A TCM Question and Answer System Based on Medical Records Knowledge Graph

Yihong Xie
The College of Information, Mechanical and Electrical Engineer
Shanghai Normal University
Shanghai, China
e-mail: 1000465597@smail.shnu.edu.cn

***Abstract*—In recent years, the combination of knowledge map and question and answer system are more and more applied in professional fields. In this paper, we firstly constructed a knowledge graph of traditional Chinese medical records which contains four entities: herbs, prescriptions, diseases, symptoms and their relationships; Secondly, we use Node2vec algorithm to represent the graph; Finally, we match questions and answers based on KNN. Experiment result shows that with the output of five prescriptions, our system's accuracy reaches 93.4%, which achieves good results.**

***Keywords*—TCM, knowledge graph, Q&A, node2vec, KNN**

I. INTRODUCTION

Traditional Chinese Medicine (TCM) dates back more than 5,000 years ago, and plenty of studies show many benefits of the therapies[1]. A TCM case is a complete record of the process of dialectical treatment performed by TCM clinicians, which is the cradle of the continuous development of traditional Chinese medicine theory.

Nowadays, there are a large number of traditional Chinese medicine resources on the Internet, which can be further collected and processed. At the same time, we hope that the knowledge of TCM can better accessed by users. So a dedicated Q&A system is desperately needed in TCM field. In recent years, research shows the knowledge graph-based question answering system can answer users' questions more accurately and quickly. However, there is a room for improvement in the existing question answering system when it comes to the field of Chinese medicine:

First, the data crawled from the website may contain errors and null value. Secondly, the accuracy of the results can be further improved.

Therefore, in this paper we firstly constructed a knowledge graph of traditional Chinese medical based on four entities and three relationships: cure between prescription and disease, inclusion between prescription and herb, representation between disease and symptom; Secondly, we use Node2vec algorithm to represent the nodes in the graph; Thirdly, we match questions and answers based on word segmentation and KNN. Finally, with the recommended accuracy rate as indicators, we conduct experiments that shows high accuracy up to 93.4%.

The main contributions of this thesis are summarized as follows:

1. A question and answer system based on TCM medical case knowledge graph is proposed;

2. Node2vec algorithm is used to realize vector representation of knowledge graph;

3. Based on knowledge graph and KNN algorithm, the question and answer algorithm are implemented.

II. RELATED WORK

A. Knowledge Graph Research

The concept of knowledge graph was first proposed by Google. It is a structured semantic knowledge used to describe concepts and their relationships. Its basic unit is the "entity-relation-entity" triplet. The entities are connected with each other through relationships to form a networked knowledge structure.

Knowledge graphs can generally be divided into open domain knowledge graphs and knowledge graphs for vertical domains. The open domain knowledge graph usually contains general domain knowledge such as YAGO, Freebase, NELL, etc[2]. The vertical domain knowledge graph is oriented to a specific field, usually based on a

specific field or industry. This type of knowledge graph has specific use objects and scenarios, such as the medical knowledge graph involved in the research of this article, but also includes the financial domain knowledge graph, the legal domain knowledge graph and so on. This paper involves the latter.

B. *Question Answering System*

Molla et al. Defined the question answering system in 2007 as an automaton that can answer questions in any natural language form [3]. Generally, the question answering system is regarded as inputting natural language questions and outputting a concise answer or a list of candidate answers. Each type of question answering system has three basic research questions: question analysis, information retrieval, and answer extraction.

Traditional question answering systems often use a large number of questions answering pairs to extract answers. While currently question answering system based on knowledge graph gives answers by two main types. One is to achieve a structured representation of the problem through semantic analysis; One is to structure the problem through vector representation [4].

