Construction of Deposit Model-oriented Knowledge Graph

Yan Qun1, a, Xue Linfu 1, b*, Liu Zeyu2, c, Gao Xin1, d, Wang Rui1, e, Dai Junhao 1, f
1. College of Earth Sciences, Jilin University, Changchun, Jilin, 130012, China
2. Tianjin Academy for Intelligent Recognition Technologies, Tianjin 300457, China
First author’s e-mail: yanqun19@mails.jlu.edu.cn
*Corresponding author’s e-mail: xuelf@jlu.edu.cn

Abstract: It has become a hot research direction under the background of artificial intelligence to effectively organize geological big data and to mine corresponding potential value. The knowledge graph, as a new form of knowledge representation, can effectively organize and represent data. This paper presents the deposit model ontology according to the triune framework of "ore-forming geological body, ore-forming structure and structural plane, and mineralization mark" by centering on the deposit model, sets up the knowledge graph with the "branch intersection, top-down action and step-by-step layering" method, expresses and mines big data for geological text in the form of knowledge graph, and establishes a special knowledge graph for deposit model with semantic processing capacity and open interconnection capacity, providing a routine to match geological data from new perspective and showing the integrity and relevance of genesis features of deposits on deeper level. The results show this method can effectively realize semantic correlation and constraints between deposit model ontology, as well as retrieval and reasoning of correlation objects, and prove the usefulness of the research workflow in this paper.

1. Introduction
Multi-source heterogeneous data, loose organizational structure and other issues in in the big data environment bring new challenges to the knowledge organization. The knowledge graph presents a new way to explore the matching information data resources from new perspective, and provides users with more intuitive and rapid way to obtain information. The knowledge graph technology has become an important component of artificial intelligence technology with the development of artificial intelligence and natural language processing technology, delivering great application value in intelligent information services involving the intelligent question answering and personalized recommendation [1]. Generally, knowledge graph is mainly divided into general knowledge graph and specialized knowledge graph, general knowledge graph focuses on the breadth, emphasizes the integration of more entities, and normally shows insufficient accuracy. While specialized knowledge graph is mainly depending on the data of specific industries, is featured with specific characteristics of industry knowledge, normally the entities are complexly correlated to relations, the data organization modes are abundant, and it is more conducive to deeply mining potential industry knowledge, delivering important application value for effective organization and development of relevant big data information [2]. Google, Microsoft, Baidu and other Internet companies have taken the lead in constructing large-scale general knowledge graphs to provide semantic searches based on entities and relations, and the construction of specialized knowledge graphs in financial and medical fields.
provided important supports for deeply mining potential industry knowledge. Meanwhile, a great deal of resource description framework (RDF) data of semantic network resources is published and shared, linked open data (LOD) projects have been carried out in an all-round manner\textsuperscript{[3-4]}. It has become a hot topic in fields to construct a complete knowledge graph. It remains to be a challenge to organize and understand existing text information effectively in the field of Geosciences. It is discovered in the analysis and geological interpretation based on geological quantitative data that new knowledge is the frontier direction of intelligent development. Accordingly, this paper is intended to study the method to construct specialized knowledge graph for the deposit model, express and mine permanently accumulated geological text big data in the form of knowledge graph, store abundant geological heterogeneous information data in more intensive manner, demonstrates the integrity and correlation of genetic characteristics of types of deposits in deeper level, and provides supports to further mine hidden information of traditional geological data.

2. Method

2.1. Chinese Word Segmentation
Word segmentation of Jieba, a method suitable for Chinese word segmentation, mainly supports three segmentation modes including exact mode, full mode and search engine mode, and it supports traditional word segmentation, custom dictionary and MIT authorization protocol. The principle of Jieba segmentation is to firstly construct a prefix dictionary based on statistical dictionary, then segment the input sentence with prefix dictionary to get all the segmentation possibilities, construct a directed acyclic graph according to the location of segmentation; calculate, maximum probability, and locate maximum segmentation combination based on through dynamic programming algorithm. For unlisted words, HMM model based on Chinese character word-forming ability is adopted, Viterbi algorithm is used for calculation and part-of-speech tagging, and keywords are respectively extracted based on TF-IDF and textrank models\textsuperscript{[5]}. 

