Knowledge map construction of multi-source heterogeneous contaminated site data

Xingchen Li[1, 2], Jianqin Zhang[1, 2*], Lina Fan[3], Xinzhi Li[1, 2], Huizhong Jiang[1, 2] and Nan Lu[3]

[1]School of Geomatics and Urban Spatial Information, Beijing University of Civil Engineering and Architecture, Beijing, 106216, China
[2]Key Laboratory of urban spatial information, Natural Resources Ministry, Beijing, 106216, China
[3]Information Center of Ministry of Ecology and Environment, Beijing,100029, China.

Author Profile: E-mail: 2108570020056@stu.bucea.edu.cn
* Corresponding author: E-mail: zhangjianqin@bucea.edu.cn

Abstract. The retirement and relocation of urban industrial enterprises has led to the retirement of a large number of contaminated sites. Aiming at the problem that the data related to the contaminated site comes from many different sources and has different structures, and it is difficult to explore the potential correlation between the data through the existing management methods, this paper proposes a knowledge map construction method for multi-source heterogeneous data of the contaminated site. According to the different structures of contaminated site data, we use the knowledge construction theory to select appropriate entity recognition, relationship recognition and knowledge fusion methods to extract various types of information of contaminated sites and establish semantic networks. The knowledge map constructed for a contaminated site in Northeast China contains 3840 contaminated site entities including site information, enterprise information and soil information, and the corresponding association relationship is 4768. Practice has proved that the proposed knowledge map construction method can effectively and intuitively represent the potential association relationship between contaminated site data, and provide corresponding technical support and decision-making information for the relevant departments of contaminated site restoration and management.

Keywords: soil pollution, multi-source heterogeneity, knowledge atlas, visualization

1 Introduction

With the continuous progress of urbanization, since the 1990s, the phenomenon of closing or relocating urban central polluting industrial enterprises has first appeared on a large scale in China's more developed cities, and has increasingly spread to small and medium-sized cities. However, most of the polluting enterprises in urban centers were built very early, and some of them did not even have environmental protection facilities

V. Escudero et al. (Eds.): BADS 2022, AHCS 11, pp. 78–85, 2023.
https://doi.org/10.2991/978-94-6463-102-9_11

at that time. Therefore, after many years of production activities, most of the sites left after the closure or relocation of enterprises are seriously polluted, and blind redevelopment and utilization of these sites will easily cause serious health risks to future residents and surrounding residents [1]. Therefore, it is necessary to strengthen the environmental risk management of the relocation of sites left by industrial enterprises. Among them, the polluted site data gradually presents the characteristics of multi-scale, multi-source, multi-dimensional, multi type, and large amount of information. It is urgent to realize the deep mining and efficient management of site pollution information through digitization and informatization.

Knowledge graph [2] is a series of different graphs, showing the development process and structural relationship of knowledge. Visualization technology is used to describe knowledge resources and their carriers, mining, analyzing, constructing, drawing and displaying knowledge and their relationships. The traditional data service knowledge base has defects in semantic association, and cannot semantically associate debris data such as contaminated sites, soil pollution and production activities. Therefore, it cannot more effectively manage contaminated sites, and therefore cannot more effectively manage contaminated sites.

At present, relevant studies on contaminated sites in China mainly focus on the analysis of the current situation, migration simulation, pollution characteristics and remediation countermeasures, such as Li Dawei [3], Ge Feng [4], and Hou Deyi [5], all of whom conducted macroscopic analysis of the current situation and outlook of contaminated sites; Wu Junjie[6] et al. analyzed the contamination of decommissioned sites by analyzing the proposed decommissioning analyzed the contamination of the decommissioned site by analyzing the physical and chemical property monitoring items, scatter distribution pattern of arsenic and heavy metal pollutant concentrations of the soil; In contrast, there is a single research direction in the combination of knowledge mapping and contaminated sites, and only knowledge mapping has been used to analyze the published papers related to contaminated sites, such as Yan Kang et al [7] used knowledge mapping tools to analyze the main countries (regions) and institutions, main publishing journals, main research scholars, important literature and research hotspots and their changing trends in the Web of Science core collection database, etc. Therefore, the purpose of this paper is to construct a structured knowledge base of contaminated sites by applying knowledge graphs to contaminated sites, to store complex relationships among entities by connecting concepts, entities, attributes and interrelationships in contaminated site data through a mesh knowledge structure, to fuse discrete data in contaminated sites, and to support the intelligent search and analysis of contaminated site data. intelligent search and deep mining of contaminated site data.

