Contents lists available at ScienceDirect

Ore Geology Reviews

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

# Visual analytics and information extraction of geological content for text-based mineral exploration reports

Bin Wang[a], Kai Ma[e], Liang Wu[a,b,c,d], Qinjun Qiu[b,c,d,*], Zhong Xie[a,b,c,d], Liufeng Tao[b,c,d]

[a] School of Geography and Information Engineering, China University of Geosciences, Wuhan 430074, China
[b] School of Computer Science, China University of Geosciences, Wuhan 430074, China
[c] Hubei Key Laboratory of Intelligent Geo-Information Processing, China University of Geosciences, Wuhan 430074, China
[d] National Engineering Research Center of Geographic Information System, Wuhan 430074, China
[e] College of Computer and Information Technology, China Three Gorges University, Yichang 443002, China

ARTICLE INFO

*Keywords:*
Geoscience reports
Text visualization
Text mining
Natural language processing
Multidimensional scaling analysis

ABSTRACT

A large amount of continuously increasing textual geoscience data is stored and not fully utilized. Text mining enables the discovery and analysis of valuable information,and presents valuable insights hidden in geological texts. This research aims to use text mining and visualization techniques to obtain content words -for the purpose of visually analyzing geological reports. The framework proposed in this study can enable researchers to quickly understand key information and improve the transmission efficiency of geological reports. First, we implemented an improved keyword extraction algorithm comprising the term frequency-inverse document frequency and word length to improve the accuracy of geological keyword extraction. Second, we extracted and visualized the relative importance as well as the links between content words that can represent the key information of geoscience reports using word-level information analysis and multidimensional scaling analysis. Finally, the keyword relevance and mutual clustering relations were visualized through graphs to provide an intuitive representation of the current state of the reports.

## 1. Introduction

In recent times, the exponential growth of a considerable amount of georeferenced quantitative data (e.g., technical literature, books, and diverse types of reports) from governmental agencies and scientific organizations have contributed to a rich data source for information reuse and discovering new knowledge (Wu et al., 2017; Wang et al., 2018; Qiu et al., 2018b). For example, open and published available exploration reports provide an essential source of information for mineral explorers researching geology and mineralization in a target area. The contents of these reports contain geological ages, existing rock types, types of alteration present, and other geological, geochemical, and geophysical information. All of these contents offer essential information and knowledge for mapping and modeling ore-forming processes. However, the massive volume of texts from such reports requires considerable time and manual effort for interpretation, so the complexity of data management are increased (Khare & Chougule, 2012; Deng et al., 2016a; Deng et al., 2016b; Enkhsaikhan et al., 2021a; Enkhsaikhan et al., 2021b; Qiu et al., 2020a; Ma et al., 2021a; Ma et al., 2021b). Accordingly, approaches to ensure accuracy and improve the efficiency of information transmission are a considerable challenge for geologists (Lima et al., 2017; Ma, 2017; Qiu et al., 2019; Qiu et al., 2020b).

To discover new information and valuable knowledge hidden in geosciences reports/documents via data analysis, an effective and intuitive method is required. Computer-assisted text visualization provides an alternative way to enrich our understanding and improve the performance of the current methods used for information management (Qiu et al., 2018b; Sun et al., 2020). Through the presentation of information in a visual or graphical format, visual text analytics enables researchers and decision-makers to understand relevant geological text data and discover new knowledge hidden in geological reports in an intuitive manner, facilitating the comprehension of complex concepts or understanding of extraction results and computations. It also helps the researchers understand the relationships between the outcomes in graphical or pictorial formats (Marzouk and Enaba, 2019; Qiu et al., 2019). Appropriately designed visual graphics can improve the accuracy and comprehensibility of geographic information transmitted and provide vital information to support managers in making timely and

* Corresponding author at: School of Computer Science, China University of Geosciences, Wuhan 430074, China.
*E-mail address:* qiuqinjun@cug.edu.cn (Q. Qiu).

https://doi.org/10.1016/j.oregeorev.2022.104818
Received 14 October 2021; Received in revised form 23 January 2022; Accepted 4 March 2022
Available online 8 March 2022

accurate decisions. In contrast to inefficient and error-prone information extraction, visualization techniques with text mining and natural language processing (NLP) in the field of geoscience data management can be used to automatically obtain critical information and content from the relevant bodies of geoscience reports; subsequently, this information can be viewed using more intuitive maps and graphics, thereby providing essential transmission for managers and facilitating the elimination of errors from the subjectivity of the extraction process (Holden et al., 2019; Sun et al., 2020).

In the geoscience domain, text analytics based on text mining and NLP techniques has already been implemented to query computer-aided drawings from different sources, classify geological reports, analyze and predict the co-occurrence relationships of geological entities, and extract critical information in a structured format. The integration of text visualization analysis and text mining technology provides an efficient method to process such a massive amount of geological text; it also helps researchers to discover new knowledge, obtain relevant content, enrich their understanding of the state of the reports, and make decisions. Peters et al. (2014) and Peters and McClennen (2015) implemented a textual NLP algorithm to obtain structured paleontological information and contents from numerical documents and constructed the Paleobiology Database. Peters et al. (2017) presented a combined application of stratigraphic databases and open geoscience reports. They developed an automated machine-read platform to analyze stromatolites' prevalence, extinction, and recovery in their work.Wang et al. (2018) proposed a workflow for extracting semantic information and constructing a knowledge graph from textual geoscience data written in Chinese. They used a statistical model to extract semantic links based on content words and used bigrams and chord graphs to depict the semantic relations. The application of text mining techniques assists in understanding geological knowledge. Shi et al. (2018) developed a deep learning method to retrieve related knowledge between the deposits in Sichuan Province, China. They visualized the content-word co-occurrences and frequency statistics to present the key information. Using an automated geological reports analysis platform called GeoDocA, Holden et al. (2019) browsed and retrieved relevant geological content and knowledge from 25,419 exploration documents. Qiu et al. (2020) presented a systematic workflow to automatically extract spatio-temporal and semantic information from unstructured geoscience reports/documents based on text mining techniques. We offered a set of representative and contextual information in a knowledge graph form and searched a list of similar reports based on geological topic information. Zhuang et al. (2021) present a multi-constraint fusion feature weighted model to extract the thematic feature items from the content fragments. Wang et al. (2021) utilize the text content of the stratigraphic information description in the geological report and the corresponding digital elevation model data of the study area. It can solve the problem of the reuse of outdated data and reconstruct geological profiles at a low cost in the absence of field measurement data. Ma et al. (2021a), Ma et al. (2021b) utilize the bidirectional encoder representations from the transformers model to automatically generate the tile of given input summarization based on the constructed dataset. Enkhsaikhan et al. (2021a), Enkhsaikhan et al. (2021b) automatically extracts geological information relevant to mineralization and ore-forming conditions from such under-utilized exploration reports. NLP and deep learning methods are used to automatically extract and label geological terms with the correct entity types and establish the relationships between these entities, and then they construct two knowledge graphs for two high-quality mineral exploration reports, one for iron ore and the other for gold deposit. Although these above approaches have been extensively applied to text mining and visualization applications, there have been very few studies on text visualization about the accuracy and efficiency of keyword extraction and on the developing geoscience-related network analysis, so further research remains necessary.

Indeed, most previous studies on text visualization of geological reports have focused on visualization of keywords, co-occurrence information, and knowledge graphs in areas such as Geophysical Exploration and Remote-Sensing Geology (Wang et al., 2018) and mineral exploration reports (Holden et al., 2019), while few studies have conducted to understand the key contents of geological reports at the semantic level in more detail. Such geological report texts provide a comprehensive description of mineral resource investigations in the relevant area and usually involve multiple target contents such as geological formations, stratigraphy, rocks, and minerals. Processing and refining such report texts using text mining and visualization analysis methods can ensure that geologists or researchers can effectively access relevant information and have an intuitive and accurate understanding of the state of geological reports.

Accordingly, this paper proposes a computational framework that includes keyword extraction, visual display, and semantic similarity analysis. This framework can be used to analyze geological reports to extract relevant keywords/keyphrases and their co-occurrences and visualize the extracted information via visual mapping. The modified Term Frequency-Inverse Document Frequency (TF-IDF) algorithm is applied to extract the keywords to represent the key information and contents in the geological text. Then, by using keyword knowledge analysis and multidimensional scaling analysis, the relative importance and mutual clustering relationships of each keyword identified in the text are clarified. Finally, the geological report is visualized in the form of a tag cloud and semantic similar word view by related algorithm to provide an intuitive and easy to understand picture of the current content of the report.

The main contributions of this research are summarized as follows:

(1) From a methodological perspective, this paper contributes a computational framework that integrates NLP and text mining for visualizing the extracted information, and extracts a list of relevant keywords/keyphrases based on an improved algorithm consisting of the term frequency-inverse document frequency (TF-IDF) and word length.
(2) From an application perspective, this paper visualizes the geological contents within individual reports using word-level information analysis and multidimensional scaling analysis to represent the occurrences and links of content words based on co-occurrences in sentences and provides query suggestions for the robust querying of reports from a set of geological report repositories using the extracted keywords/keyphrases and semantic similarity approach associations.

The effectiveness of the proposed framework, which is based on text mining and visualization analysis, in processing and refining report texts, is evaluated to ensure that geologists and managers can effectively obtain relevant information and have an intuitive and accurate understanding of the state of the reports.

The remainder of this article is organized as follows. Section 2 presents the details of our proposed framework, which consists of three steps. Section 3 illustrates the visualization process for geological text. Next, Section 4 introduces the discussion. Finally, conclusions and future work are summarized.

