ISSN: 2688-8653

# Construction of Chinese Learning Dialogue System based on Knowledge Graph

Yu Yao[1], Yufeng Liu[2], Junhu Li[1], Longling Zhang[1], Zhuoqi Wei[1]

[1]School of Date science and technology, Heilongjiang University, Harbin, China, 150080, China

[2]School of Entrepreneurship Education, Heilongjiang University, Harbin, China, 150080, China

## Abstract

**As an ancient civilization in China, Chinese culture has a long history. Since the "Belt and Road" initiative was announced, China and the countries along the "Belt and Road" have been in closer contact. At present, the respect of the Chinese language to the world has achieved obvious results. The study of the Chinese language is more important at home and abroad. In the context of big data, knowledge graphs are gaining more and more attention. It has been widely used in the fields of bioinformatics, finance, and medical treatment, but it has hardly been applied in the field of Chinese learning. This article first introduces the basic concepts and main applications of the knowledge graph and then explains the necessity of studying the knowledge graph of Chinese learning. Taking the knowledge graph of Chinese learning as an example, it introduces the tools and applications of the knowledge graph. Finally, the application of the knowledge graph in Chinese learning is summarized and prospected.**

## Keywords

**knowledge map; teaching Chinese as a foreign language; knowledge quiz; Chinese learning.**

## 1. Introduction

### 1.1. Project Introduction

In recent years, with the rise of the "China fever" worldwide, the "Chinese fever" has also continued to heat up, and the cause of international promotion of Chinese language has flourished. The "Chinese language fever" reflects the great achievements China has made in its reform and opening up in the past 20 years Urgent desire. The intelligent question answering system for Chinese learning based on knowledge graph can help Chinese learners to learn progress well. The purpose of this project is to enable Chinese language learners to lay a solid foundation in Chinese language learning, understand Chinese culture, accelerate the pace of Chinese language learning, and more quickly solve doubts in Chinese language learning.

### 1.2. Introduction of Knowledge Graph

Knowledge graph technology refers to the technology of establishing and applying knowledge graph. It is a cross-research of fusion of cognitive computing, knowledge representation and reasoning, information retrieval and extraction, natural language processing and semantic Web, data mining and machine learning.

The knowledge graph describes the concepts, entities and their relationships in the objective world in a structured form, expresses the information of the Internet into a form closer to the human cognitive world, and provides a better ability to organize, manage and understand the massive information of the Internet. Specifically, the knowledge graph is to combine the theories and methods of applied mathematics, graphics, information visualization technology, information science and other disciplines with metrology citation analysis, co-occurrence analysis and other methods, and use the visual graph to visually display the disciplines. The modern theory of core structure, development history, frontier fields and overall knowledge structure to achieve the purpose of multidisciplinary integration.

### 1.3. Application of Knowledge Graph

It displays complex knowledge fields through data mining, information processing, knowledge measurement and graph drawing, reveals the dynamic development laws of the knowledge field, and provides a practical and valuable reference for subject research. Knowledge graph, together with big data and deep learning, has become one of the core driving forces driving the development of the Internet and artificial intelligence. The knowledge graph has highlighted the increasingly important application value in the following applications: knowledge fusion, semantic search and recommendation, question answering and dialogue system, big data analysis and decision-making.

### 1.4. Summary

This article mainly implements the application of the question answering and dialogue system in the knowledge graph. The performance of the knowledge question answering system usually depends on the size and quality of the knowledge base. Extensive, diverse and complex structures, how to extract effective information from it, fuse fragmented knowledge, and form a structured knowledge base is the research focus and difficulty in knowledge engineering. In fact, domestic and foreign experts have been devoted to the research of high-quality, large-scale knowledge base construction. Research on the construction of Chinese learning knowledge graph, using Chinese knowledge database as the data source of knowledge graph. The following will introduce the data description and preprocessing, entity recognition, relationship extraction, the use of neo4j to generate visual knowledge maps and the construction of question and answer systems.

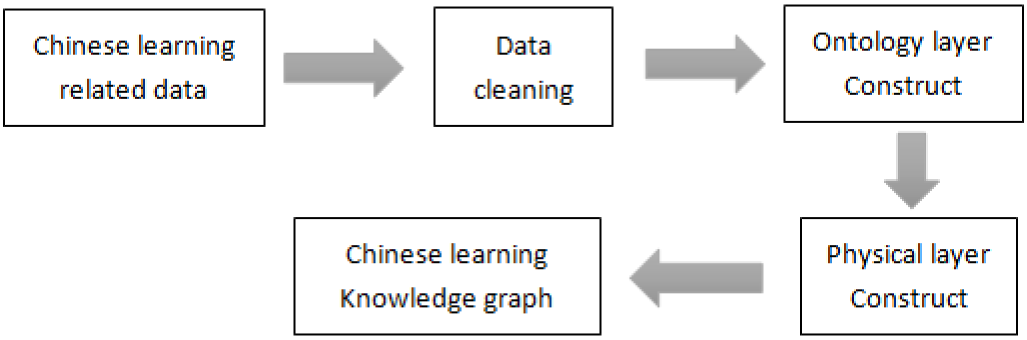
Chinese learning related data
Data cleaning
Ontology layer Construct
Physical layer Construct
Chinese learning Knowledge graph

