Construction of a Knowledge Graph for China's Nuclear Safety Laws and Regulations

Chengjie Yan, Rui Shi, and Zhengchuan Wang

College of Mechanical Engineering, Sichuan University of Science & Engineering, Zigong 643000, China

Abstract

A method is proposed to organize the knowledge of nuclear safety laws and regulations so that the relevant knowledge can be organized in a more orderly manner. As the data source of nuclear safety laws and regulations, we select the syllabus of the examination for registered nuclear safety engineers, the national laws and regulations database, and the laws and regulations in China's general database of legal knowledge resources, screen and obtain the laws and regulations related to nuclear safety, extract the corresponding formulating authority, timeliness, type of regulations, date of publication, date of implementation, and year of information, and preprocess the obtained data; construct ontologies, define entities and relationships; extract entities and relationships from the obtained text of laws and regulations based on the rule-based method, and use Python to organize the relevant knowledge in a more orderly way. The ontology is constructed, entities and relations are defined, the entities and relations are extracted from the obtained laws and regulations by rule-based method, and the data are converted into entity-relationship triples by using py2neo in Python, and stored in Neo4j graph database to complete the storage of the knowledge map of the laws and regulations on nuclear safety. A total of 41 laws and regulations related to animal husbandry were collected, and a knowledge map of animal husbandry laws and regulations with 6752 entities and 9621 relationships was constructed. This method can be applied to the organization of the knowledge of nuclear safety laws and regulations, which can make the organization of knowledge more orderly and more closely related.

Keywords

Nuclear Safety; Knowledge Mapping; Laws and Regulations; Neo4j.

1. Introduction

With the wide application of nuclear energy technology and the continuous development of the nuclear field, the issue of nuclear safety[1] has attracted more and more attention from the international community. The construction of a comprehensive and accurate knowledge[2] map of nuclear safety laws and regulations is of great theoretical and practical significance for promoting the management of regulations in the field of nuclear safety and enhancing the level of legal intelligence. This dissertation aims to construct a comprehensive and organic knowledge map of nuclear safety laws and regulations by integrating regulatory data from multiple sources such as the syllabus of the Certified Nuclear Safety Engineer Examination, the National Database of Laws and Regulations, and the General Database of China's Legal Knowledge Resources, and by utilizing advanced knowledge map construction techniques. In this process, by extracting entities and modeling relationships from the text of regulations, the data are efficiently stored in the Neo4j[3] graph database through the py2neo[4] library in Python, forming a huge network structure containing multiple entities and relationships. Through this knowledge graph, we are able to understand the regulatory system in the field of nuclear safety in a more in-depth way, and promote better application and intelligent management of regulatory knowledge. This study not only provides new ideas for the systematic collation and organization of nuclear safety laws and regulations, but also provides practical experience for the in-depth research in the field of regulatory knowledge graph construction.

However, the application in the field of nuclear safety has not been reported yet. Combining the problems of nuclear safety knowledge popularization and the current status of domestic and international application of knowledge graph in the vertical field, it is proposed to construct a knowledge graph in the field of nuclear safety by taking the nuclear safety laws and regulations as the research object, and at the same time, build a Q&A[5] platform for nuclear safety knowledge. In order to verify the feasibility of knowledge mapping in the field of nuclear safety and promote the popularization of nuclear safety knowledge.

2. Overview of Research Related to Knowledge Graphs

In 2012, Google introduced the concept of Knowledge Graph (KG) with the aim of enhancing the search results of its search engine by enabling conceptual retrieval through reasoning and graphically presenting categorized structured knowledge to users. Knowledge graph is a method of describing and modeling knowledge with a graph model, which is essentially a semanticized network that effectively integrates fragmented knowledge by means of triples (usually composed of entities, attributes, and relationships) to clearly reveal the relationship between things and their relationships. The more well-known ones in foreign countries are DBpedia [6], YAGO [7], and Freebase [8].DBpedia, started in 2007, is a large-scale multi-language encyclopedic knowledge graph containing 28 million entities, and it can be regarded as a database version of the multi-language Wikipedia.YAGO is an integration of the data of Wikipedia, WordNet, and GeoNames data integration, containing more than 1 million entities.Freebase has more than 68 million entities and 1 billion relationships. The larger domestic knowledge graphs include XLORE [9] and Zhishi.me [10]. XLORE is a large-scale Chinese and English knowledge graph constructed by Tsinghua University, containing nearly 149 million entities. Zhishi.me is constructed by Shanghai Jiao Tong University, and owns nearly 125 million triples. With the rapid development of the Internet, the information resources in the network are increasing, and the scale of the generalized knowledge graph is also expanding.

