can utilize a single filter to transform a sentence matrix into a feature map. Actually, a set of filters that work in parallel can be applied to the sentence to generate multiple feature maps and give the sentence a richer representation. For example, in this study, there are three filters with differing window sizes. In this case, the set of filters forms a filter bank and produces three feature maps.

Thirdly, the outputs from the convolutional layer are then passed to the pooling layer, which aggregates the information and reduces the representation through common statistical methods, such as mean, maximum and the L2-norm.The pooling layer can alleviate the over-fitting problem and produce a vector of sentences with fixed lengths.

In this paper, we utilize the max-over-time pooling operation, which selects global semantic features and attempts to capture the most important ones with the highest value for each feature map, while obtaining the strongest excitation signal of each feature map. The computational results are as follows:

$$x_i = \max_{d_{win}\le j\le k} [r^f_{ij}], \quad 1 \le i \le H \tag{4}$$

where H is the number of convolution filters and $x \in \mathbb{R}^H$ is the resulting sentence vector corresponding to the filter.

Finally, the softmax operation is performed after the sentence vector x, such as $(a_1, a_2, a_2, a_4, \ldots)$, is mapped to $(b_1, b_2, b_2, b_4, \ldots)$, and $b_i (i = 1, 2, 3, \ldots) \in (0, 1)$. The output of the softmax function is the probability distribution of the four categories labels, namely "geology," "geophysics," "geochemistry" and "remote sensing."

2) CLASSIFICATION EVALUATION METRICS

The macro-averaging of the evaluation metrics (including the precision, recall and F1-score) of all the categories are used to evaluate the multi-class classification algorithm. In this paper, there are three kinds of text scales (word-level, sentence-level, paragraph-level) and each text scale is further classified into four categories. Thus, at each scale, we calculate the macro-averaging of three kinds of metrics based on the four classification confusion matrixes, as shown in Fig. 4.

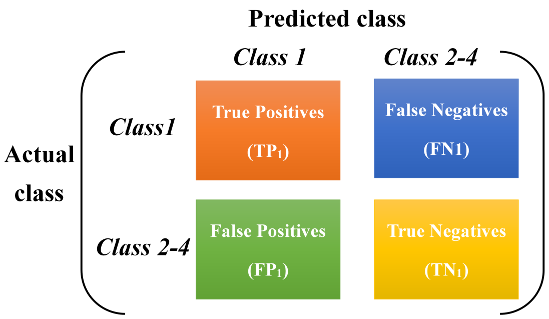

FIGURE 4. The calculation method of class 1's confusion matrix based on the whole classification confusion matrix in one scale.TP$_1$ (True Positives) is the number of testing texts that belong to class 1 and are correctly classified as class 1. FN$_1$ (False Negatives) is the number of texts that belong to class1, while being falsely classified as class 2-4. FP$_1$ (False Positives) represents the number of texts that belong to other classes and were falsely classified as class 1. TN$_1$ (True Negatives) represents the number of texts that belong to other classes and were correctly classified as belonging to them. Note, TN$_1$ is not used in this case.

Each class has three metrics, which are calculated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

The *precision* is used to represent the proportion of actual positive samples (e.g., class 1 is the positive class in this example) of the total number of samples that are classified into the positive class. *Recall* measures whether the classifier can identify all samples of this class. Both indexes should be taken into account in the study of geoscience text, and can be represented by $F1$, given by:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{7}$$

Finally, in macro-averaging, we average the performances of each individual class:

$$Macro\_P = \frac{1}{n}\sum_{i=1}^{n} P_i \tag{8}$$

$$Macro\_R = \frac{1}{n}\sum_{i=1}^{n} R_i \tag{9}$$

$$Macro\_F = \frac{1}{n}\sum_{i=1}^{n} F_i \tag{10}$$

where P_i is the precision of each category (namely geology, geophysics, geochemistry, and remote sensing), R_i is the recall of each category, F_i is the F1-score, and n = 4. *Macro_P, Macro_R, Macro_F* are the arithmetic means of the four categories.

C. STATISTICS AND VISUALIZATION

The text data classified by the CNN model is statistically analyzed to extract content-words and visualized to better present the found prospecting information, which is helpful for geological research. This mainly includes content word extraction and semantic relation extraction. The resulting prospecting information is then visualized by word clouds and knowledge graphs separately.

1) CONTENT-WORDS EXTRACTION AND VISUALIZATION

The content words represent the four categories discussed above of the prospecting information that are of interest in the geoscience text, which in turn mainly consist of four parts: terminology, technical methods, data processing methods and descriptive words [14] (see Tab. 2). Terminology refers to words related to geology, geophysics, geochemistry, remote sensing and the study area. Technical methods mainly describes words related to mineral exploration methods. Data processing refers to the specific data processing method adopted, and descriptive texts are essentially descriptive words.

