Chapter

Text Mining to Facilitate Domain Knowledge Discovery

Chengbin Wang and Xiaogang Ma

Abstract

The high-precision observation and measurement techniques have accelerated the rapid development of geoscience research in the past decades and have produced large amounts of research outputs. Many findings and discoveries were recorded in the geological literature, which is regarded as unstructured data. For these data, traditional research methods have limited functions for integrating and mining them to make knowledge discovery. Text mining based on natural language processing (NLP) provides the necessary method and technology to analyze unstructured geological literature. In this book chapter, we will review the latest researches of text mining in the domain of geoscience and present results from a few case studies. The research includes three major parts: (1) structuralization of geological literature, (2) information extraction and visualization for geological literature, and (3) geological text mining to assist database construction and knowledge discovery.

Keywords: text mining, word segmentation, geological literature, visualization, knowledge discovery

1. Introduction

Geoscience is a knowledge-intensive discipline. It has not only domain-specific terminology but also a deep intersection with mathematics, chemistry, and physics, which form a series of distinctive subdisciplines, such as geophysics, geomathematics, geochemistry, paleobiology, and more [1–3]. Thanks to the rapid development of detection techniques in the micro- and macroscales in the past decades, both the volume and quality of geoscience data have been improved greatly. A feature of detection-based research is using the extrapolation method to explore the Earth. For instance, geochemists use local geochemical data to invert the process of Earth evolution and geodynamics [4, 5]. The diverse big data and improved computer software and hardware enable an opportunity to understand the evolution of Earth system using simulation and data mining methods [6].

Many geoscience research outputs are recorded in the form of literature, making text data an integral part of geoscience big data [7]. Important information and knowledge are recorded in unstructured textural form and thus hidden in the geological literature. Nowadays, the advanced Web technologies promote the publication process of academical literature and accelerate literature exchange globally. Researchers can easily assemble publications of focused topics. In this regard, geological literature has become a big “mineral resource” for data mining and provides tremendous opportunities for new knowledge discovery. In recent years, the open data initiative has promoted government agencies, scientific organizations, and academic publishers to provide literature archives for nonprofit reuse; some are even open and free. For instance, the US Geological Survey (USGS) and China Geological Survey (CGS) have published outputs of geological survey investigation online [8, 9]. Elsevier and Springer have provided application programming interfaces (API) for developer and scientists to access metadata, full text, and conduct text mining [10, 11]. We anticipate that more geological literature will be made available by publishers, government agencies, research organizations, and individual scientists in the coming years.

In a recent review article [12], Gil and other scholars proposed a research agenda of intelligent systems that will result in fundamental new capabilities for understanding the Earth system. Automated information extraction and integration from published literature is listed as a key research direction in the agenda. Domain-specific text mining can be regarded as a topic in interdisciplinary fields, such as geoinformatics, ecoinformatics, and bioinformatics. Conventionally, text mining is a research topic in computer science. The new development in interpreted programming language and the wide-spreading open-source packages and libraries enable scholars in various disciplines to quickly learn the latest algorithms and apply them to their domain-specific researches. There are many widely used open and free libraries in text mining, such as TensorFlow [13], DeepDive [14], Caffe [15], CNTK [16], and MXnet [17]. Even if a researcher has only the basic skills in programming, he or she will be able to make a deep research using these libraries.

Text mining contains the following major steps: data collection and preprocessing, identification of entities and their links, and knowledge representation. Data collection can take place in many forms. For example, one can require permission to get data from a database or publisher and can also retrieve data form the Web by a data extractor. The obtained data from different sources may be recoded in diverse formats, such as text files and scanned images. It is necessary to transform the data into an organized, computer-readable format. For instance, we can use the optical character recognition (OCR) to identify characters and words from the scanned images of a book or paper. After the preprocessing, the next step is to analyze the information and meaning of the text data. In the early stage, many researchers have tried to use automatic text summarization to extract a concise and informative abstract that covers the key information of a text document [18–20]. Nevertheless, due to the limitation of poor readability, automatic text summarization has yet to achieve satisfactory results.

Knowledge graphs, as proposed by Google, are semantic networks with directed graph structure, which have provided new ideas to extract and represent the text information. The words representing the major entities and relationships carry the key information in a document. Therefore, a text document can be represented by a knowledge graph to show a list of entities and their relationships. The structured knowledge graph is a specific data base and can be further analyzed and visualized by graph methods. Every entity is regarded as a graph node, and the relationship between two nodes is represented as an edge. The graph visualizes the nodes and edges to represent the implicit information network of a document. In recent years, many open knowledge graphs have been constructed based on text information, such as Google knowledge vault [21], DBpedia [22], Freebase [23], YAGO [24], Wikidata [25], OpenIE [26], and NELL [27]. These knowledge graphs devote to acquire entities and their links for various topics during the construction. In contrast, some domain-specific knowledge graphs only focus on one or a few topics. For instance, the MusicBrainz [28], UniProtKB [29], and GeoName [30] are knowledge graphs in the music, biology, and geography fields, respectively. The recent development of NLP and semantic technologies also provide new methods and tools for building knowledge graphs [14, 31, 32].

