Contents lists available at ScienceDirect

Computers and Geosciences

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

Research paper

Information extraction and knowledge graph construction from geoscience literature

Chengbin Wang [a,b], Xiaogang Ma [b,*], Jianguo Chen [a,**], Jingwen Chen [a]

[a] State Key Laboratory of Geological Processes and Mineral Resources & Faculty of Earth Resources, China University of Geosciences, Wuhan 430074, China
[b] Department of Computer Science, University of Idaho, Moscow ID 83844, USA

ARTICLE INFO

Keywords:
Geological corpus
Knowledge graph
Geoscience literature
Chinese word segmentation
Chord and bigram graphs

ABSTRACT

Geoscience literature published online is an important part of open data, and brings both challenges and opportunities for data analysis. Compared with studies of numerical geoscience data, there are limited works on information extraction and knowledge discovery from textual geoscience data. This paper presents a workflow and a few empirical case studies for that topic, with a focus on documents written in Chinese. First, we set up a hybrid corpus combining the generic and geology terms from geology dictionaries to train Chinese word segmentation rules of the Conditional Random Fields model. Second, we used the word segmentation rules to parse documents into individual words, and removed the stop-words from the segmentation results to get a corpus constituted of content-words. Third, we used a statistical method to analyze the semantic links between content-words, and we selected the chord and bigram graphs to visualize the content-words and their links as nodes and edges in a knowledge graph, respectively. The resulting graph presents a clear overview of key information in an unstructured document. This study proves the usefulness of the designed workflow, and shows the potential of leveraging natural language processing and knowledge graph technologies for geoscience.

1. Introduction

Research of mathematical geoscience focuses on processing georeferenced quantitative data (e.g., rock parameters, geochemical tests, geophysical surveys and satellite imagery) for discovering new knowledge (e.g., Cracknell and Reading, 2014; Lima et al., 2017; Wang et al., 2016a, 2016b; Xiao et al., 2016). For a geological topic, we can discover new information by data analysis and geological interpretation, and enrich our understanding by comparing and connecting relevant works. Findings generated from geoscience studies are often recorded in technical reports, journal papers, books and other types of literature. In recent years, open data initiatives have promoted governmental agencies and scientific organizations to publish data online for reuse (Cernuzzi and Pane, 2014; Ma, 2017). For example, many national geological survey agencies (e.g., USGS and CGS) have published outputs of geological investigation online. Geoscience literature is a key part of those open data and provides tremendous opportunities for further research. Extracting structured information and discovering knowledge from textural geoscience data is a topic that has been less studied in mathematical geoscience and desires more work. In particular, the processing of geoscience literature in Chinese faces a harder situation, because there is no space between words in the Chinese language, and it is difficult for a computer to identify the boundary of a meaningful word or phrase in Chinese (Gao et al., 2005; Huang et al., 2015).

The fast-growing technologies in natural language processing (NLP) (Manning and Schütze, 1999) and knowledge graph representation (Singhal, 2012) can be adapted for textual geoscience data processing. Traditional methods of document representation are often based on the vector space model with the drawbacks of missing semantic links between words and lacking word sequence information (Turney and Pantel, 2010). The knowledge graph proposed by Google provides a new perspective and corresponding techniques for representing knowledge in a document (Schuhmacher and Ponzetto, 2014). Using that approach, a document can be represented by a list of entities within the document and their semantic links rather than character strings. Recently, Giboin et al. (2013) used description-oriented and document-oriented methods to build ontologies for text data. Peters and his colleagues (Peters et al., 2014; Peters and McClennen, 2015) used NLP method to extract structured paleontological

[*] Corresponding author.
[**] Corresponding author.
E-mail addresses: max@uidaho.edu (X. Ma), jgchen@cug.edu.cn (J. Chen).

https://doi.org/10.1016/j.cageo.2017.12.007
Received 29 August 2017; Received in revised form 10 December 2017; Accepted 15 December 2017
Available online 16 December 2017

information from millions of documents and developed the Paleobiology Database (https://paleobiodb.org). Morrison et al. (2017) used network analysis methods to display and explore the hidden knowledge among mineral species, localities and observations.

Text mining for literature in Chinese requires some pre-processing work. A necessary step is word segmentation. The methods of Chinese word segmentation were classified as dictionary-based, statistically-based, and hybrid (Gao et al., 2005). Most recently, Huang et al. (2015) used the Conditional Random Fields to segment geological documents written in Chinese. In their work, a self-defined geological corpus was used to train the segmentation rules and improve the quality of results.

This paper presents a workflow and a few empirical case studies for extracting information and building knowledge graphs from geoscience literature, with a focus on documents in Chinese. In this study, we first built a hybrid corpus consisting of the generic terms and geological terms to improve the performance of word segmentation. The geological terms were labeled based on a *geology dictionary* (Ministry of Geology and Mineral Resources, 2005) and *The Terminologies and Classification Codes of Geology and Mineral Resources* (TCCGMR) (Wang et al., 1999; Ma et al., 2010), a widely-used data standard in China. Second, the hybrid corpus was used to train the rules of Chinese word segmentation in the Conditional Random Fields model. Third, we removed the stop-words from the word segmentation results and analyzed the word frequency distribution. Forth, we extracted and visualized key nodes and the links derived from the content-words for representing the key information of geoscience literature. To our knowledge, this is the first study to extract information and build knowledge graphs from unstructured geoscience literature in Chinese.

2. Methods

2.1. Conditional Random Fields model

Conditional Random Fields (CRF) model is a typical discriminative model. At first, it was proposed for processing sequence data based on the maximum entropy and the hidden Markov model (Lafferty et al., 2001). It has been widely used in the domains of natural language processing and bioinformatics (e.g., Huang et al., 2015; McCallum and Li, 2003; McDonald and Pereira, 2005; Pinto et al., 2003; Sato and Sakakibara, 2005). In the Chinese word segmentation, CRF model addresses label-bias of the maximum entropy model and has a higher accuracy than the hidden Markov model (Lafferty et al., 2001; Pinto et al., 2003).

CRF model can be regarded as an undirected graph model with the conditional probability of a node at given nodes. Given a graph $G = (V, E)$, V is set of nodes, E is the set of undirected boundaries.

$$Y = \{Y_v | v \in V\} \tag{1}$$

where Y_v denotes random variables corresponding to the node $v \in V$. If Y_v obeys the Markov property:

$$P(Y_v | X, Y_w, w \neq v) = P(Y_v | X, Y_w, w \sim v) \tag{2}$$

(X, Y) is a conditional random field, where X denotes an observed sequence, $w \sim v$ indicates all the neighbor nodes of w connected with node v in G graph.