TABLE 2. Classification of the content-words extracted from the geoscience-related text used in this paper.

Classification	Geology	Geophysics	Geochemistry	Remote Sensing
Terminology	Structure, Chalcopyrite	Aeromagnetic	ore-forming element	Texture, Image
Technical method	Zircon shrimp dating	IP Sounding	Superimposed Halo	Alteration information
Data processing method	Cluster	Derivation, Inversion	principal component analysis	Band fusion
Descriptive words	Nervation, Schistose	High-remanence	high-abundance	ring-shaped, linear

Although valuable information contained in geoscience text tends to be related to high-frequency content words [37], there are still some low-frequency words that are also important [38]. TF-IDF is a method of extracting low-frequency words that contain important information while evaluating the importance of a word to a certain document in the document sets. TF is the word frequency (occurrence frequency of a word in a document) while IDF is the inverse document frequency. In other words, a word holds a certain discriminative capability when it occurs many times in a document, but rarely in others. These content words can be named as keywords. Functional words, on the contrary, occur frequently in each document. Therefore, TF-IDF is proposed to effectively distinguish between keywords and functional words. The TF-IDF of the i^{th} word in the j^{th} document is expressed as:

$$TF_{i,j}, -IDF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \cdot ln\frac{|D|}{\left|\{j : t_i \in d_j\}\right|} \tag{11}$$

where $TF_{i,j}$ is the frequency of the i^{th} word in the j^{th} paragraph (document), |D| is the number of all documents, $\left|\{j : t_i \in d_j\}\right|$ is the number of documents that contain the i^{th} word, and $n_{i,j}$ is the number of the i^{th} word in the j^{th} paragraph (document).

We conduct frequency statistics on content words in word-level text and TF-IDF statistics in paragraph-level text. The word frequency of the top-ranked high-frequency words is plotted as shown in Fig. 5. The threshold n can be selected through manual intervention and the font size of words is directly proportional to the word frequency. High-frequency words (when the frequency is greater than the threshold n) are drawn in word clouds that enable us to read prospecting information from geoscience text simply and clearly.

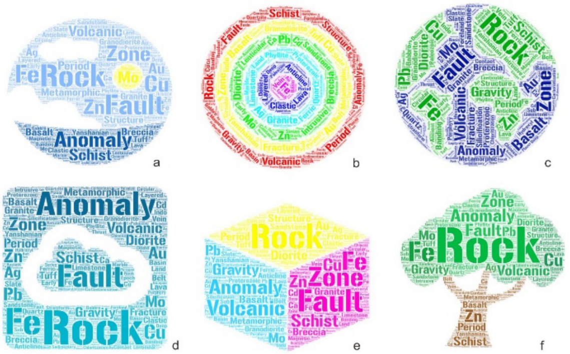

FIGURE 5. Examples of word clouds developed from words extracted from the geoscience text.

2) RELATION EXTRACTION AND VISUALIZATION

The key content words of geoscience text were extracted by the method described in the previous sections while the relationships between content words are extracted using the co-occurrence matrix. In text data, a sentence can be divided into content words and semantically ambiguous functional words [39]. Content words are the main entities in the Chinese corpus and serve as the carriers of key information, while functional words connect words into sentences and occur frequently. The words in a sentence are not only controlled by grammar, but are also restricted by word collocation, hence the fluency and meaning of sentences are realized based on words and their co-occurrence [40]. The precision of the geoscience text' classification shows that we cannot completely discard the key functional words, which is also the weakness of previous research. In this study, when training the CNN classification model, we only delete some of the functional words of the training sets, and retained a small portion of the key functional words. However, in the process of co-occurrence extraction, only content words are extracted and counted, and some functional words are properly discarded for clear expression. Two content words are co-occurrent when they are adjacent in the corpus. The co-occurrence preserves the orders of adjacent words and stores it in an $N * N$ matrix. In this paper, we statistically analyze the co-occurrence matrixes, and extract content words with high co-occurrence frequencies visualized by knowledge graphs.

A knowledge graph is a semantic network with directional structures, by which we can express the whole knowledge structure of the corpus vividly and reveal the dynamic development rules. Zhong *et al.* [41] constructed geographic names' co-occurrence networks on the basis of webpage text and realized the capacity to extract remarkable features or traffic features. As structured semantic knowledge bases, knowledge graphs include a series of nodes, edges and attributes. The basic model of knowledge graphs is generally ternary (entity-relationship-entity).