In this chapter, we will review the development of text mining in the domain of geoscience in recent years and present results of a few case studies. Comparing with other disciplines, the domain of geoscience still has limited applications of NLP and text mining. We hope the presented work will be of interested to the text mining community, and we anticipate more innovative text mining applications will appear in geoscience and other disciplines in the near future.

2. Structuralization of geological literature

Text data are usually consisted of sentences written by authors with personal understandings and opinions. Compared to metadata, text data are characterized by ambiguity, polysemy, and irregular input in the natural language. It is difficult for computers to read and understand. It is necessary to segment a piece of text into semantic word sequences for further computer processing. English and other Latin languages have relatively simple morphology, especially inflectional morphology, and are segmented by spaces between words naturally. For those languages, it is often possible to ignore the word segmentation task entirely. In contrast, there is no space between words in a few other languages, such as Chinese. It is difficult for a computer to identify the boundary of a meaningful word or phrase in Chinese [33, 34]. The methods of Chinese word segmentation were classified into dictionary-based, statistically based, and hybrid approaches [33]. The statistically based methods include machine learning and deep learning methods, such as hidden Markov model (HMM), maximum entropy Markov model (MEMM), conditional random fields (CRF), and long short-term memory (LSTM).

From another perspective, the methods of word segmentation can be divided into generic and specific domain methods according the usage scenarios. In the generic domain, because of the shortcomings of word segmentation rules, some new words, especially the professional terms, will be regarded as out-of-vocabulary and cannot be identified correctly. Geology, as a knowledge-intensive discipline, has a systematic domain-specific terminology. Most of geologic terms are not familiar with the public. Geological literature including the geological terms has their unique characteristics. For instance, the geological literature is always organized according to some fixed format and contains lots of professional geologic terms that only people with a background knowledge can read and understand. The geological literature is dominated by descriptive sentences and has little ambiguity in information expression. In geological literature written in Chinese, it is also featured by mixed writing of Chinese and English terms as well as compound terms consisted of multiple geological terms [2, 7]. The text data in the natural language are sequence data; the word usage and combination are only influenced by the context. Based on the characteristics of text data, machine learning method (e.g., CRF) and deep learning method of neural network (e.g., neural network (CNN), LSTM) have been introduced to segment geological literature in Chinese in recent 2 years with successful results [7, 34–36].

2.1 Conditional random fields

For a random vector (e.g., in NLP), the joint probability is a high-dimensional distribution, which oversteps the processing power of an ordinary computer and is difficult to monitor during data processing. To reduce the data size, the high-dimensional distribution is divided into a series of production of conditional probability based on the independence hypothesis [37]. The probabilistic graphical model is a graph to describe independence relationship between multivariate in a high-dimensional probabilistic model, thus to reduce computer load. The probabilistic graphical model includes both directed and undirected models. The directed graphical model indicates there is a causation relationship between the variables, such as Bayesian networks. The variables in undirected graphical model have dependency with each other, such as Markov networks and CRF, which is different from the causation relationships.

CRF model is a discriminative graph model, while HMM is a generative graph model. The role of CRF model is to create the discriminant boundaries similar to the support vector machine model, which has a wide usage in the fields of NLP and bioinformatics. Compared with HMM and the maximum entropy model (MEM), the CRF model improves the accuracy and addresses the drawback of label bias [38, 39]. Text data are unstructured sequence data. The structuralization of geological text is a process of word segmentation or named entity recognition (NER), which divides the geological text into a series of semantic words. For natural language, the text is only influenced by the context, which is consistent with the assumption condition of the CRF model. The assumption condition is that multivariable obeys the Markov property. In other word, the label of part of speech at *n* position in NLP only has relationship with the word or character at *n-1* position. From the point of view of the graph model, *Yv* is a subset of *V* nodes set in the graph $G = (V, E)$; the following equation is established:

$$p(Y_v|X, Y_w, w \neq v) = p(Y_v|X, Y_w, w \sim v) \tag{1}$$

where $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$ is the words or characters of text data in NLP, $w \sim v$ denotes neighbor nodes of node v in the graph, and Y is the label set of part-of-speech $\{B, E, M, S\}$. For NLP, the graphical structure is chain-structured (**Figure 1**) [14–16].