For Chinese word segmentation, observed sequence X of Chinese character and label set Y have the same graphical structure of a chain-structured Conditional Random Field (Ruokolainen et al., 2013; Xu et al., 2008; Zhang, 2008) (Fig. 1). The label set is used to mark the word boundary in a sentence (Xue, 2003). Based on the Chinese word-building, we used "B" to mark the beginning character of a word, used "M" to mark the middle character, used "E" to mark the ending characters. For the word with single character, we used "S" to mark it. Y is a set consisting of B, E, M and S, which has the same length with the observing sequences X.

The probability of candidate label of Y can be obtained by equation (3):

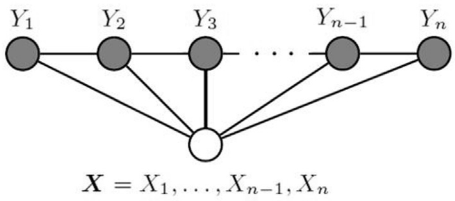

Fig. 1. Graphical structure of a chain-structured Conditional Random Field (Wallach, 2004).

$$P(Y|X) = \frac{1}{Z(x)} \exp\left(\sum_{i,k} \lambda_k f_k(y_{i-1}, y_i, x) + \sum_{i,k} \alpha_k g_k(y_i, x)\right) \tag{3}$$

where $Z(X)$ is normalization factor, and can be represented by equation (4)

$$Z(X) = \sum_{y} \exp\left(\sum_{i,k} \lambda_k f_k(y_{i-1}, y_i, x) + \sum_{i,k} \alpha_k g_k(y_i, x)\right) \tag{4}$$

where f_k is a feature function and denotes the characters of output nodes of observed sequences at location of i and $i-1$. The g_k denotes the characters of input and output nodes at location of i. λ and α represent the weights of the feature functions. Furthermore, training data is used to build the word segmentation rules of CRF model with maximum entropy. We used parameters of *precision*, *recall* and *F-score* to test the performance of Chinese word segmentation.

$$precision = \frac{n_p}{n_t} \tag{5}$$

$$recall = \frac{n_p}{n_c} \tag{6}$$

$$F - score = \frac{2 \times precision \times recall}{precision + recall} \tag{7}$$

where n_p denotes the number of words that are properly segmented, n_t denotes the number of segmented words, n_c denotes the number of words in the corpus.

2.2. Term frequency-inverse document frequency

In a natural language corpus, the frequency of a given word has a logarithm inverse relationship with its frequency ranking (Powers, 1988). Term frequency (*tf*) is a statistical variable to calculate occurrence frequency of a word in a document. Most of time, the words with high frequency are function words, such as "the", "is", "in" "of" in English and "了 (Le)", "的 (De)", "是 (Shi)" in Chinese. These function words cannot represent the information contained in a document. In order to decrease the weight of function words, inverse document frequency (*idf*) was introduced into natural language processing to identify some important words that have a low frequency in documents. The *idf* is defined by Equation (8). The variable of *tf-idf* is the product of tf and idf.

$$idf = \ln\left(\frac{n_{documents}}{n_{containing\ word}}\right) \tag{8}$$

2.3. Knowledge graph

Knowledge graphs as semantic networks with directed graph structure have been widely used to improve the search engines of Google, Baidu and Yahoo. Each node in a graph represents an entity, and an edge represents the linking relationship between two entities. The graph visualizes the nodes and the links between nodes to represent the

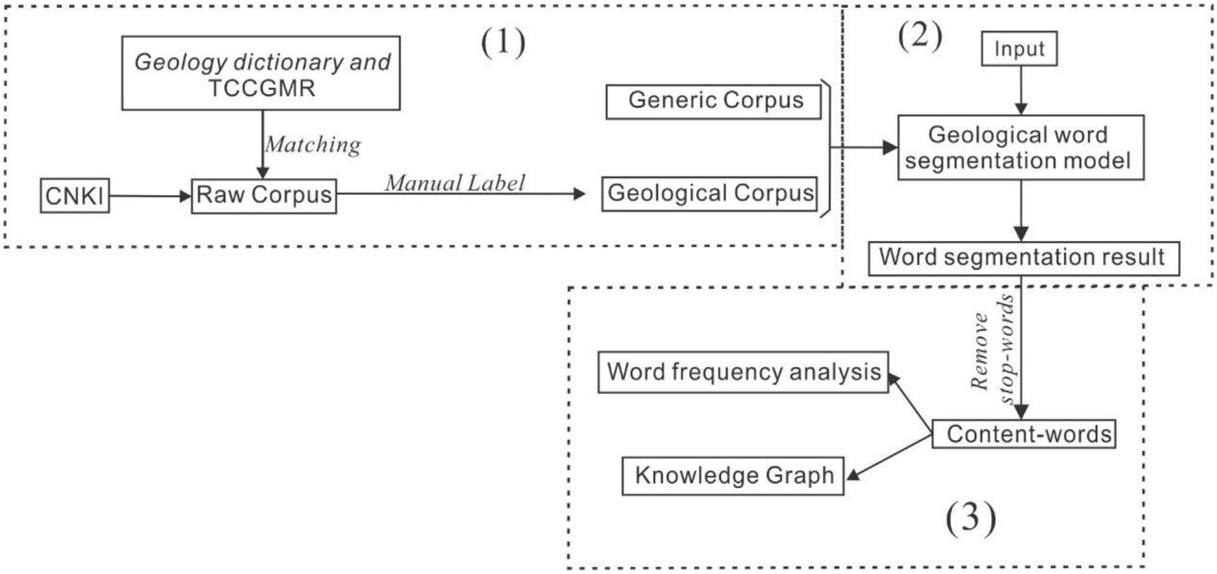

Fig. 2. Technique procedure in this study. TCCGMR: Terminologies and Classification Codes of Geology and Mineral Resources, CNKI: China National Knowledge Infrastructure.

information network and semantic relationships in a specific domain. In academic databases Web of Science and CNKI (China National Knowledge Infrastructure), it also has been used to analyze and display the citation map and other information links. Furthermore, some researchers try to extract the structured information from the unstructured web to build the knowledge base, such as DBPedia, OpenIE (Open information extraction) and NELL (Never-ending language learning), and transfer the text information into a network (e.g., Ehrlich et al., 2007; Paranyushkin, 2011).