## 2. Methodology

The primary objective of this study is to help information managers efficiently analyze and obtain effective geological content when reviewing a single geological report. We proposed a visualization framework to depict the critical content of the geological text quickly. The framework includes three stages, i.e., information extraction, visual display, and semantic similarity analysis. The modified TF-IDF algorithm was associated with the word length for keyword extraction to improve the accuracy, and its improvement over the traditional TF-IDF algorithm was experimentally evaluated. By extracting geologically relevant keywords and their co-occurrences, visual analytics (e.g., visual tag cloud and keyword centrality analysis) was utilized to visualize a geological text as a case study.

### 2.1. Methodology overview and the proposed approach

The framework used for extracting and visualizing essential information (e.g., extraction keywords/keyphrases) from unstructured geoscience reports is illustrated in Fig. 1. It consists of three components, namely keyword extraction, visual display, and semantic similarity analysis.

As a primary step, keyword extraction develops the core of geological status text information visualization. Using the modified TF-IDF algorithm (discussed in Section 2.2), a series of keywords that contain a list of core vocabulary was extracted to represent the key information and contents in the geological text. In this process, the critical processing steps include sentence splitting, tokenization, word frequency statistics, calculation results of the improved TF-IDF, and word filtering. Generally, the title of a paragraph or chapter can represent the key content of the paragraphs and chapters. However, some paragraphs and headings do not accurately and precisely represent key content, especially section headings (e.g., only simple terms such as stratigraphy and magmatic rock are used to indicate the content of the chapter). Accurately extracting and presenting key information from a geological report remains necessary.

Subsequently, the co-occurrence information was calculated to convert the geological text data for further visual processing using visual mapping. The geological information and content can be structured and standardized to display enhanced geological details (e.g., semantic links) by discovering the semantic relations within the key phrases. The social network analysis approach was selected for this procedure. By extracting the centrality of keywords and implementing multidimensional scaling along with semantic similarity analysis (e.g., BERT representations clustering), the co-occurrence links of the key phrases in the geological text enable the direct illustration of the extracted information and content words.

Next, the visual display was utilized to depict the geological information and content words to the researchers/decision-makers because pictures and graphics (Bourne and Weaver, 2018; Holden et al., 2019; Sun et al.,2020) are more sensitive and intuitive to human beings than a series of dense pages (a geological report usually has more than 100 pages) and list of dull words; additionally, information converted with pictures is more lasting and intuitive.

### 2.2. Keyword extraction based on improved TF-IDF

The Chinese text in geoscience reports contains no spaces between words, unlike western languages; therefore, Chinese word segmentation (CWS) is an essential step for further processing and recognizing the boundaries of meaningful words in sentences (Gao et al., 2005; Huang et al., 2015a; Huang et al., 2015b; Wang et al.,2018; Qiu et al., 2018a; Qiu et al., 2018b). In this study, we first developed a hybrid corpus that includes a generic corpus and a domain-geoscience corpus; then, we applied the hybrid corpus to train and build a deep learning-based domain-geoscience word segmentation model. The main structure (deep learning CWS model) was formulated based on Bidirectional Long Short-term Memory (BiLSTM) (Qiu et al., 2018b).

After obtaining the segmented text, the next step involved extracting Chinese keyword information based on the TF-IDF algorithm (Salton et al., 1974; Chen 2017). The TF-IDF algorithm calculates the frequency of a phrase in a particular document from a collection of documents. Thus, it helps determine the importance of a phrase or word in an input document in a group; the higher the frequency, the better the relevance.

The TF-IDF measure for phrase *P* in a document is calculated as follows:.

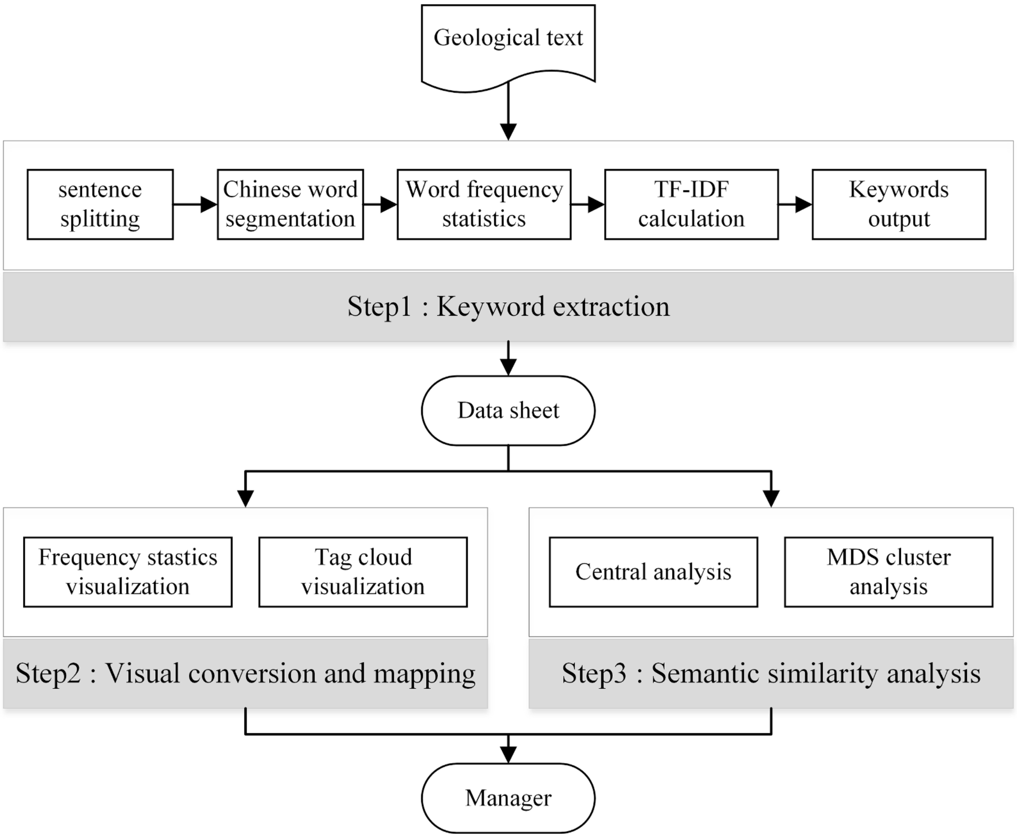

Geological text
sentence splitting
Chinese word segmentation
Word frequency statistics
TF-IDF calculation
Keywords output
Step1 : Keyword extraction
Data sheet
Frequency stastics visualization
Tag cloud visualization
Central analysis
MDS cluster analysis
Step2 : Visual conversion and mapping
Step3 : Semantic similarity analysis
Manager

**Fig. 1.** Visualization framework based on the proposed methods. The framework consists of three steps, there are keyword extraction, visual conversion and mapping and semantic similarity analysis. MDS: Multidimensional.

$$TF-IDF(P) = TF(P)\cdot IDF(P) = \frac{frequency(P)}{wc(D)}\cdot log\frac{|C|}{df(P)} \tag{1}$$

where *frequency(P)* denotes the frequency of *P* in *D*, *wc(D)* denotes the number of words in *D*, $|C|$ denotes the number of documents in document collection *C*, and *df(p)* denotes the number of documents in *C* that include *P*.

In particular, extracting keywords using the traditional TF-IDF algorithm in a specific domain is more challenging. Due to this method's lack of semantic information, it is not suitable for obtaining information from geological texts (Arroyo-Fernández et al., 2019; Qiu et al., 2019; Mee et al.,2021), some targeted revisions are required to improve its accuracy and suitability for this task. Furthermore, some studies have demonstrated that sentence structure plays a vital role in extracting keywords using text mining (Gao et al., 2012). Hence, in this step, the traditional TF-IDF computing method was revised to improve the accuracy and performance of keyword extraction considering the geological vocabulary and phrase length characteristics.

Notably, a long word can express and reveal more content and information of a sentence or document than a short term (Yu et al.,2016; Sun et al., 2020). Moreover, compared to the different effects of word features in keyword extraction, the impact of word length is significantly enhanced for other features (Berend and Farkas, 2010). Therefore, the word length information was applied to increase the weight value of words in keyword extraction. Accordingly, a weighting equation was formulated using a dynamic approach, as follows:

$$weight_{len}(w_i) = \frac{len(w_i)}{max_{len}} \tag{2}$$

where *len*($w_i$) denotes the length of the current word, $w_i$. Meanwhile, to prevent the dominance of long phrases, the denominator, $max_{len}$, was normalized; it denotes the length of the longest words in the geological reports.

Integrating Equations (1) and (2), keyword scoring can be defined using the following equation:

$$Score(w_i) = TF{*}IDF + weight_{len}(w_i) \tag{3}$$

After computing *Score*($w_i$), a set of candidate keywords was extracted from the input reports whose scores were calculated. Then, the keywords were ranked by their scores; consequently, the top *N* candidate keywords were selected as the core and meaningful keywords.

### 2.3. Visual mapping based on keyword relevance

Based on a list of extracted keywords/keyphrases, the process of visual mapping is conducted to transform the data so that it could be processed into a visual structure (Yang et al.,2018). Using visual mapping, the geological information extracted by the modified TF–IDF algorithm and related processes could be standardized and structed and can illustrate more detailed status information by mining the deeper relations between keywords. In this research, we use the social network analysis approach for this process. By analyzing the keyword centrality and the multidimensional scaling and clustering effect, the relationship between the keywords in the geological text can be determined in order to provide the structure of the extracted status information.