![Figure 1. Routes of Word Segmentation](image)

2.2. Keyword extraction
The term frequency inverse document frequency (TF-IDF) is a common information retrieval and data mining technology, in which statistical method is used to calculate the importance of a word in the
document set \(N\). Let \(N\) be the number of a given document set, TF-IDF is used to calculate the weight of Entry \(I\) in the document for a given document \(F\), and the formulas are shown as follows:

\[
wt = TF \times IDF
\]

\[
IDF_i = \log \frac{N}{n_i}
\]

Where, TF denotes the frequency of \(I\)'s occurrence; IDF denotes the reverse frequency of Document \(f\); \(n_i\) is the frequency of \(I\) occurring in the document set, obviously it is difficult to reduce the influence of the term frequency on keyword extraction based on IDF alone. In order to reduce the impact of IDF, the concept of term frequency should be added, and complete TF-IDF expression is shown as follows:

\[
wt = TF_{i,j} \times IDF_i
\]

\[
TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}
\]

Where, \(TF_{i,j}\) represents the frequency of \(i\) in Text \(j\), and normally (4) is used to normalize (3). The denominator of Formula (2) is meaningless when Term \(I\) does not exist in the document set, so it needs further processing, the formula as follow:

\[
IDF_i = \log \frac{N}{n_i + 1}
\]

According to formula (3), it is concluded that TF-IDF value of the term is greater if such term appears frequently in the document but less in the whole document set, and the words with larger document indexing is advised to be selected to filter out common terms.

2.3. Construction of Deposit Ontology

Ontology is the expression of concepts of knowledge in a specific domain and relation forms between such concepts. Abstract concept set is used to describe common characteristics of all things in a certain field, the semantic relations between these concepts are analyzed, and the ontology is constructed into a knowledge system with hierarchy and semantic relevance by establishing the relations between concepts \([7-12]\). The contents of the deposit model ontology mainly include two parts, namely conceptual ontology set and relational ontology set, which are respectively mapped to the entity set and relation set in the knowledge graph for deposit \([13]\). The deposit ontology model is constructed by the method of "intersecting by parts, acting from top to bottom, and layering step by step". The core concept ontology set is extracted by centering on ore-forming geological body, ore-forming structure and structural plane, and mineralization mark, the primary layer is filled with "ore-forming elements" and "element characteristics", and the auxiliary layers are successively filled as the supplementation or annotation content of the primary layer. Relational ontology set not only covers the affiliation relation between basic sets but also includes complicated relations between geological things. Relational ontology sets are divided into four categories in this paper, namely functional relation set, structural relation set, attributive relation set and spatial relation set. Functional relations indicate key roles of ore-forming elements on the formation and location of the deposit, such as ore controlling structure, ore-bearing strata, controlled by, etc. Structural relations indicate the relations between ore-forming elements, such as intersection, involvement, interpenetration, etc. Attributive relations indicate relationship between entities, such as is, attribute to, include, etc. Spatial relations indicate the relations in terms of spatial orientation, such as located in, adjacent to, border on, etc.
3. Results

3.1. Construction of corpus
The main data for construction of ore deposit model corpus include journal papers, reports and books in CNKI. The *Theories and Methods of Prospecting Prediction in Exploration Area* is used as main data in this paper to extract information and establish the corpus for the deposit model. This book consists of two parts including general discussion and specific discussion. The compilers mainly work on practical services of mineral exploration, and implement substantial mineral exploration verification services by acquiring types of data from macro to micro perspective. The theories and methods of ore-forming prospecting prediction model in triune exploration area for "ore-forming geological body, ore-forming structure and structural plane, and mineralization characteristic mark" are proposed in general discussion \[14-15\]. This paper mainly introduces the geological models for prospecting prediction of the main types of deposits in China, such as sandstone type uranium deposits, hydrothermal sedimentary lead-zinc deposits, magmatic hydrothermal ore deposits, metamorphosed sedimentary iron deposits and marine volcanic copper lead-zinc deposits.