For unstructured datasets, such as the geoscience literature, a key preprocessing step for building a knowledge graph is to extract the nodes and links from the text by using NLP method. In the text document, a sentence can be categorized into content-words and function words with an ambiguous meaning (Fries, 1952). The content-words representing the main entities are the carrier of key information in a literature. The function words have a high frequency for forming a sentence in the literature. In NLP, function words and punctuation marks constitute the stop-words that should be removed from word segmentation result. We can use matching algorithm to remove stop-words that are recorded in a stop-word library.

In this paper, chord and bigram graphs were selected to visualize the content-words and links as nodes and edges in a knowledge graph. The graphs were built based on "*from*", "*to*" and "*weight*" three variables. "*from*" variable means where the node comes from, "to" variable means where the node goes towards. The variables of "*from*" and "*to*" were defined based on the sequence of content-words. If two content-words are adjacent in a corpus, their relationship was regarded as "co-occurrence". In the content-word pairs with high co-occurrence frequency, the former content-word is "*from*", the latter is "*to*". The weight is defined by the co-occurrence frequency of two content-words in the whole literature. If we use the edge transparency to represent edge weight in the bigram graph, due to the existence of outlier of co-occurrence frequency, the edge color with lower co-occurrence frequency will be very light and difficult to distinguish in the figure. Although chord graph cannot express the term direction, the node width of chord graph gives a visual representation for weight.

2.4. Workflow

In order to make a knowledge graph from unstructured geoscience literature, the workflow was defined (Fig. 2). It was divided into three phases: hybrid corpus creation, Chinese word segmentation and knowledge graph construction.

2.4.1. Hybrid corpus creation

(a) Create a raw corpus derived from the geological literature in CNKI;
(b) Use terms of *geology dictionary* and *TCCGMR* to match the raw corpus and label them using label set $\{B, E, M, S\}$. This step introduces the professional knowledge contained in *geology dictionary* and *TCCGMR* into the raw corpus.
(c) After the dictionary matching, unmatched words are labeled using label set manually to build the geological corpus.
(d) Build the hybrid corpus by combining a generic corpus and the geological corpus together. In this study, the *People's Daily Corpus* released by Institute of Computational Linguistics, Perking University was selected as a generic corpus.

2.4.2. Chinese word segmentation

(a) Use the hybrid corpus to train the rules of word segmentation and build a CRF-based geological word segmentation model. The rules training and word segmentation were carried out based on the CRF++ 0.58 toolkit.[1]
(b) Use the geological word segmentation model to segment the input geological text document.

2.4.3. Knowledge graph construction

(a) Remove the stop-words from the word segmentation results by matching method and get the content-words. The stop-word library used in this study was an extended version of the stop-words of Harbin Institute of Technology.
(b) Analyze content-words frequency in individual chapters and the whole report.
(c) Select the chord and bigram graphs to visualize the content-words and links as nodes and edges in a knowledge graph.

[1] https://taku910.github.io/crfpp/.

3. Results

3.1. *Text document of geosciences literature*

Geoscience literature mainly contains journal papers, books, research reports and other types of literature. Geoscience literature is usually organized to describe the motivation, state of the art, methods, discussion and conclusion for a focused scientific topic. In China, geological reports are always organized according the requirements of China Geological Survey and characterized by prescribed chapter arrangement and simple rhetoric and terminology. Sometimes Chinese geoscience literature also adopts English words, such as author name, specific terms and abbreviations of terms.

In this study, we selected the geological report of *Advanced Technologies Studies of Geophysical Exploration and Remote-Sensing Geology in the Covered Area* as a case study to extract information and build a knowledge graph. The project behind the report was supported by China Geological Survey to employ methods of geophysical exploration and remote sensing to address the challenges of mineral exploration in the special geographic landscapes with few bedrock outcrops (i.e., grassland, and Gobi Desert covered area). This geological report includes 10 chapters and more than 100,000 Chinese characters. Chapter 1 is the introduction to the project. Chapter 2 describes state of the art in the domain of geophysical exploration and remote-sensing geology for mineral exploration. Chapter 3 shows the research ideas and technical workflow. Chapter 4 shows the method of extracting anomaly information from geophysical data and remote sensing in the covered area. Chapters 5–7 show geological interpretations of geophysical data and remote sensing image of Gobi Desert covered area of Xinjiang, forest covered area of Fujian and grassland covered area of Inner Mongolia, respectively. Chapter 8 is the conclusions of the project. Chapter 9 is acknowledgements. The last chapter is references. In this study, the last two chapters were not considered in text mining.

3.2. *Chinese word segmentation results*

The CRF-based Chinese word segmentation method is a discriminative model and needs abundant domain knowledge to train delimitation rules of Chinese words. In this study, we used a *geology dictionary* and *TCCGMR* to build the hybrid corpus. The *geology dictionary* we used is an integrated dictionary that includes more than 11,000 terms of forty geological disciplines. *TCCGMR* is constituted of more than 80,000 terms describing the objects and properties of geology and mineral resources in 35 geological sub-domains.

In order to verify the method proposed in this study, we randomly selected 1000 sentences in the domains of geology and mineral deposit from CNKI and labeled them as a test corpus. Furthermore, we used the test corpus to analyze the performances of the different training corpus. The training corpora consist of the generic Chinese corpus created by Peking University, the geological corpus created in this study and the hybrid corpus combining the generic and geological corpora together. The test results of the three corpora are shown in Table 1. The performance of Chinese word segmentation of the hybrid training corpus is better than the pure generic corpus and geological corpus. The segmentation precision of the hybrid training corpus is 7.84% and 0.52% greater than that of the generic corpus and geological corpus, respectively. The recall of hybrid corpus is 9.30% and 0.41% greater than the generic and geological corpora, respectively. The *F-score* of the hybrid corpus rises by 8.60% and 0.46% than the generic and geological corpora.

Table 1
Performances of CRF model in different corpora. CRF-PKU: generic corpus of Peking University, CRF-GEO: geological corpus, CRF-GEO + PKU: the hybrid corpus combining generic and geological corpora.

Method	Precision/%	Recall/%	F-score/%
CRF-PKU	86.30	82.10	84.15
CRF-GEO	93.62	90.99	92.29
CRF-GEO + PKU	94.14	91.40	92.75

Table 2
Classification of content-words in the geoscience literature used in this paper.