As an essential priority of keyword analysis centrality, the three representative indicators of keyword centrality include degree centrality (DC), closeness centrality (CC), and betweenness centrality (BC) (Akiyoshi, 2008; Sun et al., 2020); the calculation formulas of these three indexes are shown in Equations (4), (5), and (6), respectively. In a social network, the DC of nodes is composed of terms/phrases that are absolute and relative. Absolute centrality analysis can be described as the other nodes in a network directly connected to the current keyword/keyphrase. Their values are higher, implying that the closer the current keyword is to the central location, the greater its capacity to influence the remaining keywords. Relative centrality refers to the percentage of absolute centrality analysis relative to the absolute center of the network. The CC reflects the importance of the current keyword in a social network by calculating the relative distance range of its node from other relevant nodes. Finally, BC is used to represent the degree of the keyword node that controls the network resources. In other words, if a given node belongs to the shortest path, it has a higher magnitude.

$$C_D(N_i) = \sum_{j=1}^{g} x_{ij}, (i \neq j) \tag{4}$$

$C_D$ *(N*$_i$*)* denotes the standard degree of node *i*, where *i* represents the number of direct connections between node *i* and the other *g-1 j* nodes. Further, $i \neq j$ indicates that the relation between *i* with itself is excluded; that is, the value of the main diagonal can be ignored.

$$C_C(N_i) = \frac{1}{\sum_{j=1}^{g} dis(i,j)}, (i \neq j) \tag{5}$$

where *dis(i, j)* represents the distance from node *i* to node *j*. The larger the value of $C_C(N_i)$, the higher the closeness centrality.

$$C_B(N_i) = \sum_{j,k=1}^{g} sd(j,i,k), (j \neq k) \tag{6}$$

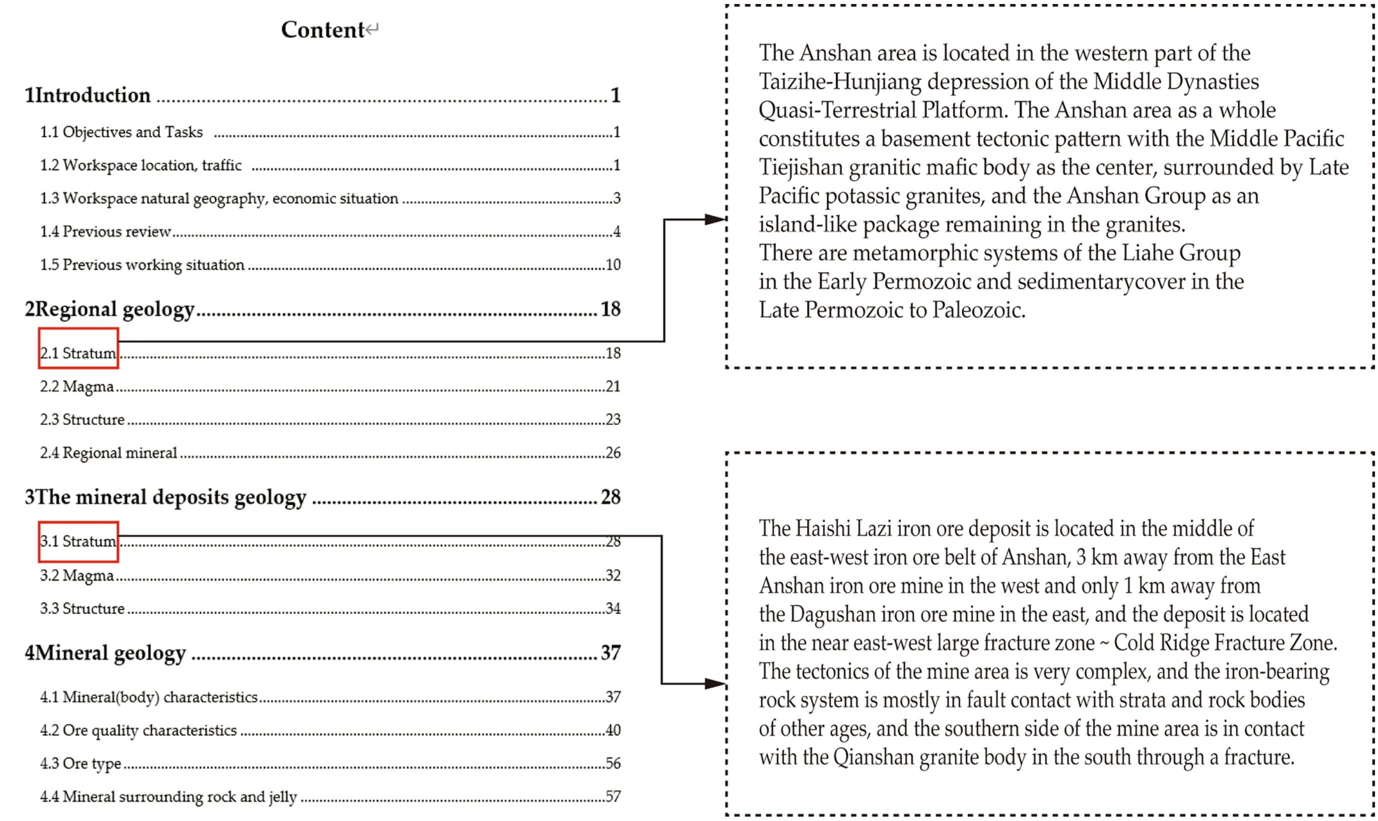
where *sd (j, i, k)* refers to that the shortest path from *j* to *k* passing through *I*; that is, *i* lies on the shortest path from *j* to *k*.

## 3. Visualization process for geological text

### 3.1. Case study project profile

In this paper, we focus on a geological report written in Chinese. Chinese geological report is an important technical document that comprehensively reflects the results of geological survey work, and is prepared according to the systematic collation and comprehensive study of various information obtained from the existing and this survey work after the completion of all the tasks given or after a phase. It generally consists of the main text of the report and various pictures, tables, attachments, etc. According to the items of geological work, the geological report can be divided into the mineral geological report, hydrological geological report, engineering geological report, environmental geological report, regional geological survey report, physical and chemical exploration report, petroleum geological report, etc. In China, geological reports are always organized according to the requirements of the China Geological Survey.

In this study, the geological report "*Exploration Report of Heishilazi Iron Deposit in Anshan City, Liaoning Province*" was selected as a geoscience case study. The project behind this report was supported by the China Geological Survey, which mainly conducted supplementary exploration of the deposit by means of drilling to identify in detail the geological, tectonic, and magmatic rock characteristics of the mine area, the scale, morphology, production, ore quality characteristics and technical properties of ore processing, and the hydrological, engineering and environmental geological conditions of the mine area in detail. This geological report consists of 10 chapters with more than 150 pages and 120,000 Chinese words. The introduction of the project is presented in Chapter 1. Chapter 2 describes the regional geology that includes the strata, magmatic rocks, structures, and regional minerals. Chapter 3 describes the mineral deposits geology, including the strata, magmatic rocks, and structures. Chapter 4 presents the description of ore body geology. Chapters 5–6 present the technical performance of ore processing and the technical conditions for mining deposits. Chapter 7 delivers the geological survey work and its quality. Chapter 8 is the resource estimation. Chapter 9 describes the study on the economic significance of deposit development. Chapter 10 concludes the project. Fig. 2 shows the structure of the selected geological report section of this



Content

1Introduction ........ 1
1.1 Objectives and Tasks ........ 1
1.2 Workspace location, traffic ........ 1
1.3 Workspace natural geography, economic situation ........ 3
1.4 Previous review ........ 4
1.5 Previous working situation ........ 10
2Regional geology ........ 18
2.1 Stratum ........ 18
2.2 Magma ........ 21
2.3 Structure ........ 23
2.4 Regional mineral ........ 26
3The mineral deposits geology ........ 28
3.1 Stratum ........ 28
3.2 Magma ........ 32
3.3 Structure ........ 34
4Mineral geology ........ 37
4.1 Mineral(body) characteristics ........ 37
4.2 Ore quality characteristics ........ 40
4.3 Ore type ........ 56
4.4 Mineral surrounding rock and jelly ........ 57

The Anshan area is located in the western part of the Taizihe-Hunjiang depression of the Middle Dynasties Quasi-Terrestrial Platform. The Anshan area as a whole constitutes a basement tectonic pattern with the Middle Pacific Tiejishan granitic mafic body as the center, surrounded by Late Pacific potassic granites, and the Anshan Group as an island-like package remaining in the granites.
There are metamorphic systems of the Liahe Group in the Early Permozoic and sedimentarycover in the Late Permozoic to Paleozoic.

The Haishi Lazi iron ore deposit is located in the middle of the east-west iron ore belt of Anshan, 3 km away from the East Anshan iron ore mine in the west and only 1 km away from the Dagushan iron ore mine in the east, and the deposit is located in the near east-west large fracture zone ~ Cold Ridge Fracture Zone. The tectonics of the mine area is very complex, and the iron-bearing rock system is mostly in fault contact with strata and rock bodies of other ages, and the southern side of the mine area is in contact with the Qianshan granite body in the south through a fracture.

**Fig. 2.** The structure of the selected geological report section of this study.

study.

### 3.2. *Experimental evaluation for Chinese word segmentation and keyword extraction*

#### 3.2.1. *Chinese word segmentation*

As described in Section 2.2, we used the hybrid corpus comprising the generic corpus (mainly the People's Daily Corpus) and geoscience corpus (specifically, geology dictionary and geological reports/documents) to build a BiLSTM-CRF-based geological word segmentation model. Providing the description of this methodology is outside the scope of our study and can be found in the work of Qiu et al. 2018b. After establishing the trained model, we used the geological word segmentation model to segment the input geological text.