2 Contaminated site knowledge graph construction

2.1 Knowledge graph construction process

Through the association between knowledge maps, data from multiple sources and structures can be integrated to reveal the potential relationships and laws between data in a multi form and multi-level manner. In view of the fact that the contaminated site data contains structured knowledge and a large number of open linked data, this paper decided to use the combination of top-down methods to build the knowledge map of the contaminated site. The entire construction process is shown in Figure 1. First, divide the data. Next, build an ontology to define the hierarchy, relationships and attributes of classes. Then, according to the different data structures, different methods are selected for knowledge extraction, and the acquired entities are summarized as ontology constructed by ontology. Then, the same entities or concepts from multiple sources are fused through knowledge fusion, and duplicate and redundant knowledge is eliminated to form a comprehensive knowledge map of the contaminated site, which is stored in the form of Neo4j graphic database.

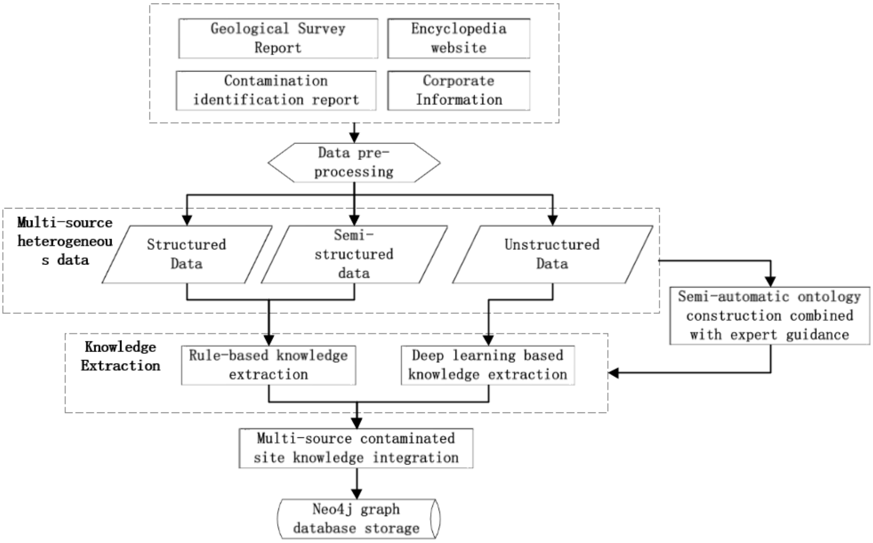
Geological Survey Report
Encyclopedia website
Contamination identification report
Corporate Information
Data pre-processing
Multi-source heterogeneous data
Structured Data
Semi-structured data
Unstructured Data
Semi-automatic ontology construction combined with expert guidance
Knowledge Extraction
Rule-based knowledge extraction
Deep learning based knowledge extraction
Multi-source contaminated site knowledge integration
Neo4j graph database storage

Fig. 1. Flow chart of knowledge map construction of contaminated sites

2.2 Key technologies for knowledge mapping of contaminated sites

2.2.1 Entity identification of contaminated sites based on BERT-BiLSTM-CRF model. The underlying data of contaminated sites mainly includes structured data, semi-structured data and unstructured data, and different methods are applied to identify the contaminated site entities for the variability of different structured contaminated site information in this paper. Among them, Unstructured data is the main manifestation of contaminated site information, mainly plain text data, which can be annotated with unknown text according to the previously acquired knowledge units, and the sequence annotation algorithm [9], which combines long and short-term memory network (BiLSTM) and conditional random field model (CRF) based on BERT pre-training model, is selected to identify contaminated site entities, and its model framework is shown in Figure 2 shows.

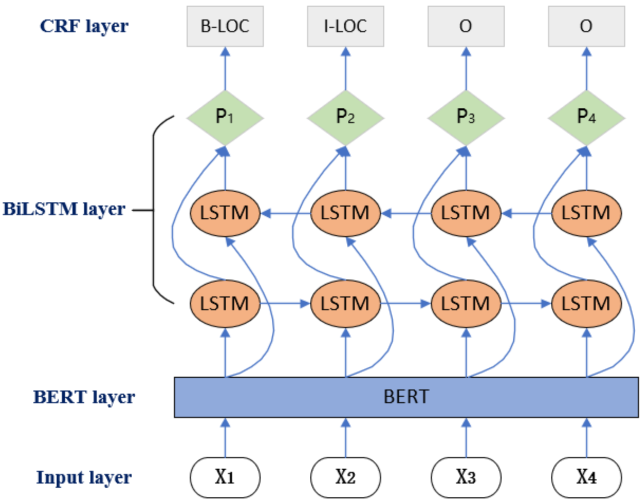

Fig. 2. BERT-BiLSTM-CRF model framework

2.2.2 Entity relationship extraction based on pattern matching. Entity relationship extraction mainly faces unstructured contaminated site text information to obtain the association relationships among contaminated site entities. Because the entity relationships in the contaminated site domain are relatively fixed, this paper adopts a pattern matching-based [10] approach for the extraction of contaminated site entity relationships.