Classification	Examples
Geology-mineral resource	Structure (构造, Gouzao), Fault (断裂, Duanlie), Intrusive rock (侵入岩, Qinruyan), Mineral deposit (矿床, Kuangchuang).
Technique method	Remote sensing (遥感, Yaogan), Gravity (重力, Zhongli), Magnetic (磁, Ci), Aeromagnetic (航磁, Hangci)
Data processing method	Anomaly (异常, Yichang), Enhancement (增强, Zengqiang), Euler (欧拉, Oula), Tilt, Derivative (导数, Daoshu)
Descriptive words	Result (结果, Jieguo), Research (研究, Yanjiu), Data (数据, Shuju), Activity (活动, Huodong)

Note: Structure (构造, Gouzao): "构造" is a term in Chinese, **"Structure"** is the corresponding translation in English, **"Gouzao"** is Chinese phonetic alphabet represents the pronunciation of Chinese word.

3.3. *Content-words extraction*

The content-words represent the main information of the literature, which can be divided into four classes geology, technical method, data processing and some descriptive words in the study (Table 2). The words of geology class are terms associated with geology and mineral resource. The class of technique method mainly describes words related to mineral exploration methods in the covered area. The class of data processing method includes terms associated with data processing. Besides, words in the fourth class are some content-words that don't have a direct relationship with the topic of literature. But they are necessary for Chinese word-building with professional term and representing the literature information (Table 2). The function of these words is similar to function-word, but they have significant meanings. Hence, these words were named as descriptive words.

The valuable information contained in the literature has a correlation with the content-words with a high frequency (Hovy and Lin, 1998). However, there are some informative words with a low frequency (Piantadosi, 2014). *tf-idf* is a method, which was introduced into natural language processing to extract informative words with low frequency. The extraction results were showed in Fig. 3. The *tf-idf* method not only extracted the informative content-words with low frequency, but also introduced some function words into the results (Fig. 3), such as *big one* (一大, Yida), *next* (下一步, Xiayibu), *have* (有, You), must (必, Bi), *second* (第二个, Dierge) and the axis label "*x*", "*y*".

Before we calculate the frequency of content-words, we first removed the function words and some common words by word-matching method with a given stop-words library. The results of content-words frequency in individual chapters and in the literature as a whole were showed in Fig. 4 and Fig. 5. The word cloud was constituted of content-words with their frequencies greater than 40, which gives a brief visualization of information in geoscience literature. Compared with the *tf-idf* method, the results extracted by content-words frequency have a better representation for the given geological report. The results extracted by *tf-idf* contain some function words or common words, which dilute the extracted content-words. Those function words and common words have a negative impact on the result of the representative information of a geological report. Therefore, we selected the results of content-words frequency after removing stop-words to build the knowledge graphs.

3.4. *Knowledge graph of the literature*

The content-word is the carrier of literature information and knowledge. The key nodes and links derived from the content-words represent the knowledge of the literature. We used the bigram graph and chord graph to visualize the relationships of content-words. The visualization of knowledge network describes the relationship between key content-

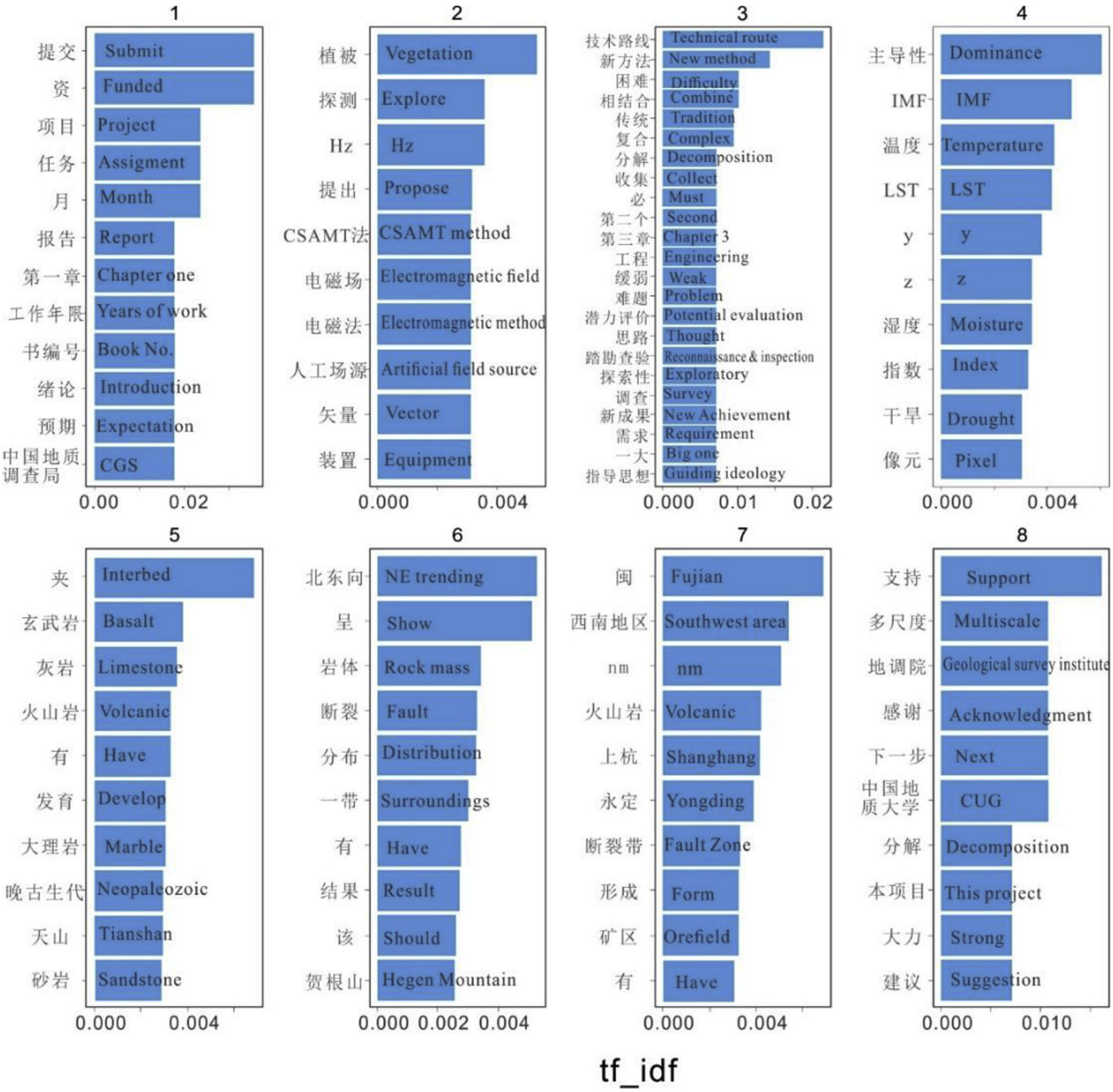

Fig. 3. Top 10 content-words extracted by *tf-idf* method. CGS: China Geological Survey, IMF: Intrinsic model function, LST: Land surface temperature, CUG: China University of Geosciences.

words (Figs. 6 and 7). It consists of knowledge and information in the domains of geology, geophysics, data processing, and is the essence knowledge of the literature.