**Figure 1.** Construction process of knowledge graph

## 2. Data Description and Preprocessing

### 2.1. Data Crawling

The data used in this article is mainly from github open source data sets, interactive encyclopedia and Baidu encyclopedia, including radicals, pinyin, Chinese educational institutions, Chinese textbooks, Chinese character culture, Chinese culture, writers, poets, Chinese history, dynasties, animals Thesaurus, medical thesaurus and other aspects of Chinese learning related data. The data set in this article is mainly crawled through the Scrapy framework. Scrapy is an application framework written for crawling website data and extracting structural data, combined with regular expressions and xpath methods to crawl and store Chinese-related data.

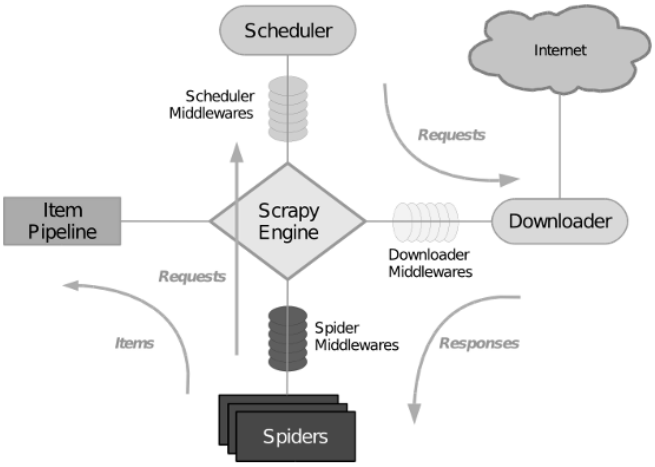

**Figure 2.** Scrapy overall architecture diagram

Finally, the name of the entity and the relevant introduction of Baidu Encyclopedia and Interactive Encyclopedia are obtained, as shown in Table 1.

**Table 1.** Data obtained by crawling

| Entity name | Related introduction |
|---|---|
| 李白 | 李白（701年—762年），字太白，号青莲居士，又号“谪仙人”，唐代伟大的浪漫主义诗人，被后人誉为“诗仙”，与杜甫并称为“李杜”，为了与另两位诗人李商隐与杜牧即“小李杜”区别，杜甫与李白又合称“大李杜”。 |
| 扌 | 扌，同“手”。用作偏旁。俗称“提手旁”。 |
| 儒家 | 儒家是孔子所创立、孟子所发展、荀子所集其大成，之后延绵不断，为历代儒客推崇。 |
| …… | …… |

## 2.2. Data Preprocessing

Data preprocessing is an important step in the fields of data mining, machine learning, and the construction of knowledge graphs. He directly affects the quality of the research results of the entire project. In today's big data era, complex and multi-source heterogeneous data is becoming more and more massive, so data preprocessing has become an indispensable and important process in this field. In this paper, data preprocessing can provide entity extraction and relationship extraction in the future. great help.

Because the data used in this article is for the structure of the web page, the method of generating a wrapper is selected, and the data is extracted using the Scrapy crawler framework in conjunction with regular expressions, so the data cleaning part is no longer performed, the entity extraction part is directly performed, and the data preprocessing part It just processes the data into a format that facilitates entity extraction.

The data preprocessing in this paper mainly uses pyltp. The language technology platform (LTP) has been continuously developed and promoted by the Social Computing and Information Retrieval Research Center of Harbin Institute of Technology for 11 years. It is the most influential Chinese processing basic platform at home and abroad. The functions it provides include the removal of stop words, Chinese clauses, word segmentation, part-of-speech tagging, named entity recognition, dependency syntax analysis, and semantic role tagging. This article is mainly used to remove stop words, Chinese clauses, participles, part-of-speech tagging, and named entity recognition.

## 3. Entity Recognition

Entities are things in the objective world and are the basic units that constitute the graph of knowledge (here entities refer to individuals or instances). Entities are divided into limited categories of entities (such as commonly used names of people, places, etc.) and open categories of entities (such as names of diseases, etc.). Entity recognition is to identify entities of a specified category in the text. Entity linking is to identify the word or phrase (referred to as entity mentioning) mentioned in the text and link it with the corresponding entity in the knowledge base.