C. *Medical Knowledge Graph*

The research on the knowledge map and question answering system of traditional Chinese medicine is currently focusing on the digitization of traditional Chinese medicine literature, while deep data analysis and application are relatively few. Tong Yu et al. Built a knowledge service platform for TCM health preservation, in which the graphical display of the medical concept was realized [5]. In terms of specific TCM question answering systems, CHEN Cheng et al. constructed a set of TCM knowledge question answering systems based on knowledge graphs [6]. The data were stored in a SQL SERVER database and the concept query was based on CRF word segmentation and rule matching methods. Ziqing Liu and others used deep learning to focus on CBOW and BILSTM + CRF for word vector training and relationship extraction and optimized the human-computer interaction interface of the retrieval module [7].

Generally speaking, there is relatively few researches on the knowledge graph of Chinese medicine and its question answering system. For the existing question answering system based on knowledge graph, their performances can be prompted. Therefore, in this paper we have adopted a new system, which has improved at a certain level in the accuracy.

III. Proposed Q&A System

A. *System Review*

The proposed Q&A system is divided into three modules, namely knowledge access module, semantic analysis module and user interaction module (Fig. 1). First, we obtain medical records from authoritative medical websites via web crawlers and store them as semi-structured text. Then we use the text extraction algorithm based on dictionary and BiLSTM + CRF to extract the relationship pairs we need and store them in the graph database [8]. For the data in the database, we segment and use node2vec to vectorize it [9]. When processing user requests, we will also perform the same operation on the user's questions, and match the dictionary with the information in the database according to the KNN algorithm [10], and return the corresponding data to achieve the question and answer function.

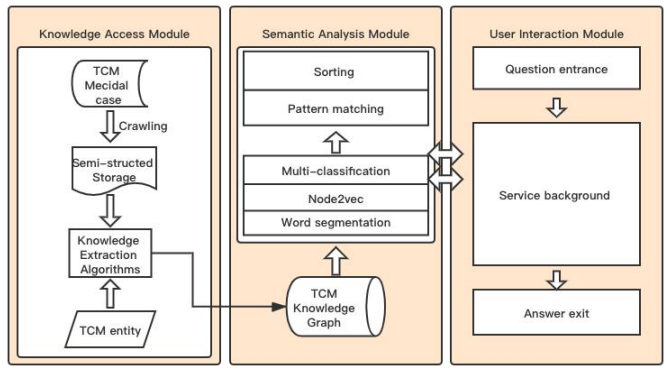

Fig. 1 Q&A System Structure

B. *Data acquisition of TCM medical records*

The data collection function is to collect the knowledge needed to construct the knowledge graph. Different sources and types of data are the knowledge base for constructing the knowledge graph in related fields, and also the data foundation of the knowledge question answering system in this article. The data used in the graph is grabbed from the medical website "National Service Platform for Famous Old Chinese Medicine Experience" (http://www.gjmlzy.com). Using the Python, we write a crawler script to grab relevant information page to page.

Then we parsed the basic information as herbs, prescriptions, diseases, symptoms and their relationships. After data cleaning, all entities and relationships data are separately captured and stored in a format of text (Fig.2).

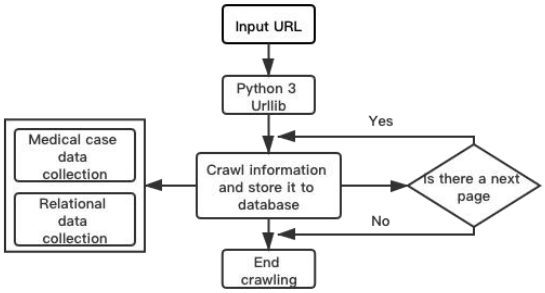

Fig. 2 The Crawler for TCM Records Acquisition

C. TCM Knowledge Graph Construction

In this part, we defined four entities and three relationships, and extracted from the text. Based on this, we construct the TCM knowledge graph.