In this study, we used the NLP and knowledge graph (the bigram and chord graphs) methods to extract and visualize key information from unstructured geological Chinese corpus. This proved that the methods we selected for this paper are appropriate and powerful enough for extracting prospecting information from unstructured geological literature. We express content-words and the frequency of co-occurrence as nodes and edges of knowledge graphs. These are constructed on the basis of three variables: "from," "to" and "weight." Wherein, "from" represents the starting word and "to" represents the ending word, both of which are defined according to the order of the content words. "Weight" is determined by the co-occurrence frequency of two content words in the corpus. Fig. 6a is an example of a bigram graph, where edge arrows indicate word orders and point to the next content words. The figure displays key information in text data and the parameters on the edges are the co-occurrence frequency of content words in the text. Fig. 6b is an example of a chord graph, which describes the relationship between content words in the analyzed text. The chord width is scaled according to the content-words frequency with arbitrary colors.

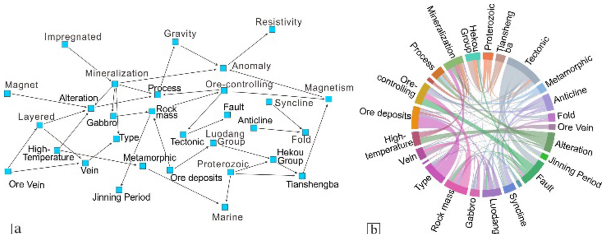

FIGURE 6. Examples of knowledge graphs: (a) bigram graph, (b) chord graph.

IV. EXPERIMENT

A. DATASETS CONSTRUCTION

This paper acquires corpus from LANs and PDNs, taking the Lala copper deposit, a volcanic-sedimentary hydrothermal copper deposit in the Sichuan Province, as an example. The logical structure tree is formed according to the genesis of the mineral deposit and the corpus of this research is established on the basis of the PDN and LAN data. The details of the datasets are given in Tab. 3.

Currently, there are some public datasets used for text classification, including the 20News, FuDan, and SST dataset. However, there are no dataset specifically designed for the classification of geoscience text. Therefore, this paper classifies the corpus of "volcanic-sedimentary hydrothermal cop-

TABLE 3. Statistics of the corpus collection.

	PDNs		LANs		
	Webpages	Papers	Reports of mineral exploration	Monographs	Papers
Numbers	14508	1590	17	15	98
Sum	16098		130		

per deposit'' collected from PDNs and LANs into the discussed four categories for normalized labeling.

The training sets employed during the research consists of 3,205 labeled sentence-level samples and 2,283 labeled paragraph-level samples. 1,000 samples are randomly selected from the training sets in each training session, 70% of which is used as the training sets and 30% used as the validation sets. Each sentence-level sample contains an average of 18.6 content words while each paragraph-level sample has an average of 76.8 content words, which indicates that the training sets includes a total of 2.35 × 105 content words. The dictionary size is 32,066. The testing sets of this experiment is composed of 400 word-level, 400 sentence-level and 400 paragraph-level samples randomly selected from the filtered corpus of ''Lala copper deposit in Sichuan Province'' (see Tab. 4).

TABLE 4. Statistics describing the numbers of each category and levels in the training and testing sets.

	Levels	Geology	Geophysics,	Geochemistry	Remote sensing	Sum
Training sets	Sentence	1862	579	425	339	3205
	Paragraph	1092	499	400	292	2283
Testing sets	Word	100	100	100	100	400
	Sentence	100	100	100	100	400
	Paragraph	100	100	100	100	400

B. EXPERIMENT RESULTS

The optimal parameter combination is obtained by experimental comparison and analysis. The maximum length of the sentence- and paragraph-level samples is set to 216 bytes and 766 bytes, respectively (samples that exceed the maximum length are cut off and those less than the threshold are padded with zeroes). The word vector dimension $d = 128$ and the filter window size $k_{win} = 3, 4, 5$. Stochastic gradient descent is adopted for the updated iteration of the weights in accordance with the preset number of cycles (500) and model validation is performed every 10 iterations, so as to train the optimal classification model.

TABLE 5. Comparisons of the accuracy and the loss of models trained by the datasets with different operation of stop-words.