To validate our approach, we constructed 2,000 sentences for the domain-geoscience corpus using regional geological reports and annotated them as a test corpus. In addition, we used the generic Chinese corpus developed by Peking University (called PKU), the geological corpus developed in this study (called GEO), and the hybrid corpus with the generic and geological corpora to train our deep learning model. The experimental results are presented in Table 1. The results of the Chinese word segmentation task using the hybrid training corpus are superior to those in the case where the single generic and geological corpora were used. Furthermore, the proposed method shows better performance, with the best values corresponding to the increase in precision by 8.45 and 1.44 points as compared to the generic and geological corpora, respectively.

**Table 1**
Performance of BiLSTM-CRF model in different corpora. BiLSTM-CRF-PKU: BiLSTM-CRF model that trained by PKU corpus, BiLSTM-CRF-GEO: BiLSTM-CRF model that trained by geoscience corpus, BiLSTM-CRF-PKU + GEO: BiLSTM-CRF model that trained by the hybrid corpus.

| Method | Precision (%) | Recall (%) | F1-score (%) |
|---|---|---|---|
| BiLSTM-CRF-PKU | 86.10 | 87.10 | 86.60 |
| BiLSTM-CRF-GEO | 93.11 | 91.02 | 92.05 |
| BiLSTM-CRF-PKU + GEO | 94.55 | 92.55 | 93.54 |

#### 3.2.2. *Keyword extraction*

To validate the improved keyword extraction method, an experiment was conducted using 43 Chinese geological reports. These reports were elaborately selected from the corpus ensuring that they were related to geology. A manual annotation process was conducted to create the gold standard keywords for the experiments. The datasets were divided into 10 folds, and we used 10-fold cross-validation for the proposed model. Six workers with sufficient knowledge of both geosciences and keyword extraction read each sentence and annotated all keywords. A voting mechanism was applied to resolve any disagreements. The basic TF-IDF method, which is the baseline and improved TF-IDF method, was used to evaluate the extraction performance and select the top 10 keywords.

The test output results of the constructed corpus are presented in Table 2. As evident from Table 2, the performance of the improved TF-IDF method is better than that of the traditional TF-IDF method. The extraction of precision and recall as measure indexes increased by 11.41% and 15.48%, respectively; additionally, the F1-score increased by 13.18%.

**Table 2**
Experimental results for the basic and improved TF-IDF approaches.

| Method | Precision (%) | Recall (%) | F1-score (%) |
|---|---|---|---|
| Basic TF-IDF method | 58.74 | 70.19 | 63.96 |
| Improved TF-IDF method | 70.15 | 85.67 | 77.14 |

### 3.3. Word-level information extraction

#### 3.3.1. TF-IDF vs. Improved TF-IDF results

After text preprocessing, it was found that the mineral exploration report included 66,617 words, 2503 unique words, and 935 sentences. In a report, the content words that include several terms reflect the key information of the document. The valuable knowledge in a report is correlated with some high-frequency content words (Hovy and Lin, 1998; Wang et al., 2018). However, some informative and meaningful words often have a low frequency (Piantadosi, 2014).

To extract the content words, we first counted the frequency of words and ranked them. Fig. 3(a) shows the extraction of the top 10 content words in descending order. The statistical method of word frequency not only obtained the informative terms (e.g., “Iron ore,”“Ore deposit,” and “Drill hole”), but it also extracted some function terms in the results (e.g., “Of,” “By,” “And,” “Exist,v and “End”). These function terms frequently appear in a report but cannot be considered as text features and represent its topic matter (Urrahman and Harding, 2012). The efficiency of the accuracy of content-word extraction can be improved after removing such words. Fig. 3(b) shows the results of the top 10 content-word frequencies in the report, and the results have been improved intuitively.

We conducted a list of experiments to extract the information content words (Fig. 4) based TF-IDF approach. The figure on the left shows the calculation result without removing the common words, and the right figure illustrates the extracted results after deleting such function and common words. All the terms were recalculated by adding the length of the word feature before attempting to recognize the greatest candidate key phrases for improving the accuracy of keyword extraction. The experimental results and score of keywords are summarized in Table 3.

In comparison to the statistical method of word frequency, the primary TF-IDF method exhibited significantly enhanced performance for informative content-word extraction. This is because terms with low frequency often contain more valuable information (Figueroa et al.,2018; Qiu et al.,2019). Furthermore, the results also demonstrated that the function and common words/terms (e.g.,“Of,”“By,”“And,”“Exist,” and “End”) have a negative influence on content-word extraction, which represents critical information in a geological report. Therefore, the content-word extraction results were selected after deleting such functions and common words (e.g., stop-words) for further analysis.

In comparison to the basic TF-IDF approach, the results extracted using the improved TF-IDF method demonstrated better representation in a geological report. For example, in typical geological text extraction, the terms “*Lepido granoblastic texture,*”“*Sericite chlorite phyllite,*” and “*Glauconitic quartzarenite*” would have an evident effect on representing the core topic from their inferred meaning.

#### 3.3.2. Visual tag cloud results

Content words are indicators that can be used to reflect the extracted information and knowledge in a report. To clearly, quickly, and comprehensively exhibit the content and information included in the geological text, such visual perspective is required to provide and display data using the data mining results effectively. Directly demonstrating the keywords/keyphrases of a document is the most intuitive method to present geological report content. Word cloud is one of the most intuitive and commonly used techniques for visualizing words. It shows a bag of words that summarize the content of the input text data in a cloud form, in which words, with font size indicating their importance, and packed together without any overlap. A word cloud is composed of content words that exceed the threshold *n* (in our experiment, *n* is set to 50) based on the corresponding frequency statistics, which presents a clear visualization of content in a geological document/report (Fig. 5).

Using the word cloud (as shown in Fig. 5), a manager/researcher can clearly and quickly obtain the following geological information included in the geological text:

(1) Words such as “Ore deposit,” “Drill hole,” and “Ore” in the graphic are highlighted with big font size, suggesting that these words have a high representation and weight. Thus, it can be easily inferred that the main content of the geological text includes deposit-related content describing the exploration and resources.
(2) The main words written on the report included information about mineral-related investigations and content, in the word cloud layout, the font size of the terms “Fracture,” “Structure,” and “Anshan” were also relatively large, indicating the main purpose of the geological text.
(3) Quantitative research conducted on the report for which this document was written were contained due to the keywords “Thickness”, “Sample”, and “Resources” provides an intuitive speculate.

In this step, the keywords extracted by the improved TF-IDF algorithm reflect the topic of the geological text to an apparent extent, and the word cloud shows an intuitive understanding of the report visually.

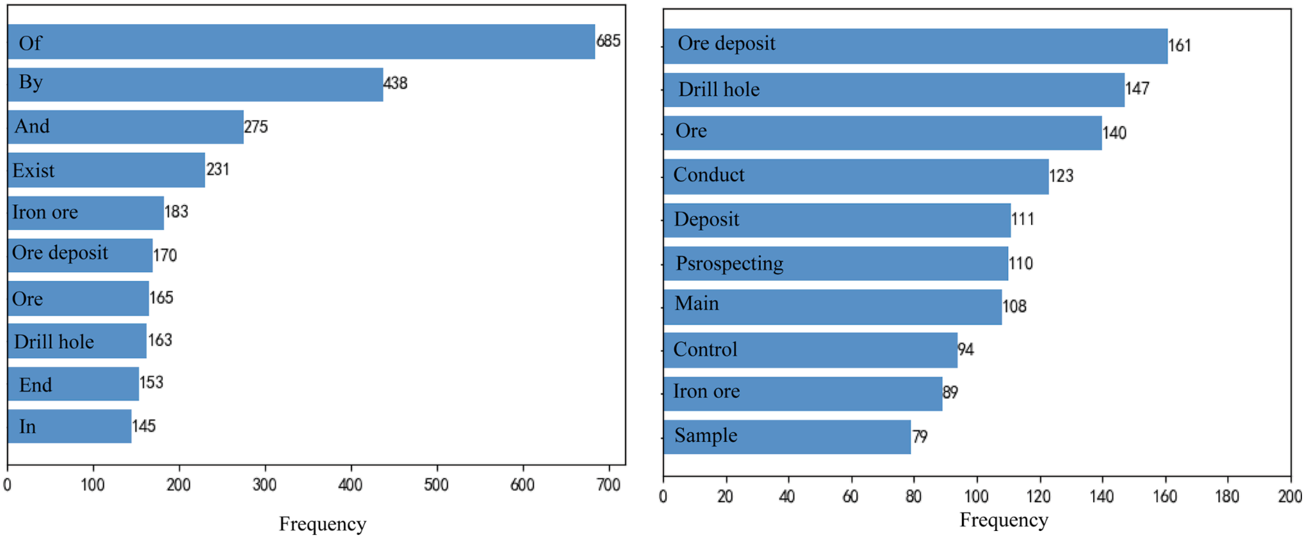


**Fig. 3.** (a) Top 10 content-word extractions based on high frequency without removing stop words. (b) Content-word extraction based on high frequency after removing stop words.