The pattern matching approach is to construct and store a lexical or semantic pattern set based on existing relational examples, guided by domain prior knowledge, and then match the elements of natural language processed statements with the pattern set to determine entity relationships. For example, from "ammonia plant mainly produces ammonia and carbon dioxide", we extract "ammonia plant", "ammonia" and "carbon dioxide ", based on the constructed pattern (site name) production (product), and identify the entity relationship between the three as production. While the accuracy of the entities represented by the identified referent objects needs to be verified according to the textual context to improve the accuracy of relationship extraction.

2.2.3 Contaminated Site Knowledge Graph Knowledge Fusion. (1) Entity alignment. Using multiple knowledge extraction techniques, relatively isolated knowledge structure units are extracted from data sets with different sources and structures, and many sub-maps with low relevance are obtained, which need to be fused to obtain a complete knowledge map of contaminated sites. In the process of entity extraction, we are faced with the characteristics of wide data sources and cumbersome structure, and the entity alignment of contaminated sites also faces some problems, such as: (1) the names of pollutants are not uniform. (2) Some entities have abbreviations. (3) The coordinates of measurement borehole points are not uniform, and the national and local coordinate systems exist simultaneously in the data from different sources of the same contaminated site. Entity alignment [11] has a natural advantage in solving these problems. By identifying and filtering different labeled attributes, we find the real-world unique corresponding entities, and at the same time fuse the obtained entity sets into unique entities to create a globally unique identity, and then integrate the entity objects into the knowledge graph.

(2) Entity associations

After entity alignment, a large number of single-category knowledge maps are obtained, and to perform in-depth knowledge mining, entity association is required [12], and the extracted entities all have certain attribute values, such as the attributes of contaminated sites include enterprise number, unified social credit code, industry, address and land area, etc. The complete entity attribute structure is shown in Figure 3.

In order to establish a more complete knowledge map, based on the constructed ontology model, entities with the same attribute values are linked by their states or attribute values, thus forming multiple triadic data sets, and clustering multiple entities with the same association relationship at the same time, which can form a knowledge map of contaminated sites with multidimensional data relationships. Figure 3 shows the association structure of some entities of the contaminated site.

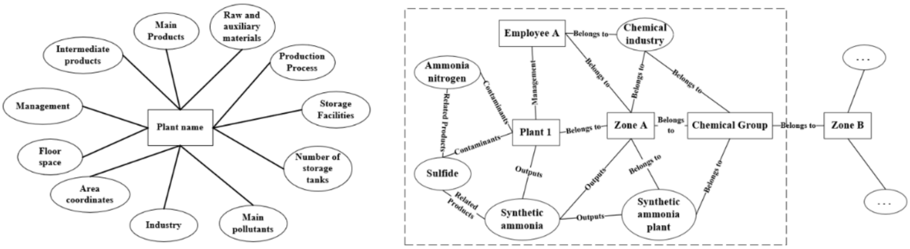

Fig. 3. Related structure diagram of some entities in the polluted site

3 Example of knowledge graph construction and analysis

3.1 Knowledge graph construction.

For the construction method of multi-source heterogeneous contaminated site knowledge graph proposed in this paper, this paper presents and analyzes examples based on the contaminated site data of a chemical group in Northeast China after relocation. A total of 3840 contaminated site entities such as site information, enterprise information and soil information are extracted, and the corresponding association relations are 4768. The overall knowledge mapping display is shown in Figure 4. As this site covers a large area, the stratigraphic distribution of the whole site, the current situation of the site, the original production process, etc., there are large differences, in order to more effectively analyze and mine the site data, according to the underlying distribution factors, production area distribution factors and late disturbance factors, this site is divided into Zone A, Zone B and Zone C. According to the situation of enterprises in each region, production activities, pollution status, soil conditions associated with this group of basic information, to facilitate the analysis of a certain pollutant pollution causes, to achieve the prevention and treatment of subsequent chemical plants.