In mineral exploration, the anomaly extracted from the geological data (e.g., geophysical data, remote sensing, etc.) is an important indicator for mineral deposits exploration (Wang et al., 2017). Therefore, the relationships between the word *anomaly* (异常, Yichang) and other content-words in the literature are important to characterize. The word *anomaly* (异常, Yichang) links with content-words *gravity* (重力, Zhongli), *magnetic* (磁, Ci), and *aeromagnetic* (航磁, Hangci), directly (Figs. 6 and 7). It also uses the link with information (信息, Xinxi) and *tilt* (tilt) to extend the network to *horizontal gradient* (水平梯度, Shuipingtidu) and *inversion* (反演, Fanyan) (Fig. 6).

4. Discussion

In this study, a hybrid corpus of geological term was built to train the delimitation rules of the CRF-based model to segment Chinese words. The *geology dictionary* and *TCCGMR* provided detailed domain knowledge of geology and mineral resources, and were used to match the geological terms of a raw corpus obtained from CNKI. It proves that the hybrid corpus not only improves the performance of identifying geological terms from a literature, but also has a high-performance identification for daily words.

In this research, all words in the raw corpus were labeled to train the word segmentation rules. The *geology dictionary* and *TCCGMR* were used to improve the precision of word segmentation. Because there are also some common words in the geological corpus, the rules of word segmentation trained by the geological corpus have a certain degree of overlap with the generic corpus. Therefore, the word segmentation performance of GEO + PKU corpus only has a light increase in terms of *precision*, *recall* and *F-score* than GEO corpus.

In a literature, the arrangement of words and chapter is organized around a topic. 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 (Firth, 1957). The words of the literature will have a lexical chain rather than are out of order (Barzilay and Elhadad, 1999; Halliday and Hasan, 1976).

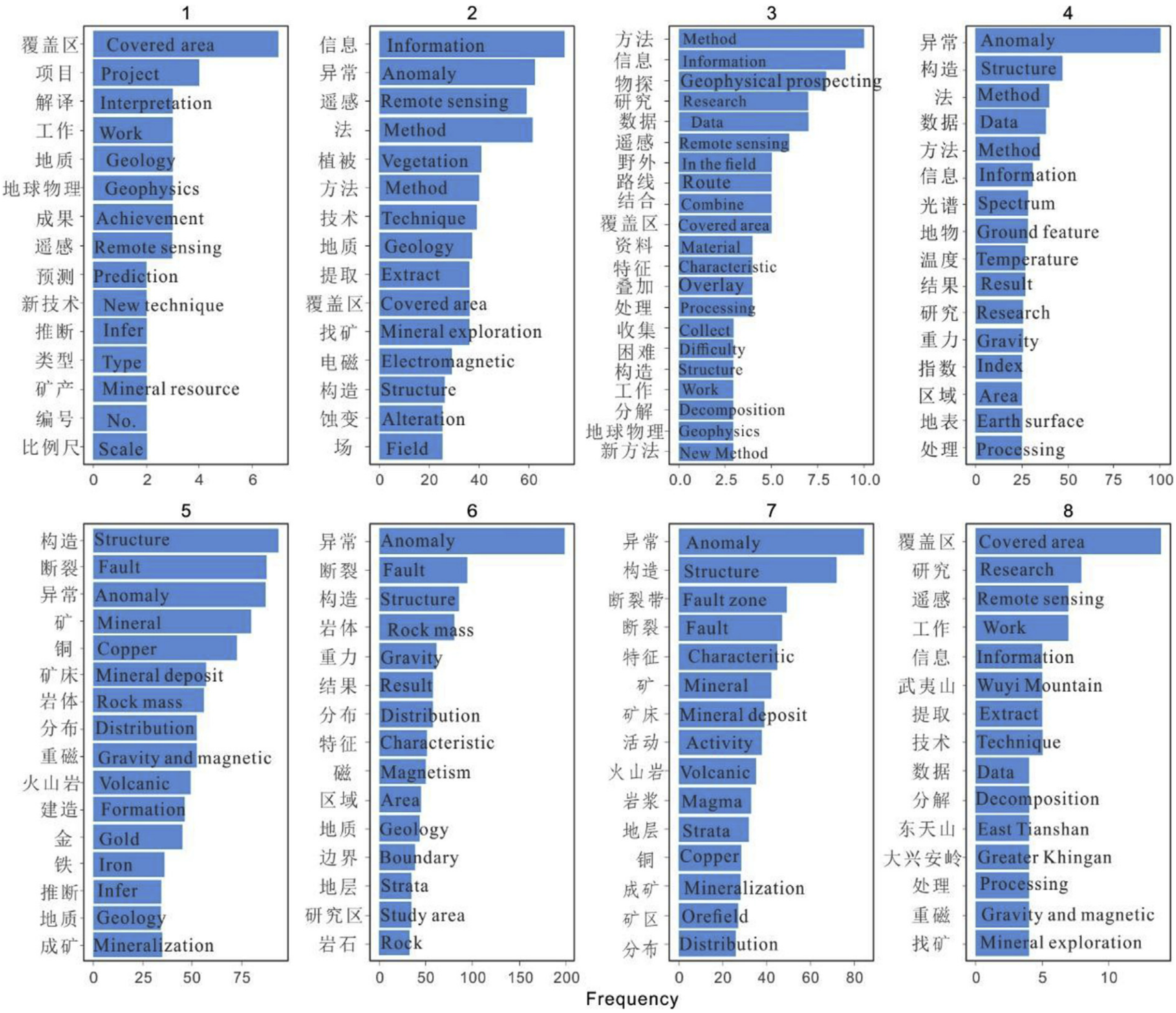

Fig. 4. Content-words with high frequency ($n > 10$) after removing stop-words. Parameter n denotes frequency of content-words in individual chapters of the geological report.

Therefore, we can use the knowledge graph extracted by the co-occurrence of content-words to represent the key information of a document. It can give us a graphical overview of a long document quickly so we do not need to read the document word by word.