According the factorization of joint probability distribution of undirected graph, the CRF model can be written as

$$p(\mathbf{Y}|\mathbf{X}) = \frac{1}{Z(\mathbf{X})} \prod_i e^{\sum_k \lambda_k f_k(\mathbf{X}, Y_{i-1}, Y_i, i)} = \frac{1}{Z(\mathbf{X})} e^{\sum_i \sum_k \lambda_k f_k(\mathbf{X}, Y_{i-1}, Y_i, i)} \tag{2}$$

In which i is the node position, k denotes the sequence number of feature function, and λ_k is the weight parameter. In Eq. (2), the feature function can be expressed in Eq. (3), which contains information of transfer and status features.

$$f = \sum_i^T \sum_k^M \lambda_k f_k(\mathbf{X}, Y_{i-1}, Y_i, i) \tag{3}$$

In CRF-based word segmentation, Wang et al. [7] designed a two-step workflow to segment geological literature in Chinese. First, a hybrid corpus was created using

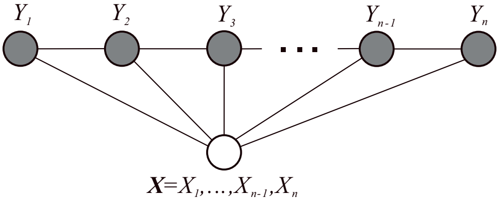

Figure 1.
A chain-structured CRF graph [7, 40].

dictionary matching and manual label methods on the basis of geological literature in CNKI, geology dictionary, TCCGMR (the terminologies and classification codes of geology and mineral resources), and a generic corpus of Peking University. Second, the segmentation rules were trained to build geological word segmentation model by the hybrid corpus, and then the trained model containing word segmentation rules was used to segment geological literature in Chinese. The workflow is shown in **Figure 2**.

In that study, a geology dictionary of 11,000 geological terms, the TCCGMR of 80,000 geological terms, and the generic corpus of Peking University were used to build the hybrid corpus. By this way, geological knowledge was introduced into the corpus to train the rules of word segmentation of geological literature. It is the most notable feature compared with other Chinese word segmentation machine. The three parameters of precision, recall, and F-scores were used to evaluate the performance of CRF-based word segmentation in that work. The result was showed in **Figure 3**. The hybrid corpus combining a generic corpus and a geological corpus has a better performance than either the generic corpus or the geological corpus alone. The precision of the hybrid training reaches 94.14%, which is 7.84% and 0.52% higher than that of CRF-PKU and CRF-GEO, respectively. The recall of hybrid corpus reaches 91.40%, which is 9.30% and 0.41% higher than that of CRF-PKU and CRF-GEO, respectively. The F-score of the hybrid corpus reaches 92.75%, which is 8.60% and 0.46% higher than that of CRF-PKU and CRF-GEO, respectively.

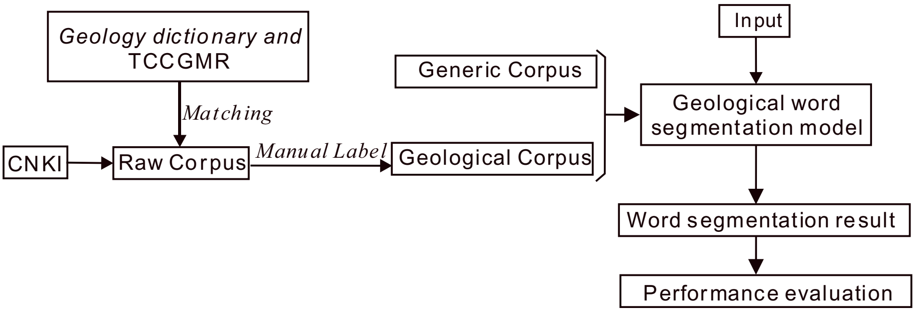

Figure 2.
Workflow of CRF-based word segmentation for geological literature in Chinese.