Operation of stop-words	Sentence/Paragraph	Training/ Validating	Accuracy	Loss
Not remove stop-words	Sentence model	Training	0.9990	0.0122
		Validating	0.8861	0.4060
	Paragraph model	Training	0.9950	0.0133
		Validating	0.8993	0.5611
Remove Jieba stop-words	Sentence model	Training	0.9970	0.0125
		Validating	0.8787	0.4209
	Paragraph model	Training	0.9940	0.1806
		Validating	0.8934	0.5364
Remove self-defined stop-words	Sentence model	Training	0.9990	0.0007
		Validating	0.9008	0.3307
	Paragraph model	Training	0.9980	0.0071
		Validating	0.8803	0.5967

In this study, the accuracy and the loss of the models are compared respectively when the stop-words are not removed, the Jieba stop-words are removed, and the self-defined stop-words are removed (see Tab. 5). It is found that the models trained by the datasets from which the self-defined stop-words are removed possess higher training and validation accuracy, especially the sentence classification model, whose validation accuracy is up to 90.08%, 2.21% higher than the Jieba one and 1.47% higher than that not being removed. Furthermore, the loss of the sentence models is lower than that of the others in general. However, the validation accuracy and the loss of the paragraph seem to have no advantage under the operation of self-defined words removal. Judging from the overall performance of the models, the CNN classification model based on multi-scale training sets with the self-defined stop-words removed that are more effective in classifying the sentence-level geoscience text.

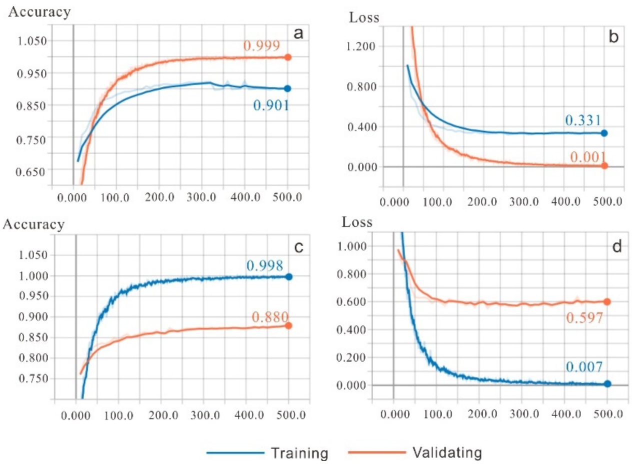

FIGURE 7. The performance contrast of two CNN models trained by multi-scale training sets according to the accuracy and the loss in the training and validating processes; (a, b) the accuracy and the loss of the model trained by sentences; (c, d) the accuracy and the loss of the model trained by paragraphs.

As shown in Fig. 7, it enjoys the classification accuracy of 99.9% and 99.8% at the sentence and paragraph levels, respectively, achieving the optimal results in both cases. The validation accuracy of the two models reaches their peak at 500 iterations and yields an approximately optimal result (90.08% and 88.03%, respectively). The classification loss of these two models tends to converge with the increase in the number of iterations. In addition, there is no overfitting. Training losses of the two models are 0.07% and

0.71% accordingly, and validation losses are 33.07% and 59.67%, respectively, which indicates that the classification loss increases with the increase in text length. Therefore, the classification model employed in this paper is more suitable for the classification of the sentence-level text. However, although the classification effect on the paragraph-level text is slightly inferior, the difference is insignificant.

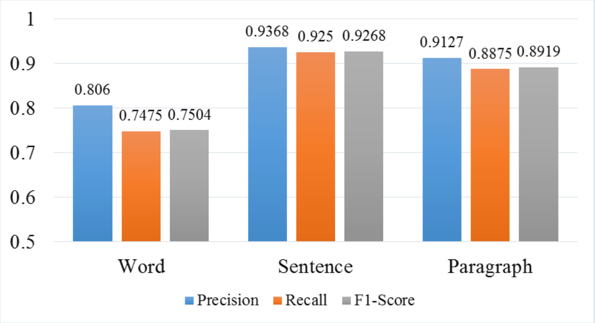

FIGURE 8. The contrast of the macro-average of precision, recall and F1-Score metrics of the classification results concerning different level testing sets.

We extract 400 test data from the testing sets for classification testing and show the comparison of precision, recall and F1- score of words, sentences and paragraphs in Fig. 8. The precision of the words, sentences and paragraphs extracted from the geoscience text are 80.6%, 93.68% and 91.27%, respectively, while the recall rates are 74.75%, 92.5% and 88.75%, respectively. According to the diagram, sentences enjoy the highest classification precision, followed by paragraphs. The 2.4% difference in precision indicates that the model is also effective in classifying long Chinese geoscience text, including the sentence-level and the paragraph-level. The classification precision at the word-level is the lowest, which is mainly due to the following reasons: although the word vector trained by the sentence-level training set preserves the semantic relations of the geoscience text, there are still some different category words existing in the same sentence. As a result, ambiguity occurs when the model is used for word-level classification because it may contradict the training sentences, and eventually lead to negative precision in the word-level classification. In addition, in the sentence training process of the CNN's classification model, semantic association is retained, while the words are relatively independent with no semantic correlation, which is different from the sentence at this point. Thus, the classification precision of the word-level testing set is lower. However, in general, the classification model based on CNN is effective in classifying multi-scale Chinese geoscience text.