Dictionary-based word segmentation is a kind of mechanical word segmentation method, and its core is to achieve the word segmentation of text by comparing entries on the basis of preset dictionary. Normally, the word length is positively correlated with the amount of information it can express. The geological dictionary-based segmentation method is mainly in this paper to train Chinese word segmentation rules\[16\]. The geological dictionary set out in this paper is an extremely comprehensive dictionary, and the contents of this dictionary includes more than 55000 terms from general geology, geophysical and geochemical prospecting, mineral deposit science, tectonics, petrology, paleontology, etc. The number of terms is still expanding. In the process of the experiment, types of stop words on the Internet including "Word Bank of Harbin Institute of Technology on Stop Words, Machine Learning Intelligent Laboratory Stop Words Bank of Sichuan University", "Baidu Stop Words" and other types of stop words are added and integrated, and a relatively comprehensive list of stop words are extracted. The stop words in the text are removed, effectively improving the quality of corpus information. The accuracy and recall are integrated, and the effect is measured with F-score.

\[
F = \frac{(a^2 + 1) \times P \times R}{(a^2 \times P) + R} \quad (6)
\]

In Formula (6), \(P\) denotes the accuracy, indicating the proportion of target words accounting for recognizable words, and it is used to determine the accuracy of the word segmentation results. \(R\) is the recall rate, indicates the proportion of target words in total words, and is used to determine comprehensiveness of target words. The importance of \(P\) and \(R\) is determined by \(a\). The value of \(a\) in this study is 1, which means that the accuracy and recall are treated with equivalent importance.
Table 1. Precision of Word Segmentation

|                          | Accuracy/% | Recall/% | F-score/% |
|--------------------------|------------|----------|-----------|
| Stuttering word segmentation | 77.41      | 73.34    | 75.32     |
| Dictionary + stuttering word segmentation | 84.47      | 81.23    | 82.82     |

The results are shown in table 1, the accuracy of word segmentation based on dictionary increases by 7.06% and F-score increases by 7.5%. However, many proper nouns are overlapping due to the particularity of geological morpheme, such as "layered structure", which is composed of two nouns including the "layered" and "structure". Therefore, the boundary between words is fuzzy in the process of word segmentation, impairing the accuracy of word segmentation, while the "layered structure", "structure" and "layered" deliver definite geological meanings in different geological context, leading to the uncertainty of word semantics and influencing the effects of word segmentation.

3.2. Construction of Knowledge Graph for Deposit Model

The knowledge graph describes knowledge resources and corresponding carriers through the graph theory and visualization technology, it can be regarded as a semantic web, and is technically represented as a technology stack, and normally it is considered to be composed of four levels: knowledge extraction, knowledge fusion, knowledge modeling, and knowledge visualization[17]. The knowledge graph represents the interrelation through the "entity key-value", and its basic unit is triple "entity-relation-entity'. The visualization of deposit model ontology is mainly implemented through Neo4j in this paper. Neo4j database is a high-performance NOSQL graphic database storing data...
online, and it is a high-performance engine for mature development of database. The database is characterized with strong consistency, it is easy to establish relations between data, and it also comes with cypher language with good scalability [18].

The deposit model ontology in this paper is a summary of ore-forming regularities from two parts including the ore deposit genesis and prospecting prediction by integrating ore-forming geological bodies, ore-forming structures and structural planes, and mineralization marks for types of deposits. The concept set and relation set of deposit model ontology are mapped to the entity set and relation set of knowledge graph, the keywords of two sets are screened and matched respectively according to the rules summarized manually, and the construction and visualization of the knowledge graph of deposit model is realized by establishing incidence relations of sets.