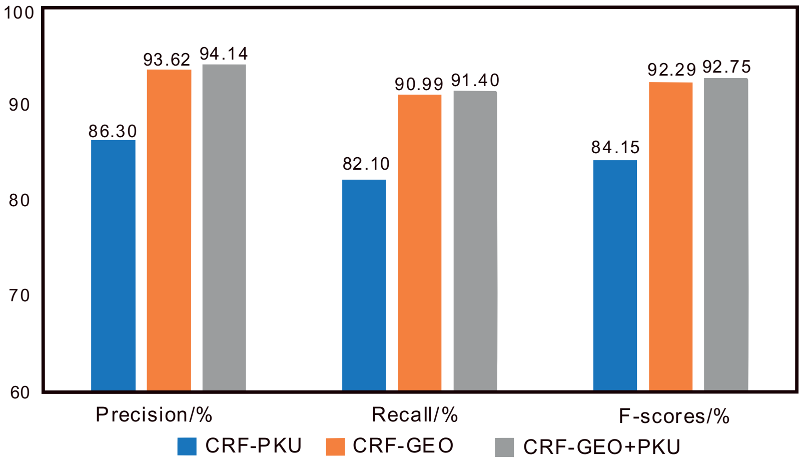

Figure 3.
Performances of the CRF model in different corpus. CRF-PKU, generic corpus of Peking University; CRF-GEO, geological corpus; CRF-GEO + PKU, the hybrid corpus combining generic and geological corpora.

2.2 Long short-term memory

The text data are consisted of a series of sequential words or characters, which can be regarded as a special data of time series and can be processed by the methods used in the time series analysis. Words or characters in text data are not completely independent but are connected to and influenced by the adjacent words or characters. In the model of neural network, it contains three basic compositions: input layer, hidden layer, and output layer. The layers of ordinary neural networks are linked with each other by weights. The nodes in a same layer are independent and have no link with each other. If the ordinary neural network methods are used to process text data, the semantic information of context will be missing. Recurrent neural network (RNN) has a short memory by nodes connecting in the hidden layer, which can receive information from self-cell and other cells. RNN has been used in the fields of NLP and automatic speech recognition [41, 42]. RNN model has the drawback of vanishing gradient problem, which means RNN model only obtains the information that is limited in the adjacent node position [43]. To address this challenge, the LSTM model designed input gate, output gate, and a forget gate to obtain information of far nodes and regulate the information flow between the cells [44] (**Figure 4**).

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t]) + b_i \tag{4}$$

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t]) + b_f \tag{5}$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_c x_t + W_c h_{t-1} + b_c) \tag{6}$$

$$ot = \sigma(W_o \cdot [h_{t-1}, x_t]) + b_o \tag{7}$$

$$h_t = o_t{}^* \tanh(c_t) \tag{8}$$

in which i, f, c, and o denote input gate, forget gate, cell vector, and output gate, receptively. σ denotes the activation functions. W denotes weight matrices and bias vector parameters which need to be learned during the training.

Qiu et al. [36] proposed a geological literature segmenter based on the Bi-LSTM model. The segmenter was carried out by the following stages (more details can be seen in the reference article):

1. Corpus construction: The corpus from domain-generic and domain-specific texts is collected and constructed.

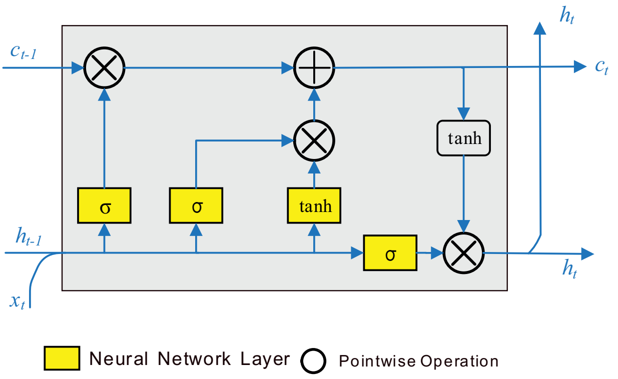

Figure 4.
The cell of LSTM [43, 45].

2. Words grouped: Each word is grouped based on frequency and a ranking algorithm.

3. Random extraction and combination: Each group of words in the previous step is extracted and joined together randomly.

4. Training: With the previous processing, sentences are formed via combination based on deep learning.

5. Testing and output: The resulting segmentation is post-processed and output.

In this research work, the significant highlight is that the training corpus is random. The segmentation rule was learned from the words and their corresponding sequences of the training corpus. The training corpus did not have any manual label information. The precision, recall, and F-scores reach 86.1%, 87.1%, and 86.6%, respectively. Compared with Wang et al. [7], the performance of CRF-based method is better than the Bi-LSTM-based segmenter based on the performance reported in their papers. But the Bi-LSTM-based method has a strong ability of identifying new words. The rate of out-of-vocabulary word identification reached 71.1%.

3. Text information visualization and knowledge discovery

3.1 Information visualization of a single geological literature

The nodes of content word and their links are the carrier of literature information and knowledge. In a large open knowledge graph, the key information was stored in in a triple format. Moreover, the bigram is also widely used in the text information representation. Wang et al. [7] used the bigram graph to represent the single geological literature.