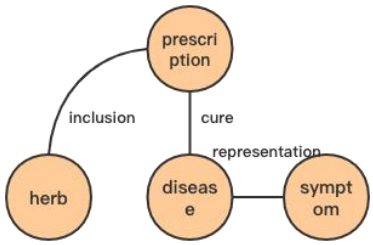

Fig. 3 Knowledge Graph Scheme

1) The construction of knowledge graph is the foundation of the question answering system. As shown in the figure3, according to the characteristics of the data, we select four entities and three relationships: cure between prescription and disease, inclusion between prescription and herb, representation between disease and symptom to define the triplet of knowledge graph.

2) After further cleaning the data, we use the dictionary-based method to extract information. Considering the particularity of medical case language, we also use in-depth learning methods, for example, BiLSTM-CRF to perform the information extraction [11]. In the end, we obtained herbs, prescriptions, diseases and symptoms, and obtained their relationships in turn.

3) After obtaining entity data and relational data, we construct corresponding RDF triplets. This system adopts Neo4j non-relational graph database to store the defined entities, relationships and their attributes.

D. TCM Question and Answer System Design

After getting a knowledge graph, we design to realize human-computer interaction. We expect users to enter their own symptoms, then we judge the disease and return the corresponding prescription. So that an interaction is completed.

1) Question and answer system modules

Q&A system can be divided into three modules, input module, matching module and output module. As shown in the figure3, the purpose of such a system is to provide a prescription P for users' question Q. In this process, for each sentence in the question, we first use the Jieba word segmentation tool to segment it. And then we extract the information to get each key word i. After that, we call KNN algorithm to find the most likely answer from embedded graph. Finally, we designed a response template that provides user the best recommended answer.

Algorithm 1 QA System

INPUT: *Question* Q
OUTPUT: *Prescription* P

```
function PROCEDURE QA(Q)
    for sentence in Q do
        J = Jieba(sentence)
        i = J.extraction()
    end for
    while i do
        Knn(i)
        if success then
            return top1 P
        else
            return False
        end if
    end while
end function
```

Fig. 4 Pseudo-code of TCM Q&A Algorithm

2) Algorithm explanation

For word segmentation, we can use the jieba word segmentation tool. For the data stored in the knowledge map, we first use node2vec algorithm to vectorize the information to represent the relationship between languages. Then we use KNN to classify the information and get the classification relationship between disease and prescription nodes (Fig.5).

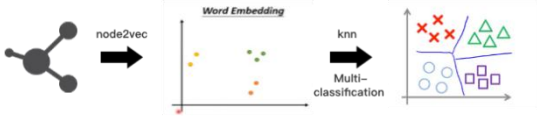

Fig. 5 Arithmetic flow

In this way, we can find the corresponding prescription based on the described symptoms and get relevant information based on the knowledge graph.

3) provide answers

In the information retrieval of the database, we write answer templates in advance. Then we find the properties in the database based on the recommended prescriptions, which will be provided to users later. Users can also conveniently view the knowledge graph independently.

IV. Experiments

A. TCM Record Data Description

The knowledge graph knowledge is crawled from the website "National Service Platform for Famous Old Chinese Medicine Experience"

the storage format of a single piece of data are:

{'url': 'data', 'basic_info': 'data', 'symptom': 'data', 'herb':' data ',' disease':' data ',' prescription ':' data',' usage ':' data '}.

After removing null and outliers, we store the information in the knowledge graph. Then we get 1709 diseases and 451 prescriptions and their relationships.

B. Experiment

In this experiment, we use KNN proximity algorithm to classify nodes. After comparing Euclidean distance and cosine similarity, we found that using cosine similarity to calculate the vector distance can achieve better results. Next, we divided 80% of the data of disease and prescription into training sets and 20% into test sets for further training. After training, from 1 to 15, we selected the number of output prescriptions in turn, calculated the accuracy of prediction, The calculation formula of Hit ratio is:

$$\text{Hit ratio} = \frac{\text{Hit}}{\text{Hit} + \text{Miss}}$$

Finally, we drew the corresponding Hit ratio curve.