C. VISUALIZATION OF THE PROSPECTING INFORMATION

The original testing corpus is automatically divided into four kinds of prospecting information after being classified by CNN, each of which contains three text scales. In view of the different scales of text, this research uses multivariate statistics and visualization techniques. It performs word frequency statistics on the word-level text and TF-IDF statistics on the paragraph-level text to extract content words, and realizes the extraction of semantic relations by co-occurrence matrix statistics on the sentence-level text. In addition, word clouds and knowledge graphs (bigram graphs and chord graphs) are employed to visualize the extracted prospecting information from the text mining of Chinese geoscience text, so that we can mine features and semantic relations more designedly and profoundly.

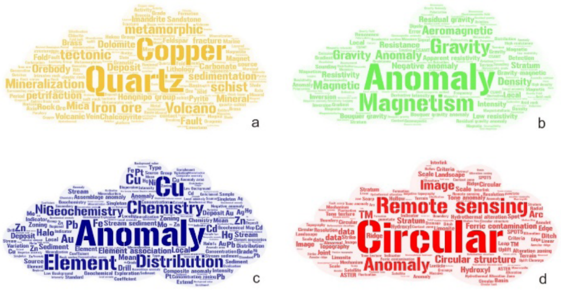

FIGURE 9. The word clouds corresponding to the Lala copper deposit: (a) geology; (b) geophysics; (c) geochemistry; (d) remote sensing. The font size of the word is scaled according to its frequency in the whole testing set.

Content-words frequency statistic is performed on the word-level classification results, which can extract the prospecting information from the four kinds of corpus. We then visualize the keywords which have a high frequency by producing word clouds as shown in Fig.9. Previously when faced with the massive amount of text data that is available, we usually found the important information by frequency statistics under the situation where we don't distinguish between geology-geophysics-geochemistry-remote sensing and other words. Thus, we usually overshadowed some important prospecting information. Word frequency statistics of the classified word-level text, however, makes the research more pertinent and makes it easier to extract effective prospecting information.

For example, copper deposits in the case study area are mainly related to quartz volcanic rocks (Fig. 9a), while the Lala copper deposits are characterized by strong gravity and magnetic anomalies (Fig. 9b). It can be seen from Fig. 9c that geochemistry focuses on the Cu anomalies and element assemblage anomalies (Fig. 9c). According to Fig. 9d, ore-forming conditions can be determined in accordance with the shape, color, topography and altered information on remote sensing images.

TF-IDF statistics were performed on the classification results of the paragraph-level text and the top ten content words were obtained (see Fig. 10). According to the comparison of word frequency and TF-IDF statistics, TF-IDF statistics extracts geological terminology (e.g., "chalcopyrite," "iron-copper" and "bedded-structure") more pertinently, while word frequency statistics extracts universal words (e.g., "quartz," "schist," "volcano" and "copper mine")

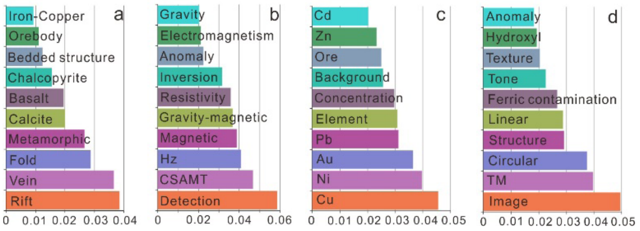

FIGURE 10. Top 10 content-words extracted by TF-IDF method: (a) geology; (b) geophysics; (c) geochemistry; (d) remote sensing.

(Fig. 10a) more effectively. TF-IDF statistics has obvious advantages in the extraction of technical and data processing methods (e.g., ‘‘inversion’’ and ‘‘CSAMT’’) in the content word extraction of geophysics (Fig. 10b), and effective in the element extraction of geochemistry (Fig. 10c). Meanwhile, it is also effective in extracting descriptive words dealing with remote-sensing text (Fig. 10d), although this is due to the strong universality of the remote-sensing text. Therefore, word-level prospecting information extracted using these two methods are generally consistent.