The visualization was built based on the "from," "to," and "weight" variables. The variables of "from" and "to" indicate the sequence of content words in the content word corpus. In the content-word pairs, the former content word is defined as "from" variable, and the latter is defined as "to" variable. Their weights were defined by the co-occurrence frequency of content-word pairs. The bigram graph was used to visualize the nodes of content words and their links.

In geological exploration, the anomaly information of geology, geochemical exploration, geophysical exploration, and remote sensing is important clues for mineral prospecting [46]. To state different anomaly information, literatures of geological exploration will have significant features in the term of word frequency. **Figure 5** shows the main information hidden in a single literature of geophysical exploration. In this visualization, geological terms (e.g., *aeromagnetic*, *gravity*, *magnetic*) and geophysical data processing terms (e.g., *inversion*, *horizontal gradient*, *information*) are all linked to the term *anomaly*. The visualization represents the hidden key knowledge in the geological literature.

3.2 Geological text mining for discovering ore prospecting clues

Geology research not only reveals the earth evolution and promotes our understanding of the Earth but also has a close relationship with the human society. One of the important roles of applied geology is to discover mineral deposits and provide raw material for economic construction and development. In the long geological

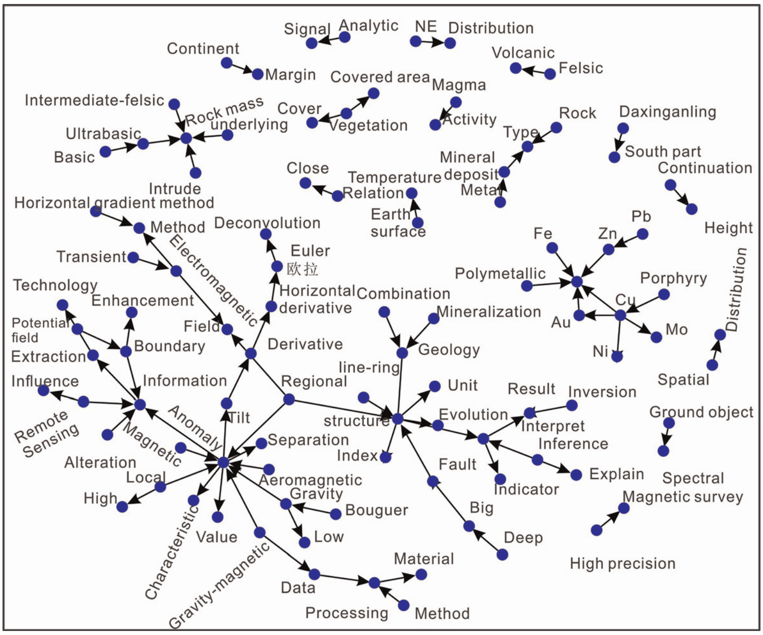

Figure 5.
Bigram graph of content words in the whole literature represents the key information in the geological report (n > 10).

history, mineral deposits were formed with large-scale geological events and were buried in the depth of Earth crust. If the mineral deposits were not broken down under the erosion of weathering after mineralization, there are ways to discover them. In the earlier days, geologists discovered mineral deposits by identifying the rock outcrops associated with mineralization. Then, along with many technological developments, the geochemical exploration, geophysical exploration, and remote sensing were also used to improve the result of mineral prospecting and mineral exploration. In recent years, GIS-based and three-dimensional mineral prospect mapping has been used in mineral exploration. Through those technologies, multisource anomalies, such as geochemical anomalies, geophysical anomalies, geological anomalies, and remote sensing anomalies, can be determined.

The anomaly information is usually derived from structured numeric data. The structured numeric data are only one part of geological big data. The majority of geological big data are unstructured, such as text and image. Previous mineral exploration mainly depends on derived information from the structured numeric data. Some important information related to the mineral prospecting and exploration is hidden in the unstructured text, such as host rock, alteration types, geological setting, ore-controlled factors, geochemical and geophysical anomaly patterns, and location. The favorable information extraction and identification from geological literature are a big challenge for conventional research methods. The NLP-based text mining provides a chance to address this challenge.

Li et al. [35] used the CNN method to classify geological text data into four categories (geology, geophysics, geochemistry, and remote sensing) on three scales (word, sentence, and paragraph). Their work extended the work on Chinese word segmentation and text preprocessing to the domain of mineral exploration. These four categories represent four types of mineral exploration information. Compared with word and paragraph scales, the sentence scale has the best performance. In their work, the *precision*, *recall* and *F-scores* of text classification reach 93.68%, 93.50%, and 92.68%, respectively. Then a co-occurrence matrix was utilized to extract content words and their relationships as nodes and links from the classification result and to visualize the information in a knowledge graph. By this way, four categories of favorable information for mineral prospecting and exploration were expressed in a bigram graph and a chord graph.