Figure 5. Display of keywords of entity set (left) and relation set (right)

The entity set is summarized as conceptual entity or physical entity in this paper, the conceptual entity is the common feature of ore-forming elements of types of deposits, and the physical entity mainly demonstrates as the ore-forming element characteristics of typical ore occurrences that have been proved. For the determination on the incidence relation between entities, conceptual entities are mainly related through attribute relations, and the incidence relations between physical entities and conceptual entities are related mainly through prior knowledge in the field of geology. By integrating the correlation of entity set and relation set, concept entities in the upper layer and physical entities in the lower layer are constructed step by step to make two branches including the deposit genesis and ore prospecting prediction subject to confluence constantly, the extension of related knowledge for the graph system is constantly expanded according to the importance of ore-forming element, the knowledge coverage and relation complexity of the graph are continuously expanded, realizing the objective to construct the geological big data correlation knowledge graph system.
Table 2. Triple of Deposit Model for "Medium & Low Temperature Hydrothermal Gold Deposit" (partial)

| Node | Label | Name                  | Attribute            | Trend       | Relation           |
|------|-------|-----------------------|----------------------|-------------|--------------------|
| 1    | Ore-forming element | Ore-forming geological body | Rock mass           | From 1 to 4 | Including          |
| 2    | Ore-forming element | Ore-forming structure  | Faulted structure   | From 2 to 5 | Developed          |
| 3    | Ore-forming element | Mineralization mark    | Mineralization feature | From 3 to    | Including          |
| 4    | Element feature    | Rock mass              | Guojialing Rock Mass | From 4 to 6 | Is                 |
| 5    | Element feature    | Faulted structure      | NE-directed fracture | From 5 to 7 | Developed          |
| 6    | Rock mass type     | Guojialing Rock Mass   | ——                  | From Rock mass type to | ——                |
| 7    | Fracture feature   | NE-directed fracture   | ——                  | From fracture feature to | ——                |

In the knowledge graph for mineral deposit model, the entity corresponds to an independent concept ontology which serves as a node, each entity shows 0 or n attributes. These attributes exist in the form of "key-value". The main content of the attribute is the explanatory or supplementary information of node semantics. By establishing semantic correlation between two nodes, the visualization can be directly realized in the form of the triple with the entity-relation-entity. The correlation and constraint of knowledge are strictly directed, "from" represents the source of the node, "to" represents the trend of the node, the "entity-relation-entity" organizes the data in the format of "from" - relationship - "to". Adds appropriate "label" to the entity according to the characteristics of hierarchical elements to highlight the hierarchy of the data set for purposes of retrieval and query. Table 2 is the triple table for "entity-relation-entity" for some deposit models. The triple form for Node 1 and Node 4 is (Node 1: Ore-forming element {Name: Ore-forming geological body, attribute: Rock mass}) - Including - (Node 4: Element feature {Name: Rock mass, attribute: Guojialing Rock Mass}).

Figure 6. Visualization of Knowledge Graph for Deposit Model

A total of 101 deposit model types are constructed in this paper, including 307 types of entities, 202 types of relation and 4946 triples with "entity-relation-entity", only limited triplet data output display is provided due to the limit of single execution output data. Figure 6(left) shows the visualization results of knowledge graphs for some deposit models, and the ore-forming information directly related to deposits can be displayed in detail through precise matching. Figure 6 (right) shows the triple information of "Daqiao Gold Deposit" in detail, a typical ore occurrence of medium & low-temperature hydrothermal gold deposit. It is possible to directly grasp the mineralization and ore-forming law of Daqiao Gold Deposit from the graph, hence the knowledge graph of deposit model constructed in this paper is more conducive to the organization and expression of data than traditional data storage mode, can also be used to analyze types of ores and mine potential relations of ore deposits, and provides a new idea for realizing the objective of knowledge driven prospecting.
prediction.

4. Conclusion
In this paper, a method of constructing knowledge graph for deposit model is proposed, which provides a new method for acquiring, storing, organizing, managing and displaying geological knowledge. The following conclusions are drawn.

(1) The knowledge graph centering on the deposit model is strongly logically, and it not only establishes relations between geological elements and deposits, but also expands a new space for mining geological information.

(2) The stop words vocabulary and geological dictionary are further expanded in the process of extracting deposit model information, which improves the efficiency of text cleaning, and makes it possible to conveniently process large-scale geological text data.

(3) The deposit model ontology is visualized effectively through Neo4j, semantic correlations and constraints between different deposit model ontology entities are intuitively showed, providing a new way to organize unstructured heterogeneous data.