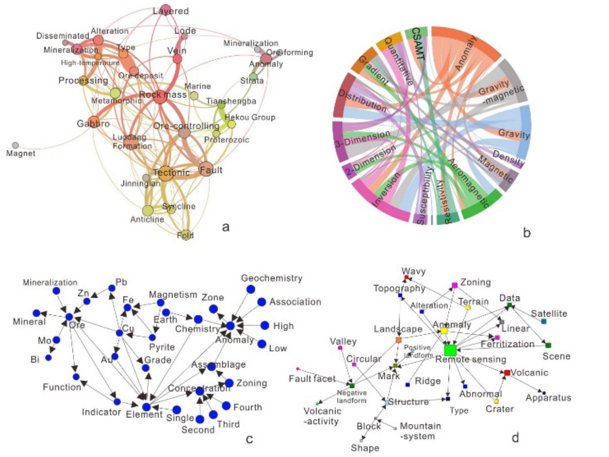

FIGURE 11. Visualization of knowledge graphs: (a) bigram graph of geology, (b) chord graph of geophysics; (c) bigram graph of geochemistry; (d) bigram graph of remote sensing.

Semantic relations of sentence-level text were extracted using co-occurrence matrix statistics and knowledge graphs of the Lala copper deposit prospecting information, which include geological, geophysical, geochemical, and remote sensing information, as shown in Fig. 11. The bigram graph of geology in Fig.11a shows the geological and mineral resource characteristics of the Lala copper deposit. For example, it can be seen from the figure that content-words such as ‘‘structure,’’ ‘‘rock mass,’’ ‘‘fault’’ and ‘‘Luodang formation’’ have high centrality, which reveals that the major ore-controlling factors include rock mass, fault structure and the Luodang formation of gabbro in the area of the Lala copper deposit. Figs. 11b and 11c show the chord graph of geophysics and the bigram graph of geochemistry, respectively, where ‘‘anomaly’’ has high centrality and frequency. ‘‘Anomaly’’ extracted from four kinds of geoscience text tends to be an important aspect of mineral deposit exploration. Geophysical anomalies of the Lala copper deposit mainly include aeromagnetic anomalies, gravity anomalies and CSAMT resistivity anomalies, while geochemical anomalies mainly refer to element assemblage anomalies such as Cu-Fe, Cu-Au and Cu-Pb-Zn. According to Fig. 11d (bigram graph of remote sensing), remote sensing is highly centralized and linked to words such as ‘‘satellite,’’ ‘‘marker,’’ and ‘‘scene’’ etc. The interpretation information obtained from the remote sensing images mainly includes textural, topographic, geomorphic, and spectral information.

Above all, this paper effectively and precisely mines and describes the prospecting information and connections extracted by CNN classifying while visualizing the prospecting information extracted from the geoscience corpus. It provides a sound basis for the associated analysis of geoscience data and the construction of prospecting models.

V. CONCLUSIONS AND FUTURE WORK

A workflow is proposed in this paper to extract key prospecting information from geoscience text data by text mining based on CNN classification. Taking the Lala copper deposit in Sichuan Province as an example, the paper completes the intelligent classification and labeling of Chinese text big data, explores the potential relations between data, and realizes the intelligent extraction of geological prospecting information. It provides a powerful basis for the automatic construction of prospecting models based on geological big data and further prospecting deployments. The following comments may be made:

(1) In the face of the massive amount of multivariate and heterogeneous PDN data, this paper develops a bi-iteration data discovery system and data pre-processing system based on keywords and URLs, thereby effectively solving technical problems for PDN data discovery and acquisition from geological big data: insufficient comprehensiveness, poor correlation and inconsistent formats.

(2) The research results show that the classification model proposed is effective in classifying multi-scale geoscience text. The classification models trained for words, sentences and paragraphs achieved better results in the classification of long text. This method effectively prevents important information from being overwhelmed by massive data and provides an appropriate workflow for the automatic extraction of geological prospecting information. In addition, we also propose a self-defined stop-word dictionary, which can retain semantic information as far as possible, thus improving the classification precision of geoscience text.

(3) TF-IDF statistics and co-occurrence matrix statistics are performed in the data mining process to extract content-words and semantic relations separately. In addition, word clouds, bigram graphs, and chord graphs were employed to vividly show keywords and semantic relations of the four kinds of prospecting information, thereby effectively and pertinently representing prospecting information

contained in the geoscience text. In this case, the research results are more scientific and reliable.

In the future, the four kinds of prospecting information can be further refined to solve the automatic construction of prospecting models based on geological big data. For example, the prospecting information can be divided into rock mass conditions, formation conditions, and structural conditions, while rock mass conditions can be further refined into wall-rock alteration and ore-bearing rock series. In addition, wall-rock alteration can be classified into chloritization, siliconization, carbonatization and sericitization. The extraction of prospecting information on the basis of geological knowledge trees requires the training sets to be more exact, and needs higher classifier precision, which is the next step to be followed.

REFERENCES

[1] J. Wu, S. Guo, J. Li, and D. Zeng, "Big data meet green challenges: Big data toward green applications," *IEEE Syst. J.*, vol. 10, no. 3, pp. 888–900, Sep. 2016.