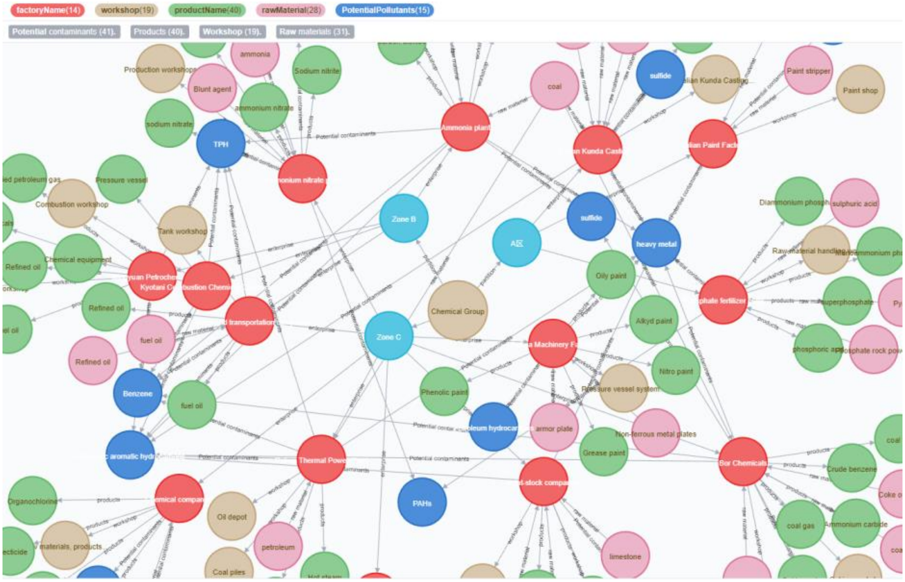

Fig. 4. Knowledge map display of contaminated site of a chemical group

4 Conclusion

In terms of deep knowledge mining and potential relationship ranking, knowledge atlas provides a new research direction for knowledge management and information intelligence of contaminated sites. Based on the characteristics of multi-source heterogeneous data of contaminated sites, the construction process of knowledge map of contaminated sites is designed, and the key technologies of entity recognition, relationship extraction and knowledge fusion of multi-source heterogeneous contaminated site data of contaminated sites are proposed. The construction of knowledge map of contaminated sites is realized for the first time, which provides a technical basis for intelligent management of contaminated site data and analysis of pollution causes of contaminated sites. Taking the polluted site left after the relocation of a chemical industry group in Northeast China as an example, the feasibility of the proposed method for constructing the knowledge map of multi-source heterogeneous contaminated sites is verified, and the semantic association between the data of contaminated sites is visualized to associate the data from different living sources. However, there are still many problems in the process of knowledge extraction in this paper. In the next work, we need to further improve the knowledge extraction method, obtain more pollution site data sources, apply the pollution site knowledge map to the research of automatic question answering and knowledge reasoning, meet the site manager's service requirements for intelligent, accurate and efficient multi-source heterogeneous pollution site data, which can meet the site manager's requirements for intelligent, accurate and efficient services for multi-source heterogeneous pollution site data, And maximize the value contained in contaminated site data..

Fund Projects

National Key R&D Program Projects(2019YFC1804903), Beijing Natural Science Foundation (8202013), National Natural Science Foundation of China (41771413).

References

1. Zhou J F, Investigation and treatment analysis of soil pollution in chemical legacy sites [J]. Resource Conservation and Environmental Protection, 2022, (02): 87-90.
2. Zhou J F, Investigation and treatment analysis of soil pollution in chemical legacy sites [J]. Resource Conservation and Environmental Protection, 2022, (02): 87-90.
3. Zhou H J, Shen T T et al., Survey of Knowledge Graph Approaches and Applications[J]. Journal on Artificial Intelligence,2020,2(2).
4. Li D W, Dong Y et al., Investigation and remediation Countermeasures of soil environment in polluted sites [J] Comprehensive Utilization of Resources in China, 2021,39 (12): 136-138.
5. Ge F, Zhang Z X et al., Analysis and Prospect of organic pollution sites in China [J] Soil, 2021,53 (06): 1132-1141.
6. Hou D Y, Zhang K K et al., Present situation and Prospect of heavy metal contaminated soil treatment in industrial sites [J] Environmental Protection, 2021,49 (20): 9-15.
7. Wu J J, Jiang N et al., Investigation on pollution characteristics of decommissioned sites of natural gas to methanol enterprises [J] Chemical Minerals and Processing, 2022,51 (03): 45-48
8. Yan K, Lou J et al., Research status and development trend of contaminated sites: Analysis Based on knowledge atlas [J] Journal of Soil, 2021,58 (05): 1234-1245.
9. Zhou Y, Liu Z et al., Construction of multi-level ontology data fusion model [J] Library and Information Work: 1-8.
10. Guillaume Lample, Miguel Ballesteros et al., Neural Architectures for Named Entity Recognition [J]. CoRR,2016, abs/1603.01360.
11. Yang X H, Zhang S W et al., Research on Chinese relation extraction technology [J] Journal of Nanhua University (NATURAL SCIENCE EDITION), 2018,32 (01): 66-72.
12. Wang Z Q, Wang Y et al., Entity Alignment Method for Power Data Knowledge Graph of Semantic and Structural Information[J]. IOP Conference Series: Materials Science and Engineering,2019,569(5).
13. Li J Y, Yue K, Discovery of related entities in knowledge atlas [J] Journal of Yunnan University (NATURAL SCIENCE EDITION), 2021,43 (06): 1079-1085.
14. Green Alastair, Guagliardo Paolo et al., Updating graph databases with Cypher[J]. Proceedings of the VLDB Endowment,2019,12(12).