There is no doubt that compound terms are more informative in human communication. However, computers cannot understand the meaning of words, and only use some variables to extract content-words. The single words are more efficient and reasonable for computers to extract key nodes and set up a knowledge graph. In the geological report selected in this paper, there are 38 geological terms in Chinese related with "磁 (magnetism)", such as 磁测 (magnetic survey), 瞬变电磁 (transient electromagnetic), 重磁 (gravity and magnetic), 磁化强度 (intensity of magnetization), 磁黄铁矿 (pyrrhotite) and more. These compound words are all constituted based on the linguistic root of "磁 (magnetism)". Although the compound words contain more informative and rigorous meaning, it is more difficult for computer to find a variable to describe the relationship between these compound words. Compared with compound words, the domain-specific single words are also

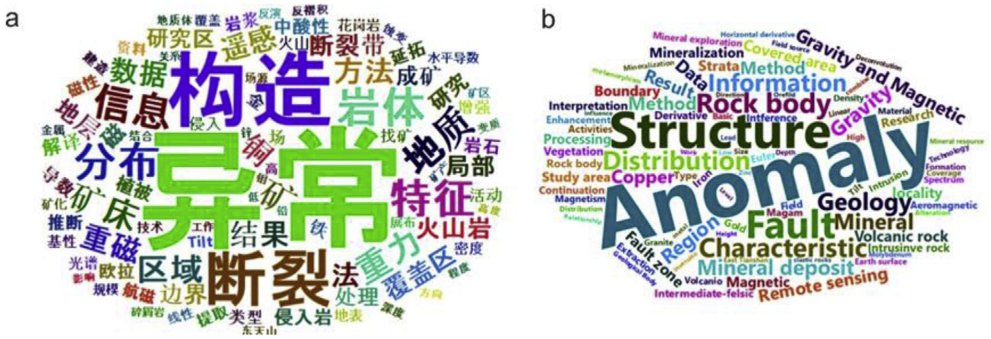

Fig. 5. Word cloud of the geological report ($n > 40$) shows a brief and visual representation of information contained in the whole literature. a-Chinese version, b-English version. Parameter n denotes frequency of content-words in the whole geological report. The font size of word scaled according to word frequency in the whole geological report.

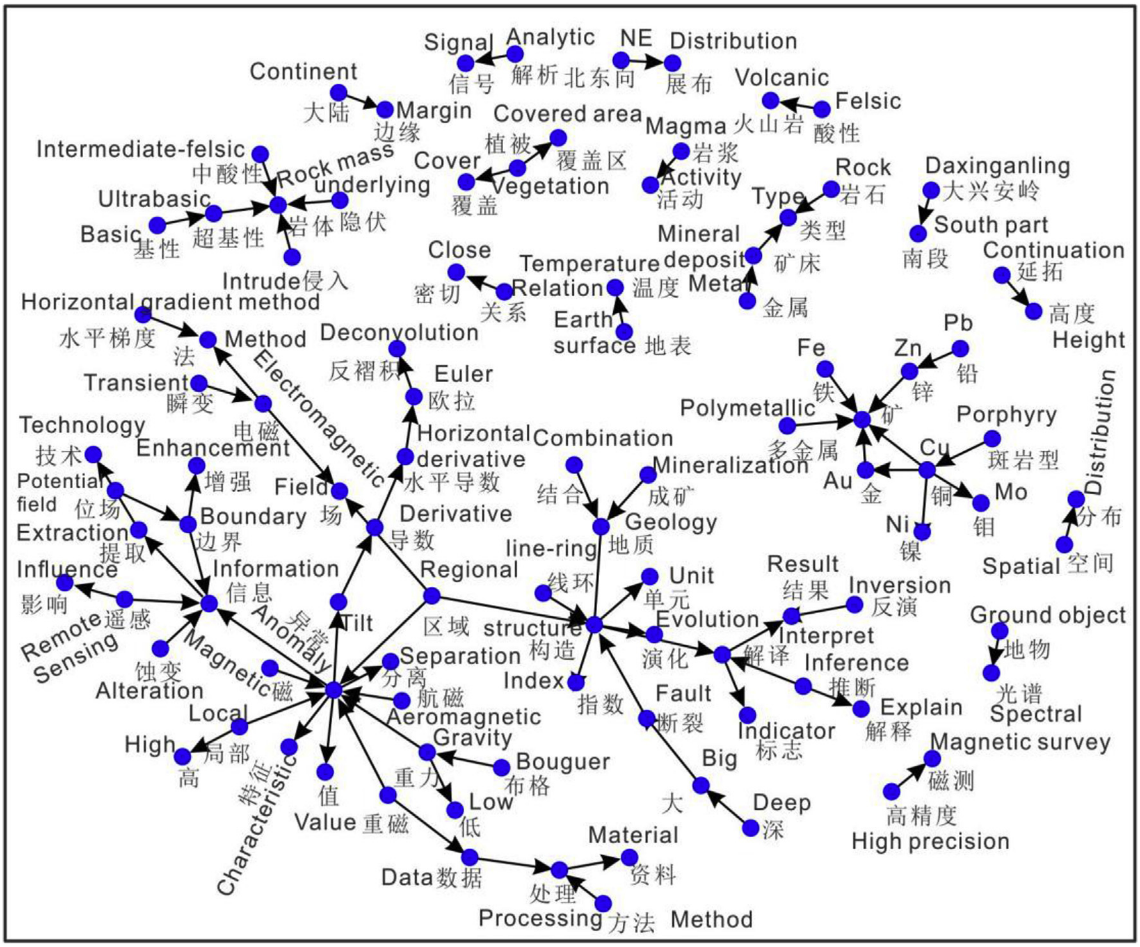

Fig. 6. Bigram graph of content-words in the whole literature represents the key information in the geological report ($n > 10$). The edge arrow indicates words sequence with the arrow pointing to the latter one of content-words pairs. Parameter n denotes co-occurrence frequency of content-words in the geological report.

informative. Furthermore, we can easily use the "磁 (magnetism)" to build the knowledge graph with other single words, such as 瞬变 (transient), 重 (gravity), by using the co-occurrence frequency of content-words. Hence, we segmented the Chinese sentence into minimum semantic units. For example, we divided the 瞬变电磁 into 瞬变 (transient) and 电磁 (electromagnetic) (Fig. 6).

The writing of geological literature with mixed Chinese and English is another negative influence factor. If we used the frequency method to extract information from literature, the writing combining Chinese and English would decrease the weight of a content-word in two types of language. For example, in terms of semantics, EMD (Empirical Mode Decomposition) is the same to "经验模态分解". These two words are identified as two different words for a computer, which will bring a low word frequency of EMD and have a negative influence on extraction of key nodes and edges for the knowledge graph.