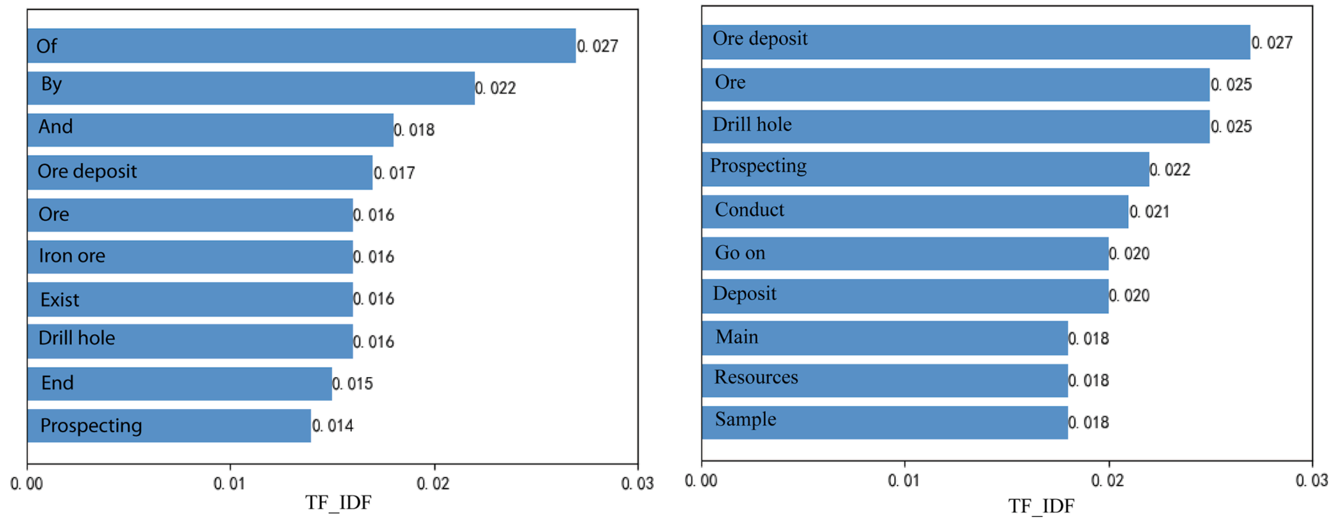


**Fig. 4.** (a) Top 10 content-word extraction based on the TF-IDF approach without removing stop words. (b) Content-word extraction is based on the TF-IDF method after removing stop words.

**Table 3**
Example of keyword extraction based on different methods for case study text analysis.

| TF<br>Keyword | Score | IDF<br>Keyword | Score | TF-IDF<br>Keyword | Score | Improved TF-IDF<br>Keyword | Score |
|---|---|---|---|---|---|---|---|
| Ore deposit | 0.012 | Drilling for mineral deposits | 6.145 | Ore deposit | 0.027 | Silky phyllite intermingled with chalcotite quartz schist | 1.000 |
| Ore | 0.011 | Mining industry | 5.740 | Ore | 0.025 | Pseudo hematite quartzite containing iron carbonate | 0.917 |
| Drill hole | 0.010 | Tectonic effect | 5.452 | Drill hole | 0.025 | Carbonate resembles hematite lean ore | 0.752 |
| Go on | 0.009 | Geological structure | 5.229 | Prospecting | 0.022 | Metamorphic tuff interlaced with marble | 0.750 |
| Deposit | 0.008 | Intrusion | 5.048 | Conduct | 0.021 | The scales have a granular, crystalline texture | 0.668 |
| Prospecting | 0.008 | Position | 4.893 | Go on | 0.020 | Dolomite chloritic quartz schist | 0.667 |
| Main | 0.008 | Attitude | 4.759 | Deposit | 0.020 | Heishilazi iron deposit | 0.587 |
| Conduct | 0.007 | Location | 4.641 | Main | 0.018 | Pseudo hematite quartzite | 0.586 |
| Iron ore | 0.007 | Geological prospecting | 4.441 | Resources | 0.018 | Institute of Geological Exploration | 0.585 |
| Sample | 0.006 | Fracture | 4.354 | Sample | 0.018 | Heishilazi Iron mine | 0.584 |

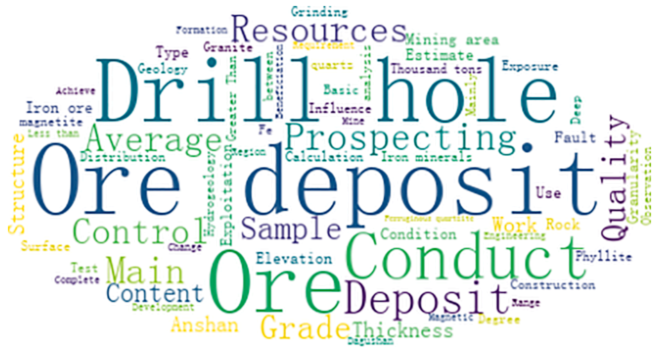


**Fig. 5.** Information visualization tag cloud of the mineral exploration reports. We have selected the top 50 words to show the visualizations. Word clouds are built from extracted words that illustrate a visual and brief representation of information stored in geological reports. The font size of the words reflects the word frequency in the geological report.

However, these methods do not illustrate the relations between the extracted content words, leading to incomplete information. Therefore, in order to present the information in the geological report text more completely, further analyses for subsequent visual depiction are necessary.

### 3.4. *Visual mapping based on keyword relevance*

The geoscience relevant information reflected by a single independent keyword/keyphrase is limited. To more completely illustrate textual information with a visual form, we need to determine the relationships between the recognized keywords. The tightness between keywords can reveal the general topic of the geological text as well as the content of the report to a certain extent. Such mapping of the knowledge domain is accomplished using a series of different graphs that depict the relationships between the knowledge development process and document structure.

#### 3.4.1. *Keyword centrality analysis*

In this study, we fed the "bipartite matrix" (as discussed previously in Section 2.3) into the UCINET software to analyze various keyword centralities (e.g., DC, CC, and BC). The results of centrality analysis are presented in Table 4. As demonstrated in Table 4, the DC and CC of the word "region" achieved the highest values of 44 and 1.105, respectively; its BC value was 1.814, which suggests that this word is near the center social network. These values indicate that the word "region" is one of the

**Table 4**
Keyword centrality statistics for the case study text.

| Keyword | Degree | Closeness | Betweenness |
|---|---|---|---|
| region | **1.705** | 1.105 | 1.814 |
| Anshan | 1.473 | 0.775 | 2.734 |
| distribution | 0.969 | 1.097 | 1.868 |
| iron ore | 0.969 | 0.854 | 1.164 |
| exposure | 0.853 | 1.101 | **3.307** |
| structure | 0.581 | 0.819 | 1.035 |
| intrusive rock | 0.504 | **1.140** | 0.509 |

cores of the social network and it can be applied as an important keyword for this geological report. The word “Luliang Period” had the lowest DC value of 0.188, revealing that this term had a minimum influence on other nodes in the social network; additionally, its CC and BC values were 0.665 and 0.004, respectively. These relatively low values indicate that “Luliang Period” is significantly far from the central network.

The visualization of centrality could assist readers/researchers in gaining an intuitive understanding of how keywords are distributed in a social network. The processed data were imported to the NetDraw software for drawing and producing a keyword map of the geological text (Fig. 5). This keyword map can clearly and intuitively reflect three types of knowledge discovery: (1) the keywords are divided into several closely related subcategories to discover new subtopics; (2) the capacity to determine the relations between these nodes provides clues for the relative importance of nodes; and (3) the capacity to ascertain the main content of each important topic and reveal the central meaning of the report.

As shown in Fig. 6, the term “Region” is linked to most other nodes and is considered to be the largest node in the social network, suggesting that it indicates the key content in the geological text representing the condition of the report. The term “Luliang Period” is located at the edge of the social network and is connected to a small node, revealing that this term weakly affects the social network. In addition, the visualization of keyword centrality illustrates the distribution of terms in the geological report. For instance, “Iron ore”, “Structure”, “Intrusive rock”, “Anshan Group,” and “Exposure” are located at the center position of the social network along with some crucial links, revealing that they compose the geological information represented by the key information of the text.

The centrality map of keywords and key phrases demonstrated in Fig. 6 shows the geological content and information by applying the intensity of different terms, expressed in terms of location (closer to the center means more likely to represent document content) and node size (larger size means more important keywords) in the social network to indicate the degree of connection between keywords. Evidently, the keywords connected to various remaining nodes with higher intensity imply their significance in the geological report and the main characteristics of the report. In addition, the association between keywords can be used to explore the current status information and the report trend. Other keywords (“Anshan” and “Iron ore belt”) in the graph (e.g., Fig. 6) can be combined to reflect more accurate and complete content and information. For instance, the keywords “Intrusive rock” and “Iron ore” express larger node sizes in the central network, suggesting that these keywords possess the maximum connections with other terms.

Motivated by this observation, a natural inference is that this geological report text focuses on iron ore and structure; thus, the related sub-themes include “Distribution,” “Anshan,””Liaohe Plain,” and “Granite.” The centrality connections of the keywords expressed by the arrow directions of several connection nodes, information content contained the representative theme, and its related relation can be delineated in the geological text. However, we cannot determine the relationship between various keywords. Based on this observation, further analyses must be conducted via multidimensional scaling.

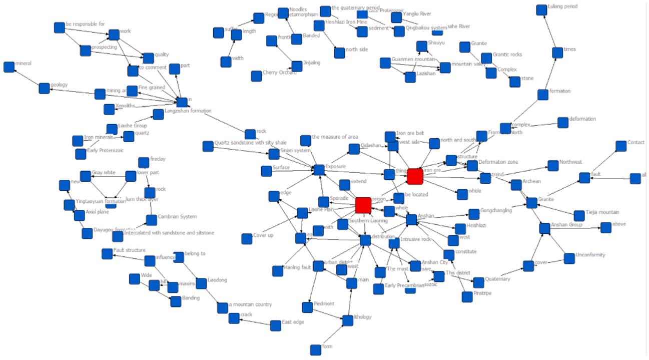


Fig. 6. Keyword centrality analysis based on the improved TF-IDF approach. Red rectangles represent more focused words and blue rectangles represent common words.