[2] J. Wu, S. Guo, J. Li, and D. Zeng, "Big data meet green challenges: Greening big data," *IEEE Syst. J.*, vol. 10, no. 3, pp. 873–887, Sep. 2016.

[3] J. S. Peng and Y. M. Shao, "Intelligent method for identifying driving risk based on V2V multisource big data," *Complexity*, vol. 2018, no. 1, pp. 1–9, May 2018.

[4] L. Cernuzzi and J. Pane, "Toward open government in Paraguay," *IT Prof.*, vol. 16, no. 5, pp. 62–64, Sep./Oct. 2014.

[5] X. Ma, "Linked geoscience data in practice: Where W$_3$C standards meet domain knowledge, data visualization and OGC standards," *Earth Sci. Inform.*, vol. 10, no. 4, pp. 429–441, Dec. 2017.

[6] B. Lent, R. Agrawal, and R. Srikant, "Discovering trends in text databases," in *Proc. 3rd Int. Conf. Knowl. Discovery Data Mining (KDD)*, Jul. 1997, pp. 227–230.

[7] G. P. C. Fung, J. X. Yu, and W. Lam, "Stock prediction: Integrating text mining approach using real-time news," in *Proc. IEEE Int. Conf. Comput. Intell. Financial Eng.*, Mar. 2003, pp. 395–402.

[8] J. Patrick, "The Scamseek project—Text mining for financial scams on the Internet," in *Data Mining*, G. J. Williams and S. J. Simoff, Eds. Berlin, Germany: Springer, Jan. 2006, pp. 295–302.

[9] A. Porter and I. Rafols, "Is science becoming more interdisciplinary? Measuring and mapping six research fields over time," *Scientometrics*, vol. 81, no. 3, pp. 719–745, Dec. 2009.

[10] R. J. Byrd, S. R. Steinhubl, J. Sun, S. Ebadollahi, and W. F. Stewart, "Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records," *Int. J. Med. Inform.*, vol. 83, no. 12, pp. 983–992, Dec. 2014.

[11] H. F. Lin *et al.*, "Visualization model for Chinese text mining," *Comput. Sci.*, vol. 27, no. 4, pp. 37–41, Jan. 2000.

[12] Z.-Q. Chen and G.-X. Zhang, "Study on the text mining and Chinese text mining framework," *Inf. Sci.*, vol. 25, no. 7, pp. 1046–1051, Aug. 2007.

[13] T L. Deng, "Analysis and implementation of text mining for call centers with big data," Beijing Univ. Posts Telecommun., Beijing, China, Tech. Rep., 2015.

[14] C. B. Wang, X. Ma, J. Chen, and J. Chen, "Information extraction and knowledge graph construction from geoscience literature," *Comput. Geosci.*, vol. 112, pp. 112–120, Mar. 2018.

[15] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[16] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 19, no. 1, Jan. 2007, pp. 153–160.

[17] R. G. Guimaraes, R. L. Rosa, D. De Gaetano, D. Z. Rodriguez, and G. Bressan, "Age groups classification in social network using deep learning," *IEEE Access*, vol. 5, pp. 10805–10816, May 2017.

[18] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press, 1999.

[19] W.-T. Yih, K. Toutanova, J. C. Platt, and C. Meek, "Learning discriminative projections for text similarity measures," in *Proc. 15th Conf. Comput. Natural Language Learn. Assoc. Comput. Linguistics*, Jun. 2011, pp. 247–256.

[20] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil, "Learning semantic representations using convolutional neural networks for Web search," in *Proc. 23rd Int. Conf. World Wide Web*, Apr. 2014, pp. 373–374.

[21] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, May 2014, pp. 1749–1751.

[22] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *J. Mach. Learn. Res.*, vol. 12 pp. 2493–2537, Aug. 2011.

[23] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[24] T. Lei, R. Barzilay, and T. Jaakkola, "Molding CNNs for text: Non-linear, non-consecutive convolutions," *Indiana Univ. Math. J.*, vol. 58, no. 3, pp. 1151–1186, Aug. 2015.

[25] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *J. Mach. Learn. Res.*, vol. 3, pp. 1137–1155, Feb. 2003.

[26] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Oct. 2014, pp. 1746–1751.

[27] J. Liang *et al.*, "Deep learning for Chinese micro-blog sentiment analysis," *J. Chin. Inf. Process.*, vol. 28, no. 5, pp. 155–161, 2014.

[28] S. Sun and Y. He, "Multi-label emotion classification for microblog based on CNN feature space," *Adv. Eng. Sci.*, vol. 49, no. 3, pp. 162–169, May 2017.