Many researchers have tried to use automatic text summarization to extract a concise, informative and short abstract that covers the key information from a text document (e.g., Hu et al., 2015; Luhn, 1958; Nallapati et al., 2016). Nevertheless, due to the limitation of short summarization and poor readability, automatic text summarization has yet to achieve satisfactory results (Mitray et al., 1997). In this study, we used the NLP and knowledge graph methods to extract and visualize key information from unstructured geological literature. It proves that the methods we selected in this paper are appropriate and powerful for extracting a visual key information from unstructured geological literature quickly.

In the results of this research, the information contained in a geologic document is represented by a knowledge graph. In such a graph, content-words are represented as nodes, and the frequency of co-occurrence of those content words in the whole literature are represented as edges. Ideally, the bigram graph and chord graph can display all the nodes of geoscience literature by setting a low threshold value for visualization. In this paper, considering the space for displaying the English translation of Chinese words, we set a higher threshold value for displaying the key nodes and the edges between them. Therefore, several important nodes for individual chapters are not displayed in the bigram graph and chord graph (Figs. 6 and 7). The current visualization in this research was based on two-dimensional diagrams. A topic of interest can be further explored is to apply three dimensional diagrams (Ma et al., 2017) to discover the co-relationships between key content-words in a document.

5. Conclusions

In this research, we employed NLP methods and knowledge graphs to extract and visualize information extracted from geoscience literature. This work provides a new perspective to reuse massive amounts of unstructured literature, which is increasingly made open and accessible in recent years. The following conclusions can be drawn: (1) The *geology dictionary* and *TCCGMR* provide content-rich domain knowledge to build a geological corpus. The geological corpus combined with the PKU generic corpus improves the performance of the CRF-based model for Chinese word segmentation. (2) Word frequency analysis displays the distribution and usage of content-words in individual chapters and literature as a whole, which is helpful for understanding the knowledge distribution and extracting information with a specific topic. (3) The visualization of information is a useful tool to represent the hidden knowledge in the unstructured literature.

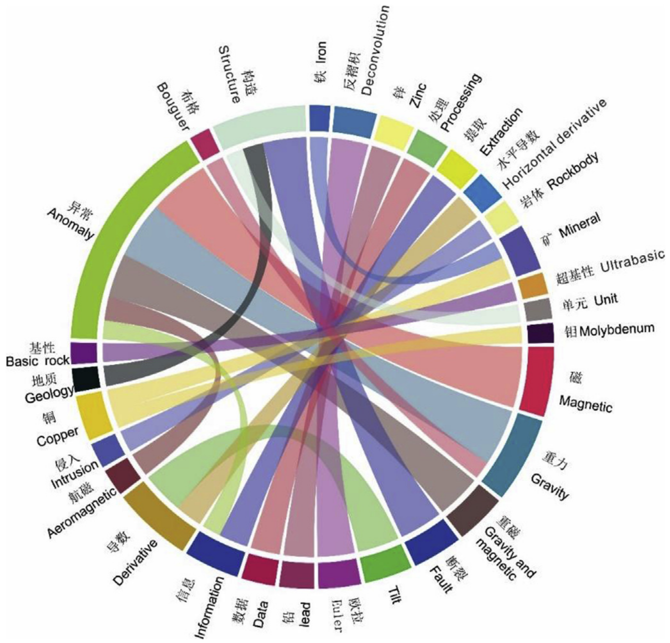

Fig. 7. Chord graph showing the inter-relationship between content-words in the geological report (n > 20). Chord width scaled according the content-words frequency, with arbitrary colors. Parameter *n* denotes co-occurrence frequency of content-words in the geological report. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

A few topics can be proposed for the future work: (1) Information retrieval between knowledge graph and original literature. Knowledge graph only show the key nodes and edges between them. A directional query function is needed to link detailed description in original literature to nodes of knowledge graph for retrieve. (2) A pilot system is needed to integrate the NLP methods, visualization, and information retrieval in original literature. (3) We can use the animatic network analysis instead of bigram and chord graph to visualize the knowledge in geoscience literature. (4) Knowledge graph construction from multi-documents and information inference.

Acknowledgement

This work was supported by the National Key R&D Program of China (2017YFC0601500, 2017YFC0601504), the China Geological Survey through the project of Research and Application of Big Data in geoscience (201511079-02), the National Science Foundation (NSF) through the NSF Idaho EPSCoR Program (IIA-1301792), and the 2017 Computers and Geosciences Research Scholarships of IAMG.

References

Barzilay, R., Elhadad, M., 1999. Using lexical chains for text summarization. In: Mani, I., Maybury, M.T. (Eds.), Advances in Automatic Text Summarization. The MIT Press, pp. 111–121.

Cernuzzi, L., Pane, J., 2014. Toward open government in Paraguay. IT Prof. 16 (5), 62–64.

Cracknell, M.J., Reading, A.M., 2014. Geological mapping using remote sensing data: a comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. Comput. Geosci. 63, 22–33.

Ehrlich, K., Lin, C.Y., Griffiths-Fisher, V., 2007. Searching for experts in the enterprise: combining text and social network analysis. In: Proceedings of the 2007 International ACM Conference on Supporting Group Work, pp. 117–126.

Firth, J.R., 1957. A Synopsis of Linguistic Theory 1930-1955, in Studies in Linguistic Analysis. Blackwell Publishers, Oxford, pp. 1–32.

Fries, C.C., 1952. The Structure of English: an Introduction to the Construction of English Sentences. Harcourt, Brace, New York, p. 304.

Gao, J.F., Li, M., Wu, A., Huang, C.N., 2005. Chinese word segmentation and named entity recognition: a pragmatic approach. Comput. Ling. 31, 531–574.

Giboin, A., Grataloup, S., Morel, O., Durville, P., 2013. Building Ontologies for analyzing data expressed in natural language. In: Perrin, M. (Ed.), Shared Earth Modeling: Knowledge Driven Solutions for Building and Managing Subsurface 3D Geological Models. Editions Technip, Paris, pp. 231–259.

Halliday, M.A.K., Hasan, R., 1976. Cohesion in English. Longman, London, p. 374.

Hovy, E., Lin, C.Y., 1998. . Automated text summarization and the SUMMARIST system. In: Proceedings of a Workshop on Held at Baltimore, Maryland: October 13-15, 1998(TIPSTER '98). Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 197–214.

Hu, B., Chen, Q., Zhu, F., 2015. Lcsts: a Large Scale Chinese Short Text Summarization Dataset. arXiv preprint arXiv:1506.05865.

Huang, L., Du, Y.F., Chen, G.Y., 2015. GeoSegmenter: a statistically learned Chinese word segmenter for the geoscience domain. Comput. Geosci. 76, 11–17.