3.3 Geological text mining to assist database construction and knowledge discovery

The microfossil at 4280 million years old found in Quebec, Canada, may be the oldest fossil as so far [47]. In the Earth's history, biological evolution has a close corresponding with the geological evolution. The existence of biology depends on specific physical and chemical conditions, such as oxygen content and temperature. In other words, different biotypes and biocenoses indicate the conditions of different earth environments. The fossils were formed along with the sedimentary environment and are the footprint left by the biosphere. Each fossil records some biological information, such as biological morphology and living environment. Paleontologists always study the fossils to explore the earth environment evolution. A single fossil cannot indicate biological and geological evolution. The conclusions of such evolution are based on a series of comparative studies of fossils in different geological times and settings.

The Paleobiology Database (PBDB; http://paleobiodb.org) contains systematic and detailed fossil information, which make it a necessary infrastructure for fossil comparative researches. The PBDB is one of the most successful fossil databases, which was founded nearly two decades ago. Now it has become an open and active community for different research agendas. In the initial stage, the fossil records in the PBDB were from original fieldworks and extracted from published literature manually. As the rapid development of digital publication, the manual data entry for fossil information became tedious and less efficient and was not able to deal with the massive amounts of new and legacy publications. To address this challenge, PaleoDeepDive [46], a machine reading and learning system, was developed to extract fossil information from literature. This system uses the factor graph and NLP technologies to identify fossil entities and their semantic relationships. The extracted results were stored in the form of triples inside a knowledge base. Compared with the manual fossil data entry, the output of PaleoDeepDive has an obvious advantage in terms of quantity. Moreover, the change trend (e.g., taxonomic diversity and genus-level turnover) has a high corresponding relationship with the manual data entry [48]. The extracted fossil records have been used to update the PBDB. Now, the PBDB is not just a paleobiology database, it also provides WebGIS-based interface for fossil information retrieval and query. It also provides R library, API, and a mobile APP for researchers and the general public to use. Based on the PDBD, a series of high-quality research papers have been published to improve our understanding about the Earth. For instance, Peters et al. [49] analyzed the rise and fall of stromatolites in North America and divided the marine environment into three phases based the change of stromatolites.

The application of GeoDeepDive is still ongoing. Macrostrat (https://macrostrat.org/), a collaborative platform for geological data exploration and integration, was constructed based on the results that GeoDeepDive extracted from massive amounts of scientific literature. By April 2018, Macrostrat has contained 33,903 properties of geological units distributed across 1474 regions in North and South America, the Caribbean, New Zealand, and the deep sea, more than 180,000 geochemical and outcrop-derived measurements, all the fossil records in PBDB, and more than 2.3 million bedrock geologic map units from over 200 map sources [50].

4. Conclusion

In this chapter, we reviewed the latest developments of NLP techniques in the domain of geoscience to accelerate knowledge discovery from geological literature and deepen our understanding about the Earth. From the review, it was concluded that the researches of text mining in geoscience are still in the early stage. Most current researches focus on the literature structuralization and simple information extraction at a single document scale. The information integration and knowledge discovery from the big data of geological literature require further work and will lead to a lot of innovative research topics and applications.

Acknowledgements

The authors thank Prof. Abdelkrim El Mouatasim and Dr. Kristina Kardum for their organization for this book and their assistance. This work was supported by the National Key R&D Program of China (2017YFC0601500, 2017YFC0601504) and Fundamental Research Funds for the Central Universities, China University of Geosciences (Wuhan) (No. 162301182382).

Author details

Chengbin Wang[1]* and Xiaogang Ma[2]

1 School of Earth Resources, China University of Geosciences, Wuhan, China

2 Department of Computer Science, University of Idaho, Moscow, ID, USA

*Address all correspondence to: wangchb@cug.edu.cn

References

[1] Wang C, Ma X, Chen J. Ontology-driven data integration and visualization for exploring regional geologic time and paleontological information. Computers and Geosciences. 2018:12-19. DOI: 10.1016/j.cageo.2018.03.004

[2] Wang C, Ma X, Chen J. The application of data pre-processing technology in the geoscience big data. Acta Petrologica Sinica. 2018;**34**(2): 303-313. (in Chinese with English abstract)