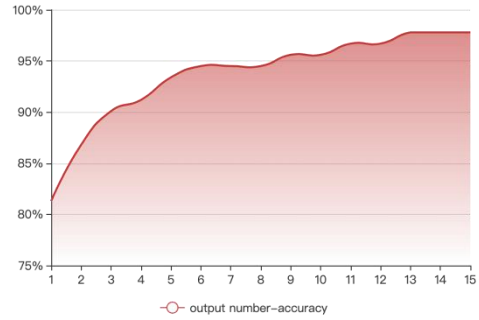

Fig. 6 Hit ratio curve

From the fig.6, it can be seen that when the output number is 1, the accuracy rate can reach 81.3%. Subsequently, the accuracy rate keeps rising and reached 97.8% when the output number is 15, achieving good results.

C. Discussion

In this sub-secession, we discussion the limitation of our work. In this paper, for the purpose of the verifying the TCM Q&A approach, the knowledge graph we constructed only selects four entities and three relationships. In subsequent applications, we hope to add more information to improve accuracy and usability.

Besides, this paper demonstrates a Q&A System in TCM field. However, these provided Knowledge Graph and algorithm can be extended to any other fields, such as financial.

V. Conclutions

This paper constructs a question and answer system based on the TCM knowledge graph. Then we uses Node2vec to realize the vector representation of the graph, and uses the KNN algorithm to classify and predict the information. The final experiment verifies that our system has good performance. In the future work, we will use the graph neural network (GNN) to train the vector, so as to further improve the accuracy of question answering system.

References

[1] Zhu, Jihe and Arsovska, Blagica and Kozovska, Kristina (2018) The impact and role of the Traditional Chinese Medicine on human health. IOSR Journal of Dental and Medical Sciences (IOSR-JDMS), 17 (7). pp. 80-82. ISSN 2279-0853

[2] Suchanek F M, Kasneci G, Iikum G. YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia[J]. Semantic Ib, 2007: 10.

[3] Molla D, Vicedo J L. Question answering in restricted domains: an overview[J]. Computational Linguistics, 2007, 33(1): 41−61.

[4] Zhiyuan Liu, Maosong Sun, Yankai Lin, and Ruobing Xie. 2016. Knowledge Representation Learning: A Review. Journal of Computer Research and Development, 53(2):247-261.

[5] Tong Yu, Jinghua Li, Qi Yu, Ye Tian, Xiaofeng Shun, Lili Xu, Ling Zhu, Hongjie Gao, Knowledge graph for TCM health preservation: Design, construction, and applications, Artificial Intelligence in Medicine, Volume 77, March 2017, Pages 48-52, ISSN 0933-3657

[6] Cheng Chen, Jie Di, Jinyu Qin, Jia Jiang, Haixia Wu, Tingting Cai. Design and Development of Intelligent Question Answering System Based on Knowledge Map of Chinese Medicine[J]. China New Telecommunications,2018,20(02):204-207.

[7] Ziqing Liu, EnIi Peng, Shixing Yan, Guozheng Li, Tianyong Hao.T-Know: A Knowledge Graph-based Question answering and Information Retrieval System for Traditional Chinese Medicine. Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, pages 15-19 Santa Fe, New Mexico, USA, August 20-26, 2018.

[8] Yipin Zhang, Bei Guan, Yinrun lu, Chong Wang, Bingchao Wu, Yongji Wang, Shixuan Bi. Study on the Entity Extraction Method of Traditional Chinese Medicine on the Basis of Deep Learning[J]. Journal of Medical Informatics,2019,40(02):58-63.

[9] Grover A, Leskovec J. node2vec: Scalable feature learning for networks[C] //Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. 2016: 855-864.

[10] Zhang N, Jia Z, Shi Z. Text categorization with KNN algorithm[J]. Computer Engineering, 2005, 8.

[11] Chen T, Xu R, He Y, et al. Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN[J]. Expert Systems with Applications, 2017, 72: 221-230.