### 3.4.2. Multidimensional scaling analysis of keywords

Multidimensional scaling analysis, a multivariate analysis approach, focuses on converting the similarity data and individual differences to present a multidimensional graph; this helps in maintaining the relative relationship between the original data points.

The data (also discussed below) were imported to the NetDraw software. Their multidimensional scaling visualization is demonstrated in Fig. 7. As shown in Fig. 7, the terms “Region,” “Distribution,” and “Iron ore” are located in the central position, suggesting the three important focal points of the geological text. The critical content of the geological report mainly includes the description and distribution of regional geological reports on iron ore deposits. The visualization results of multidimensional scaling demonstrated that “Anshan,” “Iron ore belt,” “Exposure,” “Sinian system,” and “Archean” are close to the center in the multidimensional space. Meanwhile, these terms are related to “Region” (considering the spatial information), “Distribution,” and “Iron ore,” thereby revealing that they express a domain-specific problem described in the report. Motivated by this observation, a natural inference is that one of the main scattered information data and content included in the geological text is that the exploration scope of the report is the Anshan area. The related terms such as “Anshan Group,” “Archean,” and “Quaternary” suggest the distribution and geological time of the iron ore.

The keyword network of multidimensional scaling analysis can be applied to further discover geological text. The terms expressing “structure” are relatively closer to and linked with those expressing “Iron ore,” reflecting that these two terms are positively correlated in the geological text. Therefore, it is reasonable to infer that the geological time and geological structure of iron ore are mentioned in this geological text. Similarly, “From the north” and “Structure” also present a close connection, suggesting that the temporal and spatial information of the iron ore structure distribution is contained in this geological report.

## 3.5. Semantic similar word view

Word embeddings (WE), which express a distributed word representation, have been widely applied in various NLP tasks (Mikolov et al., 2012; Le and Mikolov, 2014). In this space, the closer the distance (e.g., Euclidean distance) between words, the higher the semantic similarity between them. Based on this observation, the semantics of similar words surrounding the current word (i.e., the k-nearest neighbors) were used to present the semantic meaning of the central word.

Therefore, we used semantic word clouds to demonstrate the current

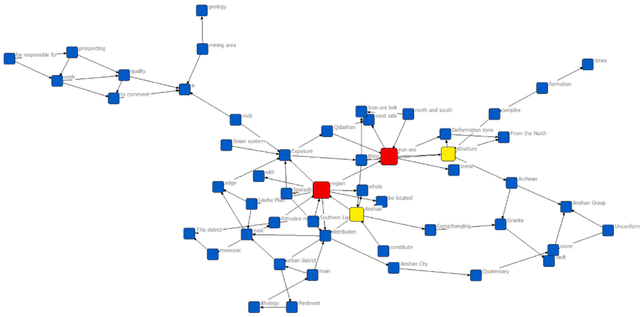


Fig. 7. Multidimensional scaling analysis results for the keywords. Red and yellow rectangles represent more focused words, and blue rectangles represent common words.

word and similar semantic words in the WE space. For the selected similar words, the cosine similarity algorithm was selected to identify the top *N* similarity words in the semantic WE space (Fig. 8). As evident in Fig. 8, the layout of the word cloud was used to demonstrate a list of neighbors around words. In the figure, the current word (marked with green) is located in the center of the graph, and the top-N similar words are surrounded by the current words using various colors. The font size of words represents the degree of semantic similarity between the current word and similar words. Different words use color to emphasize their visual contrast. For example, in Fig. 8(a), the two closest words to the current word “Iron ore” are “Mineral” and “Magnetite lean ore”; however, the nearest words to this term (“Iron ore”) are “Hematite” and “Carbonate magnetite lean ore.”.

We collected several Chinese geological texts from domain-geoscience reports/documents and pre-trained the Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2018) on the Chinese geological domain corpora. It must be noted that WE do not consider word location information. Meanwhile, at the sentence level, the word vector encoded by BERT depends on the entire sentence, which can be dynamically fine-tuned, and BERT can use the information of the following sentence to accurately represent the differences between the current encoded similar sentences; WE are therefore slightly inferior in these aspects. Therefore, BERT can better express the semantic information of sentence encoding. In this study, we used the extracted keyword as a sentence as the input of BERT. Since BERT encodes Chinese using character segmentation, our method was able to effectively reflect the similarities and differences between the keywords. We performed dimension reduction and visualization on the mineral geological report using the BERT model, as demonstrated in Fig. 9. After using the BERT, the WE space exhibited better performance for clustering. For example, the words that are most similar to “Iron ore” include “Carbonate magnetite lean ore,” “Lean iron ore,” “Iron ore belt,” “Magnetite lean ore” and “Hematite.” Based on this observation, we concluded that similar semantic words could reveal the co-occurrence information of words/terms, implying that *geological exploration* is closer in the WE space if *engineering geology* is more likely to be used.

### 3.6. Auto-generated keyword/keyphrase suggestion

Full text search and document query are also widely used to support document exploration since the very beginning of the text visualization. Instead of showing a ranked list of related documents regarding the query keywords, most of the existing visualization techniques transform the search results into a visual representation to illustrate the insight of content relationships among documents (Cao and Cui, 2016). Reports/documents can be searched using the report number as well as a set of geological keywords. We used the co-occurrence learned from the reports in the repository to provide automatic completion in the keyword search. Fig. 10 shows a user-entered term, “Siltstone,” and the resulting automatically generated suggestions. Holdena et al. (2019) only used the co-occurrence information for generating the query keyword suggestions. However, we used the co-occurrence information and the enhanced keyword extraction algorithm and combined it with keyword semantic similarity to auto-generate the keywords suggestions.

## 4. Discussion

This paper focuses on illustrating related or summarizing the content or linguistic features of a document the based on combination of text mining, NLP, and semantic oriented techniques. The proposed approach is flexible and can accommodate different types of geological reports such as regional geology reports and engineering geology reports, allowing for quick extraction of key information from the text in a visual form.

Chinese texts comprise of consecutive characters: there are no space-denoted boundaries between words (Huang et al., 2015a; Huang et al., 2015b; Deng et al., 2016a; Deng et al., 2016b; Qiu et al., 2018). Therefore, segmenting character sequences into semantically and syntactically meaningful units—words—becomes a fundamental problem for analyzing and understanding Chinese texts. The public corpus PKU includes a large amount of daily words, and the GEO corpus that we constructed comes from 98 various geological reports. The hybrid corpus was constructed to train the BiLSTM-CRF model to segment Chinese words from geological reports/documents. The experimental results proved that the model can effectively recognize the geological terms from a report and segment general terms based on the hybrid corpus.

In a geological report, words and chapters are often arranged around a single topic. The fluency and meaning of the sentences are based on the words and their co-occurrences (Firth, 1957; Wang et al., 2018). Therefore, the improved TF-IDF was used to extract a set of co-occurring content words representing the critical information of the report. This set reflected the topic of the geological text to an apparent extent, and the word cloud showed a visually intuitive understanding of the report. However, these methods cannot illustrate the relationship between the extracted words, leading to incomplete presentation of information. Note that compound terms/keyphrases are more informative and can express the topic more accurately and reasonably than single words. Ideally, the domain-specific and compound words better express the topic and content of the report. In the result of this research, compared with the traditional TF-IDF algorithm, we consider the length factor by improving the algorithm. For example, we can easily use “Heishi Lazi iron ore deposit” to express paragraph topic information. Of course, during the experiment, we also found that TF-IDF has limitations

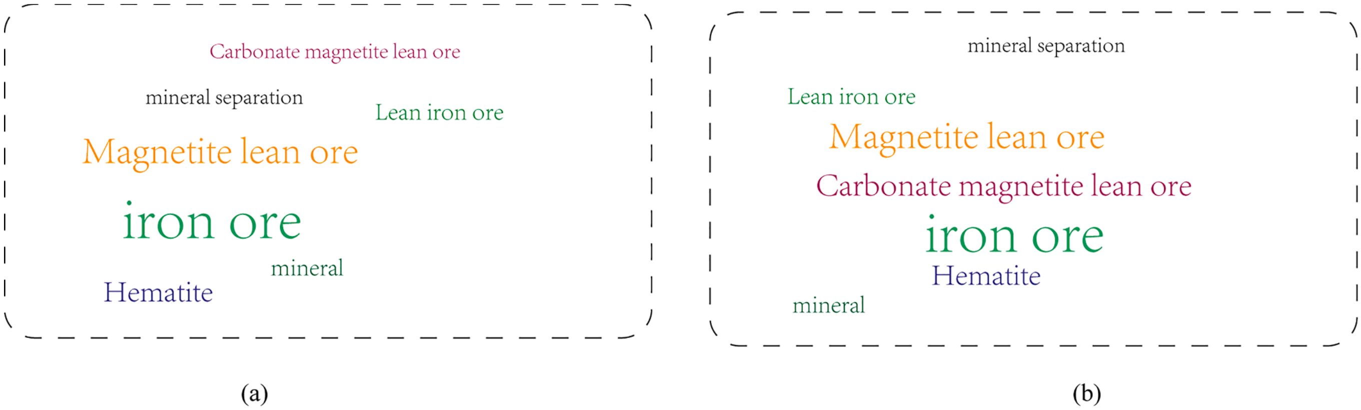


**Fig. 8.** The view of word semantic similarity. The center represents the current word part, and the other part shows the current WE space. The current word, along with several surrounding words, demonstrates the semantic meaning. (a) Semantically similar words based on WEs. (b) Semantically similar words based on BERT.