[3] Ma X. Data science for geoscience: Leveraging mathematical geosciences with semantics and open data. In: Daya Sagar B, Cheng Q, Agterberg F, editors. Handbook of Mathematical Geosciences. Cham: Springer; 2018. pp. 687-702. DOI: 10.1007/978-3-319-78999-6_34

[4] Balch WM. Calcium carbonate measurements in the surface global ocean based on moderate-resolution imaging spectroradiometer data. Journal of Geophysical Research. 2005;**110** (C07001):1-21. DOI: 10.1029/2004jc002560

[5] Liu Y, Gao S, Hu Z, Gao C, Zong K, Wang D. Continental and oceanic crust recycling-induced melt–peridotite interactions in the trans-North China Orogen: U–Pb dating, Hf isotopes and trace elements in zircons from mantle xenoliths. Journal of Petrology. 2010;**51** (1–2):537-571. DOI: 10.1093/petrology/egp082

[6] Guo H, Liu Z, Jiang H, Wang C, Liu J, Liang D. Big earth data: A new challenge and opportunity for digital Earth's development. International Journal of Digital Earth. 2016;**10**(1):1-12. DOI: 10.1080/17538947.2016.1264490

[7] Wang C, Ma X, Chen J, Chen J. Information extraction and knowledge graph construction from geoscience literature. Computers and Geosciences. 2018;**112**:112-120. DOI: 10.1016/j.cageo.2017.12.007

[8] USGS. Mineral Resources Data System (MRDS) [Internet]. Available from: https://mrdata.usgs.gov/mrds/

[9] CGS. GEOCLOUD 2.0 [Available from: http://geocloud.cgs.gov.cn]

[10] Elsevier. Elsevier Developers-Text Mining [Internet]. Available from: https://dev.elsevier.com/tecdoc_text_mining.html

[11] Springer. Text and Data Mining at Springer Nature [Internet]. Available from: https://www.springernature.com/gp/researchers/text-and-data-mining

[12] Gil Y, Hill M, Horel J, Hsu L, Kinter J, Knoblock C, et al. Intelligent systems for geosciences. Communications of the ACM. 2018;**62**(1):76-84. DOI: 10.1145/3192335

[13] Google. TensorFlow 1.12.0 [Internet]. Available from: https://github.com/tensorflow/tensorflow/releases/tag/v1.12.0

[14] Zhang C. DeepDive: a data management system for automatic knowledge base construction[thesis]. Madison: University of Wisconsin-Madison; 2015

[15] Jia Y, Shelhamer E. Caffe Tutorial [Internet]. Available from: http://caffe.berkeleyvision.org/tutorial/

[16] Microsoft. The Microsoft Cognitive Toolkit [Internet]. Available from: https://www.microsoft.com/en-us/cognitive-toolkit/

[17] Apache. MXNet A flexible and efficient library for deep learning [Internet]. Available from: https://mxnet.apache.org/

[18] Hu B, Chen Q, Zhu F. LCSTS: A Large Scale Chinese Short Text Summarization Dataset. arXiv preprint arXiv:150605865. 2015

[19] Luhn HP. The automatic creation of literature abstracts. IBM Journal of Research and Development. 1958;**2**(2): 159-165

[20] Nallapati R, Zhou B, Gulcehre C, Xiang B. Abstractive text summarization using sequence-to-sequence rnns and beyond. arXiv preprint arXiv:160206023. 2016

[21] Dong X, Gabrilovich E, Heitz G, Horn W, Lao N, Murphy K, et al. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM; 2014

[22] Lehmann J, Isele R, Jakob M, Jentzsch A, Kontokostas D, Mendes PN, et al. DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web. 2015; **6**(2):167-195. DOI: 10.3233/SW-140134

[23] Bollacker K, Evans C, Paritosh P, Sturge T, Taylor J. Freebase: A collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. AcM; 2008. DOI: 10.1145/1376616.1376746

[24] Suchanek FM, Kasneci G, Weikum G. Yago: A core of semantic knowledge. In: Proceedings of the 16th International Conference on World Wide Web. ACM; 2007. DOI: 10.1145/1242572.1242667

[25] Vrandečić D, Krötzsch MJCotA. Wikidata: A free collaborative knowledgebase. 2014;**57**(10):78-85. DOI: 10.1145/2629489

[26] Stanovsky G, Dagan I. Open IE as an intermediate structure for semantic tasks. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). 2015

[27] Mitchell T, Cohen W, Hruschka E, Talukdar P, Yang B, Betteridge J, et al. Never-ending learning. Communications of the ACM. 2018;**61**(5):103-115. DOI: 10.1145/3191513

[28] Hemerly J. Making metadata: The case of MusicBrainz; 2011. DOI: 10.2139/ssrn.1982823