3. The Application Value of Traditional Embroidery in Modern Fashion Design

The knowledge graph of nuclear safety laws and regulations is logically divided into a schema layer and a data layer. The schema layer is regarded as the core of the knowledge graph and is usually built on top of the data layer. The schema layer of nuclear safety knowledge mapping can clearly define the entities, attributes, and relationships between entities in the field of nuclear safety in order to clarify the hierarchical relationships between different concepts, and sort out and subdivide the concepts and relationships of nuclear safety laws and regulations to form a conceptual hierarchical relationship map with a clear structure. And the data layer of nuclear safety laws and regulations mapping is the instance data filling under the constraints and limitations of the schema layer. The data layer contains more information such as entities, relationships and attributes.

In this paper, a combination of top-down and bottom-up approach is used to construct the nuclear safety knowledge graph. In this process, the schema layer clarifies the goal of nuclear safety based on the understanding of nuclear safety knowledge, combs the relevant concepts of various nuclear safety knowledge ontologies top-down according to the content of the relevant nuclear safety professional dictionary, identifies the sources of the concepts as well as the boundaries of the nuclear safety domains, and constructs the corresponding comprehensive ontology library of the nuclear safety domains; and the data layer is guided by the framework of the schema layer, and adopts different processing methods for different data types, extracting the corresponding entities, relationships and attributes. different processing methods for different data types, extract the corresponding entities and relationships, and perform entity fusion. The fused nuclear safety triad will eventually be stored in the graph database, thus realizing the bottom-up construction of the data layer.

3.1 Data Sources and Access

Constructing a knowledge graph requires the support of the underlying primary data, and the accuracy and coverage of the primary data directly determine the quality of the graph. When acquiring nuclear safety knowledge, it is necessary to ensure that the data have a high degree of accuracy, authority and usability. With the development of nuclear energy in China, nuclear safety data are characterized by large scale, dispersion, high specialization and low value density. Since there is no large-scale nuclear safety knowledge base at this stage, this paper selects the objects of national laws and regulations database, and extracts the key information to construct nuclear safety knowledge mapping as follows:

Since laws and regulations usually adopt formal and accurate language and cannot be changed at will, the laws and regulations listed in the syllabus of the Nuclear Safety Engineer Examination will be used as references, which mainly include the legal texts, regulatory documents and relevant policy documents at the national level, covering the management of nuclear energy, radiation protection, and response to nuclear accidents, etc. The laws and regulations listed in the syllabus of the Nuclear Safety Engineer Examination will be used as references. Considering the timeliness of regulatory documents, some of which may have been revised, modified or repealed, the acquired laws and regulations are listed one by one in the National Laws and Regulations Database (https://flk.npc.gov.cn/), Beida Law Treasure (https://home.pkulaw.com/), and Beida Legal Intent (http://www.lawyee. org/) and other websites to ensure the accuracy of the data on laws and regulations.

For the nuclear safety related laws and regulations data, the OCR recognition tool Adobe Acrobat OCR was used to recognize the acquired nuclear safety laws and regulations. A total of 41 laws and regulations were obtained. The nuclear safety laws and regulations include the Nuclear Safety Development of the People's Republic of China and the Radioactive Pollution Prevention and Control Law of the People's Republic of China, etc.

3.2 Data Preprocessing

Most of the data related to nuclear safety laws and regulations are obtained after data acquisition, and the corpus data have different structures and more repetitive data, which need to be pre-processed by segmentation and de-duplication before they can be used to construct the atlas.