[29] S. Feng, Y. Wang, K. Song, D. Wang, and G. Yu, "Detecting multiple coexisting emotions in microblogs with convolutional neural networks," *Cogn. Comput.*, vol. 10, no. 1, pp. 136–155, Feb. 2018.

[30] K. Ehrlich, C.-Y. Lin, and V. Griffiths-Fisher, "Searching for experts in the enterprise: Combining text and social network analysis," in *Proc. Int. ACM Conf. Supporting Group Work*, Jan. 2007, pp. 117–126.

[31] D. Paranyushkin, "Identifying the pathways for meaning circulation using text network analysis," Nodus Labs, Berlin, Germany, Tech. Rep., Dec. 2011.

[32] M. Schuhmacher and S. P. Ponzetto, "Knowledge-based graph document modeling," in *Proc. 7th ACM Int. Conf. Web Search Data Mining*, Feb. 2014, pp. 543–552.

[33] F. Lu, L. Yu, and P. Y. Qiu, "On geographic knowledge graph," *J. Geo-Inf. Sci.*, vol. 19, no. 6, pp. 723–734, Jun. 2017.

[34] J. Xu, T. Pei, and Y. Yao, "Conceptual framework and representation of geographic knowledge map," *J. Geo-Inf. Sci.*, vol. 12, no. 4, pp. 496–502, Aug. 2010.

[35] Z. W. Hou, Y. Q. Zhu, and Y. Gao, "Geologic time scale ontology and its applications in semantic retrieval," *J. Geo-Inf. Sci.*, vol. 20, no. 1, pp. 17–27, Jan. 2018.

[36] S. M. Morrison *et al.*, "Network analysis of mineralogical systems," *Amer. Mineralogist*, vol. 102, no. 8, pp. 1588–1596, Aug. 2017.

[37] E. Hovy and C.-Y. Lin, "Automated text summarization and the SUMMARIST system," in *Proc. Workshop Held Baltimore*, Baltimore, MD, USA, Oct. 1998, pp. 197–214.

[38] S. T. Piantadosi, "Zipf's word frequency law in natural language: A critical review and future directions," *Psychonomic Bull. Rev.*, vol. 21, no. 5, pp. 1112–1130, Oct. 2014.

[39] C. C. Fries, *The Structure of English: An Introduction to the Construction of English Sentences*. New York, NY, USA: Harcourt, Brace & World, 1952.

[40] J. R. Firth, "A synopsis of linguistic theory 1930–1955," in *Studies in Linguistic Analysis*. Oxford, U.K.: Blackwell, 1957.

[41] X. Zhong, Y. Gao, and L. Wu, "Extract core toponyms from Web page text based on link analysis," *J. Geo-Inf. Sci.*, vol. 18, no. 4, pp. 435–442, Apr. 2016.

LI SHI was born in Shenyang, Liao Ning, China, in 1994. She received the B.S. degree in remote sensing of environmental resources from Capital Normal University, Beijing, in 2016. She is currently pursuing the Ph.D. degree, the master's doctor combined program, with the China University of Geosciences, Beijing, China.

Since 2016, she has been a Ph.D. Student with the Land and Resources Information Research and Development Key Laboratory, Beijing. She has co-authored one book and two articles. Her research interests include big data analysis and text information mining.

CHEN JIANPING was born in Beijing, China, in 1959. He received the B.S. and M.S. degrees in geoscience and the Ph.D. degree in remote sensing of environmental resources from the Chengdu University of Technology, Sichuan, in 1995 and 1988, respectively.

From 1982 to 1993, he was a Research Assistant with the Remote Sensing Geology Laboratory, Chengdu University of Technology. From 1993 to 1997, he was an Associate Professor with the Chengdu University of Technology. Since 1997, he has been a Full Professor with the Earth Science and Resources Department, China University of Geosciences, Beijing. He has authored 30 books, over 120 articles, and over 10 inventions. His research interests include mineral resources assessment and big data analysis.

Dr. Chen was a recipient of the International Association for Mathematical Geosciences Scientist Award for Excellence in 2009 and the Academician of the Russian Academy of Sciences in 2015.

XIANG JIE was born in Changde, Hunan, China, in 1990. He received the B.S. degree in geoscience from the Hunan University of Science and Technology in 2008, and the M.S. and Ph.D. degrees in geodetection and information technology from the Land and Resources Information Research and Development Key Laboratory, China University of Geosciences, Beijing, China, in 2018.

From 2016 to 2017, he was a Visiting Scholar with the University of Padova, Padua, Italy. He has co-authored three books, nine articles, and one invention. His research interests include mineral resources assessment and mapping the topographic fingerprints of humanity across earth.

Dr. Xiang was a recipient of the Best Presentation Award from the European Geosciences Union General Assembly in 2017.