According to the study described in this paper, the following prospects are proposed.

(1) The geoscience data mining based on the knowledge graph can greatly expand the knowledge coverage of knowledge graph with constant expansion of the entity scale in the geoscience knowledge graph.

(2) High efficiency of inter-entity correlation determines the strictness of logical structure of knowledge graph. The task of entity correlation will become increasingly more complex due to the inconsistency caused by the differences in the construction of knowledge graphs, and it is necessary to develop more efficient and noise-resistant entity correlation thread mining methods.

Acknowledgments
This research was funded by the Geological Survey Project of China Geological Survey (DD20160050) and was allowed to be published as its phased result. We would like to thank reviewers for carefully reading this paper and their very useful comments.

References
[1] Fu Shan, LV Ailin, Yan Shu. Concept and Application of Knowledge Graph [J], Information and Communications Technology and Policy, 2019 (05): 10-13
[2] Liu Chang, Wang Bin, Xue Jie, Chen Xu, Zhan Weiwei, Xiong Xin. A Implementation Method for Semantic Constraint of Real-time System based on Knowledge Graph [J], Journal of Chinese Mini-Micro Computer Systems, 2019,40 (12): 2644-2649
[3] Liu Qiao, Li Yang, Duan Hong, Liu Yao, Qin Zhiqiang. Overview on Construction of Knowledge Graph Technology [J], Journal of Computer Research and Development, 2016,53 (03): 582-600
[4] Yuan Man, Chu Bing, Chen Ping, Study on Semantic Standard in Construction of Knowledge Graph [J], Information studies: Theory & Application, 2020,43 (03): 131-137
[5] Zeng Xiaqin. Implementation of Chinese Stuttering Word Segmentation Technology based on Python [J]. China Computer and Communication (theoretical edition), 2019, 31 (18): 38-39 + 42
[6] Yang Ying, Dai Bin. Chinese keyword Extraction Method based on Multiple Features [J], Computer Applications and Software, 2014,31 (11): 109-112
[7] Duan Yufeng, Huang Sisi, Study on Ontology Construction Method [J], Journal of Intelligence, 201534 (11): 139-144
[8] Chinchor N, Marsh E. Muc-7information extraction task definition[C]//Proc of the 7th Message Understanding Conf. Philadelphia :Linguistic Data Consortium, 1998:359-367
[9] Rau L F. Extracting company names from text[C]//Proc of the 7th IEEE Conf on Artificial Intelligence Applications. Piscataway, NJ: IEEE, 1991:29-32
[10] PUJARA J, MIAO H, GETOOR L, et al. Knowledge graph identification [M]. Berlin: Springer, 2013.

[11] VASSILIADIS P, QUIX C, VASSILIOU Y, et al. Data warehouse process management [J]. Information System, 2001, 26(3): 205-236.

[12] STUDER R, BENJAMINS V R, FENSEL D. Knowledge engineering principles and methods [J]. Data and Knowledge Engineering, 1998, 25(1/2): 161-197.

[13] Li Ming. Study on Ontology Composite Mapping [J], Computer Engineering and Applications, 2010, 46(17): 160-162 + 184

[14] Ye Tianzhu, LV Zhicheng, et al, Theories and Methods for Prospecting Predictions in Exploration Area [M], Geological Publishing House, Beijing, 2014: 234-252

[15] Ye Tianzhu, LV Zhicheng, et al. Theory and Method of Prospecting Prediction in Exploration area (general) [M]. Geological Publishing House, Beijing, 2014: 118-130

[16] Chen Kaichang, Study on Chinese Word Segmentation in Natural Language Processing Technology [J], China Computer & Communication (Theoretical Edition), 2016 (19): 61-63

[17] Wang, Chengbin, Xiaogang, et al. Information extraction and knowledge graph construction from geoscience literature [J]. Computers & Geosciences, 2018.

[18] Yu Fanghua. APOC and ALGO, Extended Guidelines to APOC Graph Database [M]. Tsinghua University Press, Beijing, 2020: 3-5