Segmentation refers to cutting a continuous text into individual words with independent semantics. De-duplication refers to the removal of duplicate items in the text and the removal of common deactivated words. Currently, the commonly used lexical tools include THULAC, Stanford Lexer, LTP, and Jieba, etc. THULAC is a lexical tool developed by Tsinghua University, which adopts a hybrid lexical method combining rule-based and statistical-based techniques. Chinese natural language processing toolkit introduced by Harbin Institute of Technology, which contains many functions such as participle, lexical annotation and named entity recognition etc. Jieba is the most used Chinese participle tool in China at present, which adopts the participle algorithm based on the prefix lexicon structure, which greatly improves the speed of participle, supports three participle modes: precise mode, full mode and search engine mode, supports customized lexicon, and it is easy to use. It supports customized dictionaries and is easy to use. After comparison, this paper chooses Jieba tool to carry out preprocessing work on nuclear security data.

The specific steps of preprocessing are as follows:

(1) Import the acquired nuclear safety laws and regulations data into Jieba tool to realize the word splitting of nuclear safety text data.

(2) Remove the deactivated words, combining the deactivated word list of HIT and Jieba tool to remove the repeated statements, special characters and punctuation marks in the text.

(3) Output the preprocessing results.

A knowledge graph corpus is a collection of language samples used to construct and enrich a knowledge graph. Constructing a knowledge corpus in the field of nuclear safety can provide data support for entity and relationship extraction. Due to the lack of public knowledge base in the field of nuclear safety, it is necessary to annotate the text with data, in order to ensure the accuracy and credibility of the annotated data, this paper uses manual annotation, and adopts the BIO sequence annotation method to annotate some of the data in the nuclear safety information. Each word is labeled in the form of "B-X", "I-X" or "O", with "B" representing the current B" means that the current tagged element is the starting position of an entity, "I" means that the current tagged element is the non-starting position of an entity, and "O" means that the current tagged element is not an entity. Marking is done in one sentence, so before the marking work, Python program script is written to process the text in sentences, and then the unstructured data is analyzed, combined with the expert's advice, and the entity extraction type is defined, and some of the entity types is shown in Tables 1.

Table 1. Entity List of Nuclear Safety Laws and Regulations

Name	conceptual	resolve
Nuclear_safety_law	law	law
law_category	category	category
year	year	year
release_date	pub_year	pub_year
publish_department	publishment	publishment
timeliness	timeliness	timeliness
implementation_date	date	date
law_chapter	chapter	chapter
chapter_articles	article	article
articles_term	item	item

3.3 Knowledge Extraction

For nuclear safety laws and regulations text, it is very important to ensure the accuracy and legality of the data, the legal text needs to be rigorous and accurate, the abstract generalization of the legal text may lead to inaccurate, ambiguous or distorted information, which in turn affects the reliability of the knowledge graph, and the number of laws and regulations related to nuclear safety is relatively small, in order to ensure the accuracy of the data, the design of the design combines the manual and rule matching for the 41 acquired In order to ensure the accuracy of the data, the design combines manual and rule matching to extract entities and relationships from the 41 acquired laws and regulations. After constructing the ontology of nuclear safety laws and regulations, the entity and relationship types to be extracted can be obtained, among which, the entity types include "laws and regulations", "timeliness", "issuing department", "chapter", "department" and "department". Among them, entity types include "laws and regulations", "timeliness", "issuing department", "chapter" and "article", etc., and relationship types include "nuclear safety laws and regulations - category relationship of laws and regulations", Relationship types include "nuclear safety laws and regulations - the relationship between the date of publication of laws and regulations" and "nuclear safety laws and regulations - the relationship between the chapter of laws and regulations" and other ten types. Based on the defined entity and relationship types, the 41 laws and regulations are extracted manually and the extracted data are processed into a ternary format and saved in an Excel file.