[29] Boutet E, Lieberherr D, Tognolli M, Schneider M, Bansal P, Bridge AJ, et al. UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: How to use the entry view. Methods in Molecular Biology. 2016;**1374**

[30] Maltese V, Farazi F. A semantic schema for GeoNames [Internet]. 2013. Available form: http://eprints.biblio.unitn.it/4088/1/techRep004.pdf

[31] Tseng Y-H, Lee L-H, Lin S-Y, Liao B-S, Liu M-J, Chen H-H, et al., editors. Chinese open relation extraction for knowledge acquisition. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. Vol. 2: Short Papers. 2014

[32] Zheng X, Li S, Feng J, Lin M, Song H, Zhang S. FudanDNN: A Deep Learning Framework with Easy-to-use GUI [Internet]. Available from: https://github.com/FudanDNN/FudanDNN

[33] Gao JF, Li M, Wu A, Huang CN. Chinese word segmentation and named entity recognition: A pragmatic approach. Computational Linguistics. 2005;**31**(4):531-574. DOI: 10.1162/089120105775299177

[34] Huang L, Du YF, Chen GY. GeoSegmenter: A statistically learned Chinese word segmenter for the

geoscience domain. Computers and Geosciences. 2015;**76**:11-17. DOI: 10.1016/j.cageo.2014.11.005

[35] Li S, Chen J, Xiang J. Prospecting information extraction by text mining based on convolutional neural networks–a case study of the Lala copper deposit, China. IEEE Access. 2018;**6**:52286-52297. DOI: 10.1109/access.2018.2870203

[36] Qiu Q, Xie Z, Wu L, Li WJ. DGeoSegmenter: A dictionary-based Chinese word segmenter for the geoscience domain. Computers and Geosciences. 2018;**121**:1-11. DOI: 10.1016/j.cageo.2018.08.006

[37] Sutton C, McCallum A. An introduction to conditional random fields. Foundations and Trends® in Machine Learning. 2012;**4**(4):267-373

[38] Lafferty JD, McCallum A, Pereira FCN. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. 655813. Morgan Kaufmann Publishers Inc; 2001. pp. 282-289

[39] Pinto D, McCallum A, Wei X, Croft WB. Table extraction using conditional random fields. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval; Toronto, Canada. 860479. ACM; 2003. pp. 235-242

[40] Wallach HM. Conditional Random Fields: An Introduction [Internet]. Available from: http://dirichlet.net/pdf/wallach04conditional.pdf

[41] Yin W, Kann K, Yu M, Schütze H. Comparative study of CNN and RNN for natural language processing. 2017. arXiv:1702.01923

[42] Graves A, Liwicki M, Fernández S, Bertolami R, Bunke H, Schmidhuber J, et al. A novel connectionist system for unconstrained handwriting recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2009;**31**(5): 855-868. DOI: 10.1109/TPAMI.2008.137

[43] Bengio Y, Simard P, FRasconi P. Learning long-term dependencies with gradient descent is difficult. IEEE transactions on neural networks. 1994;**5** (2):157-166

[44] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation. 1997;**9**(8):1735-1780. DOI: 10.1162/neco.1997.9.8.1735

[45] Olah C. Understanding LSTM networks. [Internet]. 2015. Available form: http://colah. github. io/posts/2015-08-Understanding-LSTMs

[46] Wang C, Rao J, Chen J, Ouyang Y, Qi S, Li Q. Prospectivity mapping for "Zhuxi-type" copper-tungsten polymetallic deposits in the Jingdezhen region of Jiangxi Province, South China. Ore Geology Reviews. 2017;**89**:1-14. DOI: 10.1016/j.oregeorev.2017.05.022

[47] Dodd MS, Papineau D, Grenne T, Slack JF, Rittner M, Pirajno F, et al. Evidence for early life in Earth's oldest hydrothermal vent precipitates. Nature. 2017;**543**(7643):60. DOI: 10.1038/nature21377

[48] Peters SE, Zhang C, Livny M, Re C. A machine reading system for assembling synthetic paleontological databases. PLoS One. 2014;**9**(12): e113523. DOI: 10.1371/journal.pone.0113523

[49] Peters SE, Husson JM, Wilcots J. The rise and fall of stromatolites in shallow marine environments. Geology. 2017;**45**(6):487-490. DOI: 10.1130/G38931.1

[50] Peters SE, Husson JM, Czaplewski J. Macrostrat: A platform for geological data integration and deep-time earth crust research. Geochemistry, Geophysics, Geosystems. 2018;**19**(4): 1393-1409. DOI: 10.1029/2018GC007467