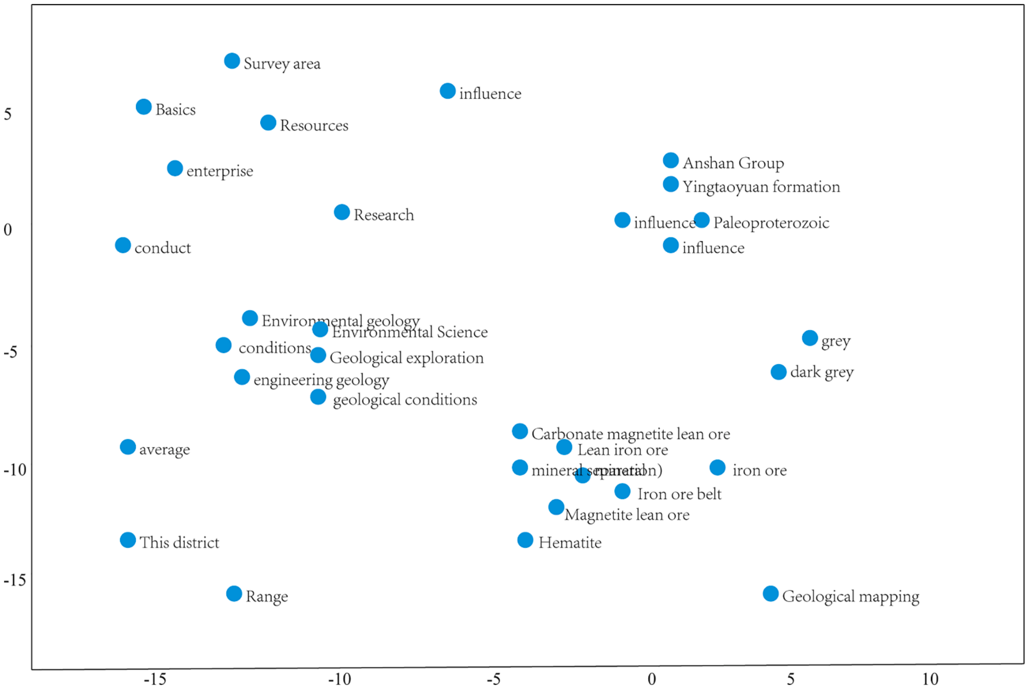
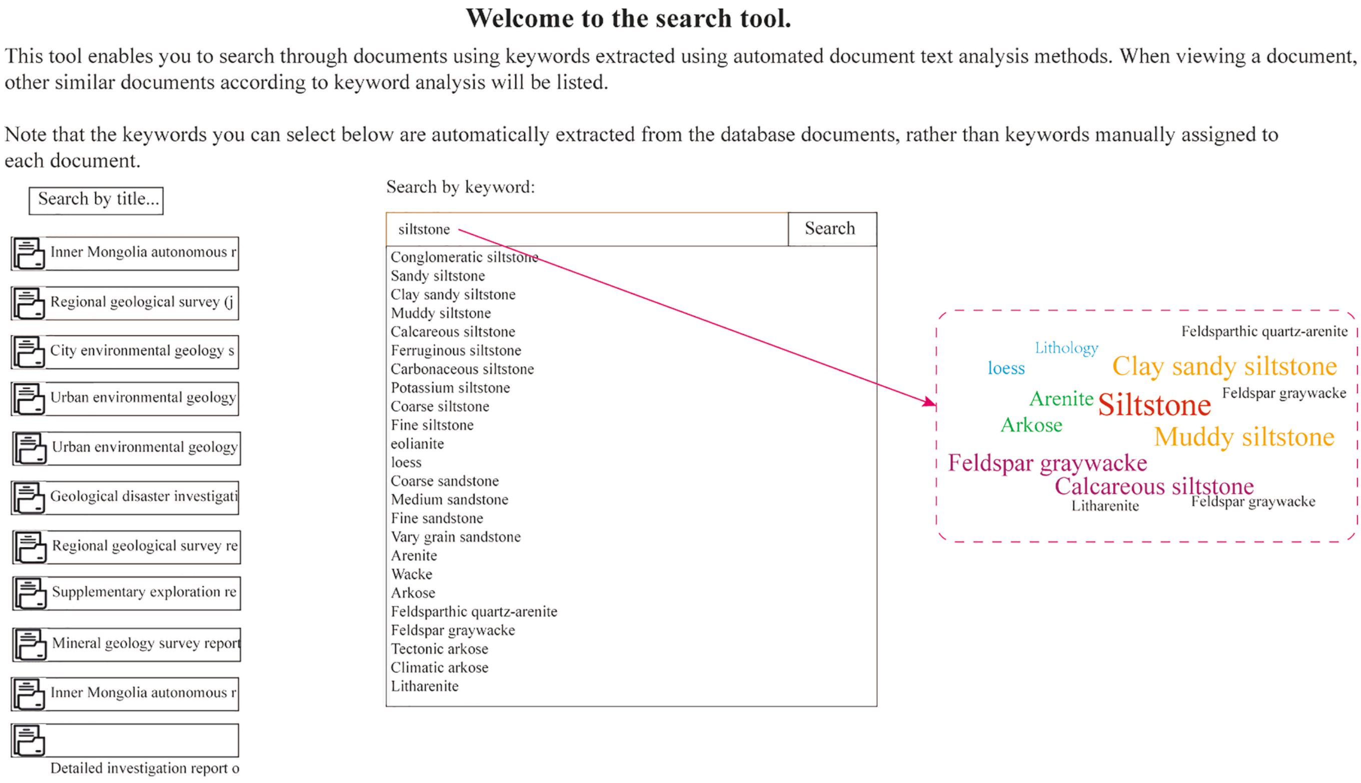


**Fig. 9.** WE visual analytics and cluster analytics based on the BERT model.



**Welcome to the search tool.**

This tool enables you to search through documents using keywords extracted using automated document text analysis methods. When viewing a document, other similar documents according to keyword analysis will be listed.

Note that the keywords you can select below are automatically extracted from the database documents, rather than keywords manually assigned to each document.

Search by title...

Inner Mongolia autonomous r
Regional geological survey (j
City environmental geology s
Urban environmental geology
Urban environmental geology
Geological disaster investigati
Regional geological survey re
Supplementary exploration re
Mineral geology survey report
Inner Mongolia autonomous r
Detailed investigation report o

Search by keyword:

siltstone Search

Conglomeratic siltstone
Sandy siltstone
Clay sandy siltstone
Muddy siltstone
Calcareous siltstone
Ferruginous siltstone
Carbonaceous siltstone
Potassium siltstone
Coarse siltstone
Fine siltstone
eolianite
loess
Coarse sandstone
Medium sandstone
Fine sandstone
Vary grain sandstone
Arenite
Wacke
Arkose
Feldsparthic quartz-arenite
Feldspar graywacke
Tectonic arkose
Climatic arkose
Litharenite

**Fig. 10.** Illustrative example of auto-generated keyword suggestions based on the input entity “Siltstone.”.

although the algorithm is faster in calculation and expresses topic information through statistical information, it cannot extract and express information from the semantic level, and further joint semantic information or domain knowledge is needed to further improve the results.

Also, automatic text summarization proves an approach to extract a short and information abstract that can be used for representing the key information from a text document (Wang et al., 2018; Ma et al., 2021a; Ma et al., 2021b). Nevertheless, since the current text automatic summarization generally uses a deep learning model and the extracted summarization is not yet an accurate representation of the document content and cannot achieve satisfactory results (Mitray et al., 1997; Narayan et al., 2018). In this research, we used text mining, NLP, and semantic-oriented techniques to extract and visualize key content from unstructured geological reports.

Note that co-occurrences and relationships between words are crucial for the complete visual presentation of the information contained in a geological report. Graphics and pictures are more descriptive and convincing than plain data and text. Therefore, we use the keyword centrality analysis to assist readers/researchers in gaining an intuitive understanding of how keywords are distributed in a social network. In the geological report chosen in this study, "Iron ore," "Structure," "Intrusive rock," "Anshan Group," and "Exposure" are located at the center of the social network and connected via crucial links. This indicates that these terms represent the main geological information of the report. It offers a quick graphical overview of a lengthy report, thereby eliminating the need to read the entire report word by word and avoiding errors caused by subjectivity during extraction. Additionally, multidimensional scaling analysis can be used to facilitate the processing of raw data, which includes various texts and symbols, by discovering structural relationships in the data. The proposed method, therefore, presents a network connecting the identified geological text keywords in the form of a graph or picture while maintaining the raw assignment data, and clusters the extracted key phrases to depict their connections as distance content in a multidimensional space.

From the results obtained, we extract the critical information from the relevant geological text and present it in the form of intuitive graphics and maps. The content words in combination with a similar semantic approach can be used to provide auto-generated keyword suggestions for searching. As such, several key information that can otherwise not be easily discovered using a simple data visualization process (or that could only be determined by time-consuming, in-depth reading) can be mined using the proposed method, and the results obtained can be utilized by managers for decision-making purposes.