3.4 Knowledge Storage

After extracting the entity relationships, the extracted knowledge should be written into the database to be persisted, so as to pave the way for further knowledge query. There are complex one-to-many and many-to-many relationships among the entities of nuclear safety related laws and regulations. However, general relational databases usually use a fixed form of storage to the table, this storage and query method is not only costly and inefficient, but also extremely inconvenient for the dynamic management of subsequent additions, deletions and modifications. Compared with relational databases, non-relational graph databases are more suitable for storing knowledge graphs.Neo4j is a high-performance NoSQL graph database developed in java language, which has become one of the most popular graph databases due to its advantages of high performance, lightweight, and multi-operating system support, etc. Therefore, Neo4j is used as the knowledge graph database for storing the knowledge graphs of nuclear security laws and regulations in this paper. Therefore, in this paper, Neo4j graph database is used as the database for storing nuclear safety laws and regulations knowledge graph.

Since the cleaned data are stored in MySQL database during the data processing stage, the data in MySQL should be converted into triples and stored in the Neo4j graph database, which is done by using python language, SQLAlchemy library in python library to manipulate MySQL database, and py2neo library in python library to manipulate MySQL database. The python library SQLAlchemy is used to manipulate the MySQL database, and the python library py2neo is used to manipulate the Neo4j database. Using python language to loop through the CQL CREATE statement to insert the data into the Neo4j database, and finally constructed the knowledge graph of animal husbandry laws and regulations, which contains a total of 6,752 entities and 9,621 relationship.

4. Conclusion

As an emerging knowledge organization technology, knowledge mapping has demonstrated its advantages in many industries, but its application in the field of nuclear safety is still immature, and there are problems such as strong specialization and fragmented domain data, etc. The paper proposes a method to construct a knowledge map of nuclear safety laws and regulations, using several authoritative law and regulation information websites as data sources. The paper proposes a method to construct a knowledge graph of nuclear safety laws and regulations, taking several authoritative laws and regulations information websites as data sources to collect nuclear safety related laws and regulations, and constructing fragmented knowledge of nuclear safety laws and regulations into a knowledge graph after data preprocessing, ontology construction, entity relationship extraction and knowledge storage, laying a foundation for establishing a more comprehensive and rich knowledge graph of animal husbandry. In the future research, we can extract the entities and relationships of the legal provisions in the nuclear safety laws and regulations in order to construct a knowledge graph with finer granularity, so as to provide a knowledge base to support the intelligent retrieval and intelligent Q&A system related to nuclear safety in the future.

References

[1] Petrangeli G. Nuclear safety[M]. Elsevier, 2006.

[2] Wang X, Chen L, Ban T, et al. Knowledge graph quality control: A survey[J]. Fundamental Research, 2021, 1(5): 607-626.

[3] Miller J J. Graph database applications and concepts with Neo4j[C]//Proceedings of the southern association for information systems conference, Atlanta, GA, USA. 2013, 2324(36): 141-147.

[4] Karangutkar S. Implementation of clustering algorithm using graph embeddings and graph data science on Yelp restaurant dataset[D]. Dublin Business School, 2020.

[5] Shah C, Oh S, Oh J S. Research agenda for social Q&A[J]. Library & Information Science Research, 2009, 31(4): 205-209.

[6] Bizer C, Lehmann J, Kobilarov G, et al. DBpedia - A Crystallization Point for the Web of Data. Web Semantics: Science, Services and Agents on the World Wide Web, 2009, 7(3): 154-165. DOI: 10.1016/j.websem.2009.07.002.

[7] Suchanek F M, Kasneci G, Weikum G. Yago: A Core of Semantic Knowledge[C]//Proceedings of the 16th International Conference on World Wide Web. ACM, 2007: 697-706. DOI: 10.1145/1242572.1242667.

[8] Bollacker K, Evans C, Paritosh P, et al. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge[C]//Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. ACM, 2008: 1247-1250. DOI: 10.1145/1376616.1376746.

[9] Wang Z, Li J, Wang Z, et al. Xlore: A Large-scale English-Chinese Bilingual Knowledge Graph[C]// Proceedings of the 2013 International Conference on Posters & Demonstrations Track-Volume 1035. 2013: 121-124.

[10] Niu X, Sun X, Wang H, et al. Zhishi.me - Weaving Chinese Linking Open Data[C]//International Semantic Web Conference. Springer Berlin Heidelberg, 2011, 7032: 205-220. DOI: 10.1007/978-3-642-25093-4-14.