Lafferty, J.D., McCallum, A., Pereira, F.C.N., 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc, pp. 282–289.

Lima, L.A., Görnitz, N., Varella, L.E., Vellasco, M., Müller, K.-R., Nakajima, S., 2017. Porosity estimation by semi-supervised learning with sparsely available labeled samples. Comput. Geosci. 106, 33–48.

Luhn, H.P., 1958. The automatic creation of literature abstracts. IBM J. Res. Dev. 2, 159–165.

Ma, X., 2017. Linked Geoscience Data in practice: where W3C standards meet domain knowledge, data visualization and OGC standards. Earth Sci. India 10 (4), 429–441.

Ma, X., Wu, C., Carranza, E.J.M., Schetselaar, E.M., van der Meer, F.D., Liu, G., Wang, X., Zhang, X., 2010. Development of a controlled vocabulary for semantic interoperability of mineral exploration geodata for mining projects. Comput. Geosci. 36 (12), 1512–1522.

Ma, X., Hummer, D., Golden, J.J., Fox, P.A., Hazen, R.M., Morrison, S.M., Downs, R.T., Madhikarmi, B.L., Wang, C., Meyer, M.B., 2017. Using visual exploratory data analysis to facilitate collaboration and hypothesis generation in cross-disciplinary research. Int. J. Geo-Inf. 6 (11), 368.

Manning, C.D., Schütze, H., 1999. Foundations of Statistical Natural Language Processing. MIT Press, p. 680.

McCallum, A., Li, W., 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003-Volume 4. Association for Computational Linguistics, Edmonton, Canada, pp. 188–191.

McDonald, R., Pereira, F., 2005. Identifying gene and protein mentions in text using conditional random fields. BMC Bioinf. 6, S6.

Ministry of Geology and Mineral Resources, 2005. Geological Handbook. Geological Publishing House, Beijing (in Chinese).

Mitray, M., Singhalz, A., Buckleyyy, C., 1997. Automatic Text Summarization by Paragraph Extraction. Compare 22215, 26.

Morrison, S.M., Liu, C., Eleish, A., Prabhu, A., Li, C., Ralph, J., Downs, R.T., Golden, J.J., Fox, P., Hummer, D.R., Meyer, M.B., Hazen, R.M., 2017. Network analysis of mineralogical systems. Am. Mineral. 102 (8), 1588–1596.

Nallapati, R., Zhou, B., Gulcehre, C., Xiang, B., 2016. Abstractive Text Summarization Using Sequence-to-sequence Rnns and beyond. arXiv preprint arXiv:1602.06023.

Paranyushkin, D., 2011. Identifying the Pathways for Meaning Circulation Using Text Network Analysis. Nodus Labs, Berlin. http://noduslabs.com/research/pathways-meaning-circulation-text-network-analysis.

Peters, S.E., McClennen, M., 2015. The Paleobiology Database application programming interface. Paleobiology 42, 1–7.

Peters, S.E., Zhang, C., Livny, M., Re, C., 2014. A machine reading system for assembling synthetic paleontological databases. PLoS One 9, e113523.

Piantadosi, S.T., 2014. Zipf's word frequency law in natural language: a critical review and future directions. Psychonomic Bull. Rev. 21 (5), 1112–1130.

Pinto, D., McCallum, A., Wei, X., Croft, W.B., 2003. Table extraction using conditional random fields. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, Toronto, Canada, pp. 235–242.

Powers, D.M.W., 1988. Applications and explanations of Zipf's law. Adv. Neural Inf. Process. Syst. 5 (4), 595–599.

Ruokolainen, T., Kohonen, O., Virpioja, S., Kurimo, M., 2013. Supervised Morphological Segmentation in a Low-resource Learning Setting Using Conditional Random Fields. CoNLL, pp. 29–37.

Sato, K., Sakakibara, Y., 2005. RNA secondary structural alignment with conditional random fields. Bioinformatics 21, ii237–ii242.

Schuhmacher, M., Ponzetto, S.P., 2014. Knowledge-based graph document modeling. In: Proceedings of the 7th ACM International Conference on Web Search and Data Mining, pp. 543–552.

Singhal, A., 2012. Introducing the Knowledge Graph: Things, Not Strings. Official google blog. http://y-x.iteye.com/admin/blogs/2012810.

Turney, P.D., Pantel, P., 2010. From frequency to meaning: vector space models of semantics. J. Artif. Intell. Res. 37, 141–188.

Wallach, H.M., 2004. Conditional Random Fields: an Introduction. University of Pennsylvania, p. 11. http://repository.upenn.edu/cgi/viewcontent.cgi?article=1011&context=cis_reports.

Wang, X., Liu, G., Yuan, Y., Han, Z., 1999. Application of" the terminology classification codes of geology and mineral resources" on geological and mineral resources point-source information system. J. Earth Sci. 5, 022 (In Chines with English abstract).

Wang, C., Chen, J., Xiao, F., 2016a. Application of empirical model decomposition and independent ComponentAnalysis to magnetic anomalies separation: a case study for Gobi Desert coveragein eastern tianshan, China. In: Geostatistical and Geospatial Approaches for the Characterization of Natural Resources in the Environment. Springer, Cham, pp. 593–598.

Wang, C., Chen, J., Xiao, F., Fode, T., Li, L., 2016b. Radioelement distributions and analysis of microtopographical influences in a shallow covered area, Inner Mongolia, China: implications for mineral exploration. J. Appl. Geophys. 133, 62–69.

Wang, C., Rao, J., Chen, J., Ouyang, Y., Qi, S., Li, Q., 2017. Prospectivity mapping for "Zhuxi-type" copper-tungsten polymetallic deposits in the Jingdezhen region of Jiangxi province, south China. Ore Geol. Rev. 89, 1–14.

Xiao, F., Chen, Z., Chen, J., Zhou, Y., 2016. A batch sliding window method for local singularity mapping and its application for geochemical anomaly identification. Comput. Geosci. 90, 189–201.

Xu, Y., Wang, X., Tang, B., Wang, X., 2008. Chinese unknown word recognition using improved conditional random fields. In: Intelligent Systems Design and Applications, 2008. ISDA'08. Eighth International Conference. IEEE, pp. 363–367.

Xue, N., 2003. Chinese word segmentation as character tagging. Comput. Ling. Chin. Lang. Process. 8, 29–48.

Zhang, C., 2008. Automatic keyword extraction from documents using conditional random fields. J. Comput. Inf. Syst. 4, 1169–1180.