## 5. Conclusions and future work

A text document refers to the data, such as a paper, a news article, or an online webpage. Visualizations designed to represent documents usually focus on illustrating how various documents are related or on summarizing the content or linguistic features of a document to facilitate an effective understanding or comparison of various documents. In this study, a framework for developing geological text visualizations was proposed and evaluated. Using information extraction, visual display, and semantic similarity analysis, the content words and information hidden in a geological report text were depicted through a list of visual graphs. In the information extraction process, the word length was comprehensively considered; subsequently, an integrated TF-IDF algorithm was used to extract a set of keywords to improve the accuracy of content information extraction, thereby representing and reflecting the key contents in the geological text. Then, using the identified keywords, the keyword centrality and multidimensional scaling were visually analyzed to present an apparent and intuitive presentation of the geological information. Finally, clustering analysis and semantic similarity analysis were performed to provide a visual perspective to realize the relationships between keywords in the geological report to ensure that researchers/managers could acquire the key contents more clearly and intuitively. In addition, some information and contents that were not easily found by a single and simple text process of visualization (or can only be decided by time-consuming deep reading) were mined, which can significantly assist managers in decision making.

The contributions of this paper can be seen from two perspectives. From the perspective of methodology, this work presents a computational framework that integrates NLP and text mining for visualizing the extracted information in the geological report. As indicated by the experiment results, integrating text mining, NLP, and semantic-oriented techniques to help visualize such content-wised information in an intuitive manner have been designed and developed. These techniques enable the discovery of actionable insights. From the perspective of the application, this work can help improve the efficiency of geological information transmission and assist decision-making, as information presented in graphical rather than the textual form can be digested in a relatively shorter time and can be more persuasive.

This study can be extended in several directions. First, other types of geological reports in the same area (such as regional geological reports, remote sensing geological reports, etc.) can be included to further explore linked information with mineral resources as the core. Currently, our experiment used only a geological report. Other data does not affect the effectiveness of the proposed computational framework, but might not express richer correlation information in a more comprehensive way. Thus, further analysis can be performed when more news geological data become openly available. Second, a more effective keyword extraction algorithm is required to combine other textual features (e.g., parts of speech) and deep learning approaches in original reports. The proposed framework has limitations in terms of semantic information extraction. Deeper natural language analysis can be performed on the textual descriptions of the geological report. Text visualization analysis should be implemented in multi-documents to integrate tables and figures in geological reports.

Computer code availability.

The original data can be downloaded from the below link: https://github.com/cugdeeplearn/Visual-analytics.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

The authors thank the editor and the two reviewers for their constructive suggestions and comments. This study was financially supported by the National Natural Science Foundation of China (42050101, 41871311), the China Postdoctoral Science Foundation (No.2021M702991), Open Research Project of The Hubei Key Laboratory of Intelligent Geo-Information Processing (No. KLIGIP-2021A01), and the Fundamental Research Funds for the Central Universities, China University of Geosciences (Wuhan) (No. CUG2106116)).

## References

Akiyoshi, M., 2008. Knowledge sharing over the network. Thin Solid Films 517 (4), 1512–1514.

Arroyo-Fernández, I., Méndez-Cruz, C.-F., Sierra, G., Torres-Moreno, J.-M., Sidorov, G., 2019. Unsupervised sentence representations as word information series: Revisiting TF–IDF. Comput. Speech Lang. 56, 107–129.

Berend, G., Farkas, R., 2010. SZTERGAK : Feature Engineering for Keyphrase Extraction. Meeting of the association for computational linguistics.

Bourne, L.M., Weaver, P., 2018. The origins of schedule management: the concepts used in planning, allocating, visualizing, and managing time in a project. Front. Eng. Manage. 5 (2), 150–166.

Chen, C.-H., 2017. Improved TFIDF in big news retrieval: An empirical study. Pattern Recogn. Lett. 93, 113–122.

Deng, K., Bol, P. K., Li, K. J., et al. (2016). On the unsupervised analysis of domain-specific Chinese texts. In Proceedings of the national academy of sciences of the United States of America: 113 (p. 6154).

Devlin J., Chang M.-W., K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.0480.

Enkhsaikhan, M., Holden, E.-J., Duuring, P., Liu, W., 2021a. Understanding Ore-Forming Conditions using Machine Reading of Text. Ore Geol. Rev. 135, 104200 https://doi.org/10.1016/j.oregeorev.2021.104200.

Figueroa, G., Chen, P.C., Chen, Y.S., 2018. RankUp: Enhancing graph-based keyphrase extraction methods with error-feedback propagation. Comput. Speech Lang. 47, 112–131.

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

Gao, J., Li, M., Wu, A., Huang, C., 2005. Chinese Word Segmentation and Named Entity Recognition: A Pragmatic Approach. Computat. Linguist. 31 (4), 531–574.

Gao, X., Singh, M.P., Mehra, P., 2012. Mining Business Contracts for Service Exceptions. IEEE Trans. Serv. Comput. 5 (3), 333–344.

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

Holden, E.-J., Liu, W., Horrocks, T., Wang, R., Wedge, D., Duuring, P., Beardsmore, T., 2019. GeoDocA – Fast analysis of geological content in mineral exploration reports: A text mining approach. Ore Geol. Rev. 111, 102919.

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.

Khare, V.R., Chougule, R., 2012. Decision support for improved service effectiveness using domain-aware text mining. Knowl.-Based Syst. 33 (33), 29–40.

Le, Q. V., & Mikolov, T. (2014). Distributed representations of sentences and documents. arXiv preprint arXiv1405.4053.

Lima, L.A., Gornitz, 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.

Ma, K., Tian, M., Tan, Y., Xie, X., Qiu, Q., 2021. What is this article about? Generative summarization with the BERT model in the geosciences domain. Earth Sci. Inf. 1–16.

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.

Marzouk, M., Enaba, M., 2019. Text analytics to analyze and monitor construction project contract and correspondence. Autom. Constr. 98, 265–274.

Mee, A., Homapour, E., Chiclana, F., Engel, O., 2021. Sentiment analysis using TF-IDF weighting of UK MPs' tweets on Brexit. Knowl.-Based Syst. 228, 107238.

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

Mikolov, T., Deoras, A., Povey, D., et al. (2012). Strategies for training large-scale neural network language models. In Automatic speech recognition and understanding (pp. 196–201). IEEE.

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

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

Peters, S.E., Husson, J.M., Wilcots, J., 2017. The rise and fall of stromatolites in shallow marine environments. Geology 45 (6), 487–490.

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.

Qiu, Q., Xie, Z., Wu, L., 2018a. A cyclic self-learning Chinese word segmentation for the geoscience domain. Geomatica 72 (1), 16–26.

Qiu, Q., Xie, Z., Wu, L., Li, W., 2018b. DGeoSegmenter: A dictionary-based Chinese word segmenter for the geoscience domain. Comput. Geosci. 121, 1–11.

Qiu, Q., Xie, Z., Wu, L., Li, W., 2019. Geoscience keyphrase extraction algorithm using enhanced word embedding. Expert Syst. Appl. 125, 157–169.

Qiu, Q., Xie, Z., Wu, L., Tao, L., 2020a. Automatic spatiotemporal and semantic information extraction from unstructured geoscience reports using text mining techniques. Earth Sci. Inf. 13 (4), 1393–1410.

Qiu, Q., Xie, Z., Wu, L., et al., 2020b. Dictionary-based automated information extraction from geological documents using a deep learning algorithm. Earth Space Sci. 7 (3), e2019EA000993.

Wang, C., Ma, X., Chen, J., Chen, J., 2018. Information extraction and knowledge graph construction from geoscience literature. Comput. Geosci. 112, 112–120.

Wu, L., Xue, L., Li, C., Lv, X., Chen, Z., Jiang, B., Guo, M., Xie, Z., 2017. A knowledge-driven geospatially enabled framework for geological big data. Int. J. Geo-Informat. 6 (6), 166.

Sun, J., Lei, K., Cao, L., Zhong, B., Wei, Y., Li, J., & Yang, Z. (2020). Text visualization for construction document information management. Automation in Construction.

Salton, G., Wong, A., Yang, C.S., 1974. A vector space model for automatic indexing. ACM.

Urrahman, N., Harding, J.A., 2012. Textual data mining for industrial knowledge management and text classification: A business-oriented approach. Expert Syst. Appl. 39 (5), 4729–4739.

Yang, J., Kim, E., Hur, M., Cho, S., Han, M., Seo, I., 2018. Knowledge extraction and visualization of digital design process. Expert Syst. Appl. 92, 206–215.

Yu, L., Lu, F., Liu, X.L., et al., 2016. A method of context enhanced keyword extraction for sparse geo-entity relation. J. Geo-informat. Sci. 18 (11), 1465–1475.

Zhuang, C., Li, W., Xie, Z., Wu, L., 2021. A multi-granularity knowledge association model of geological text based on hypernetwork. Earth Sci. Inf. 14 (1), 227–246.

Ma, K., Tian, M., Tan, Y., Xie, X., Qiu, Q., 2021. What is this article about? Generative summarization with the BERT model in the geosciences domain. Earth Sci. Inf. 15 (1), 21–36.

Wang, B., Wu, L., Li, W., Qiu, Q., Xie, Z., Liu, H., Zhou, Y., 2021. A semi-automatic approach for generating geological profiles by integrating multi-source data. Ore Geol. Rev. 134, 104190.

Enkhsaikhan M, Holden E J, Duuring P, et al. Understanding ore-forming conditions using machine reading of text. Ore Geology Reviews, 2021, 135: 104200.

Cao, N., Cui, W. (Eds.), 2016. Introduction to Text Visualization. Atlantis Press, Paris.

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

Deng, K.e., Bol, P.K., Li, K.J., Liu, J.S., 2016b. On the unsupervised analysis of domain-specific Chinese texts. Proc. Natl. Acad. Sci. 113 (22), 6154–6159.

Narayan S, Cohen S B, Lapata M. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. arXiv preprint arXiv:1808.08745, 2018.