Entity recognition and linking is the core technology of knowledge graph construction, knowledge complementation and knowledge application. Entity recognition technology can detect new entities in the text and add them to the existing knowledge base. Entity linking technology can discover new knowledge about specific entities by discovering different occurrences of existing entities in the text. The study of entity recognition and linking will provide a knowledge base for computer-like human reasoning and natural language understanding.

In this paper, the entity used when crawling data is used as the entity of the knowledge graph, and the existing entities are classified according to the named entity recognition results and regular expressions mentioned above. Entities build relationships with each other and establish triples.

**Table 2.** Classification of entities

| Entity type | Chinese meaning | Examples |
|---|---|---|
| component | 偏旁部首 | 扌、氵、忄、艹、犭、宀 |
| poet | 诗人 | 李白、孟浩然、岑参、王安石、毛泽东 |
| celebrity | 诸子百家 | 儒家、道家、阴阳家、法家、名家、墨家 |
| tradition | 传统文化 | 农历、武术、二十八宿、对联、龙、五行学说、八卦 |
| materials | 汉语教材 | 《汉语教程（俄文版）》、《汉语拼音练习册》 |

## 4. Relationship Extraction

Entity relationship describes the relationship between objects that exist objectively, and is defined as a certain relationship between two or more entities. Entity relationship learning is to automatically detect and recognize a certain semantic relationship between entities from text. Also called relationship extraction. Entity relationship extraction classification pre-defined relationship extraction and open relationship extraction. Predefined relationship extraction means that the relationship extracted by the system is pre-defined; open relationship extraction does not predefine the extracted relationship category, and the system automatically finds and extracts the relationship from the text. Entity relationship recognition is the foundation of automatic construction of knowledge graph and natural language understanding.

This article uses regular expressions to classify the poets 'names into the entities that have been classified, such as" (\ w {1,5} 代) |. *? 称. *? (\ W {1,5} 代) ". The dynasty to establish relations. Establish the relationship between the two entities, that is, the triple, in turn for different two types of entities, including the poet-dynasty relationship, the Chinese character-radical relationship, the Chinese character-pinyin relationship, the literary work-author relationship, and the Chinese character- The relationship of meaning, the relationship of Chinese characters-the idiom composed of the Chinese characters, the relationship of Chinese characters-strokes, the relationship of Chinese characters-structures, the relationship of Chinese textbooks-textbook authors, etc., Table 3 is the relationship classification table.

**Table 3.** Classification of entity relationships

| Relationship type | Chinese meaning | Examples |
|---|---|---|
| component_of | 汉字的偏旁 | <打> 偏旁 <扌> |
| pinyin_of | 汉字的拼音 | <打> 拼音 <dǎ> |
| meaning_of | 汉字的含义 | <打> 含义 <击、敲> <做、造> <放出、发出 > |
| belong | 属于的朝代 | <李白> 是 <唐朝>的、<唐三彩>盛行于<唐朝> |
| Idioms | 包含某汉字的成语 | <屈打成招> 中包含 <打>字 |
| structure_of | 汉字的结构 | <打> 是<左右结构> |

## 5. Use Neo4j to Store Knowledge Graphs

Figure 3 shows several current mainstream storage methods for knowledge graphs. Knowledge graphs are graph-based data structures. There are two main storage methods: RDF storage format and graph database. Among them, RDF and APache Jena exist as triples, NoSQL is not conducive to storing large amounts of data, and Neo4j is an open source map database.

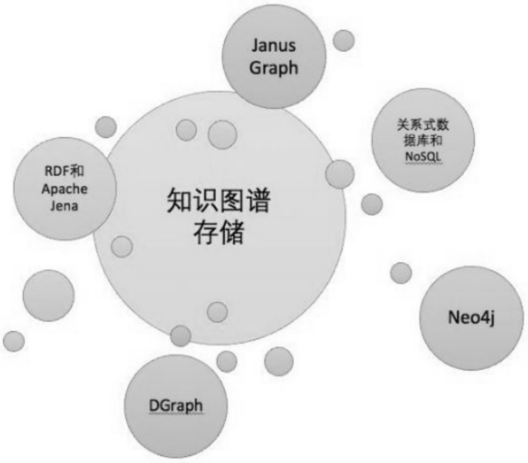

**Figure 3.** Storage of knowledge graph

In this paper, Neo4j is used to store and visualize the knowledge graph. By creating nodes based on the entities and the relationships between the entities that have been classified above, and establishing relationships between the two connected nodes, a huge and complicated complex is finally achieved. Knowledge graph. Through Neo4j's query language

Cypher, you can query a certain entity and its related entities and the sub-graphs of all entities with a certain relationship. The establishment of the graph database is very helpful for the establishment of the question-answer system later. Through semantic recognition, query To meet other entities that have a certain relationship with an entity, it can be used as an answer to the answer, thereby realizing the question and answer process.

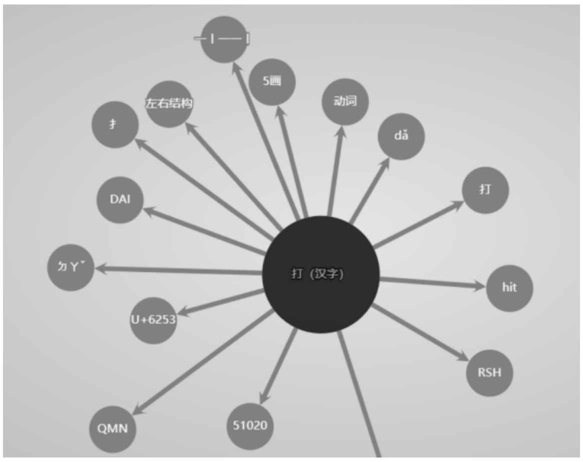

**Figure 4.** The result of querying the Chinese character "打"

## 6. Construction of Question Answering System

Question answering system refers to letting the computer automatically answer the questions raised by users, and is an advanced form of information service. The question answering system is regarded as one of the subversive technologies of information services in the future, and it is considered to be one of the main verification methods for the machine's ability to understand language. This part contains two parts: intention recognition and entity recognition module and data generation module.

The main function of the intent recognition and entity recognition module is to extract the core keywords of the questions asked by the user, build a keyword library, build a tree-like matching tree, accelerate the speed of intent recognition and entity recognition, and merge different types of entities through the matching tree , And finally carry out intent recognition integration to get the user's key intent. The extracted intentions are sent to the data generation module, and then entered into the graph database for query matching. This module is also equipped with functions such as sentence beautification, and finally the query results are processed, and finally fed back to the user to implement a question and answer system.

## 7. Summary

As China becomes stronger and stronger, China's influence is also increasing. Knowledge mining and information service sharing in Chinese learning have become an important research content in the era of big data. In the field of Chinese teaching, the application of knowledge graphs is also being tried. However, traditional search engines based on string matching cannot understand the deeper semantic information in user problems, and it is difficult to obtain accurate user requirements, nor can they meet the user's refined requirements. Therefore, designing and implementing a Chinese knowledge service system and constructing a Chinese learning knowledge map have certain significance for the development of Chinese teaching informatization.

In order to construct the knowledge graph of Chinese language learning, this paper develops data crawling, data preprocessing, data extraction, data storage, intelligent question answering, etc. In order to solve the knowledge graph expansion and improve the efficiency of knowledge question answering, tree-like intention recognition is adopted. , Further improve the performance of the algorithm, use knowledge question and answer in the system, build a Chinese knowledge service platform, realize the display of Chinese knowledge graph and knowledge question and answer.

## Acknowledgements

This project is funded by the School of Entrepreneurship Education of Heilongjiang University. The funding number is 201910212133.

## References

[1] Liu Z, Peng E, Yan S, et al. T-Know: a Knowledge Graph-based Question Answering and Information Retrieval System for Traditional Chinese Medicine[C]//Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations. 2018: 15-19.

[2] Pechsiri C, Piriyakul R. Explanation knowledge graph construction through causality extraction from texts[J]. Journal of computer science and technology, 2010, 25(5): 1055-1070.

[3] Yi Liu, Jiawen Peng, and Zhihao Yu. 2018. Big Data Platform Architecture under The Background of Financial Technology: In The Insurance Industry As An Example. In Proceedings of the 2018 International Conference on Big Data Engineering and Technology (BDET 2018). Association for Computing Machinery, New York, NY, USA, 31–35.

[4] Yanjun Z, Xiaodong Y, Yi L, et al. Research on the Construction of Wisdom Auditing Platform Based on Spatio-temporal Big Data [J][J]. Computer and Digital Engineering, 2019, 47(03): 616-619.

[5] Z. Zhao, J. Wang and Y. Liu, "User Electricity Behavior Analysis Based on K-Means Plus Clustering Algorithm," 2017 International Conference on Computer Technology, Electronics and Communication (ICCTEC), Dalian, China, 2017, pp. 484-487.

[6] Wang P, Jiang H, Xu J, et al. Knowledge graph construction and applications for Web search and beyond[J]. Data Intelligence, 2019, 1(4): 333-349.

[7] Chen Z, Bao J, Zheng X, et al. An Assembly Information Model Based on Knowledge Graph[J]. Journal of Shanghai Jiaotong University (Science), 1-11.