ISSN: 2472-3703 DOI: 10.6911/WSRJ.202301_9(1).0009

Construction of Traditional Chinese Medicine Knowledge Graph Question Answering System Based on Data Augmentation

Chengming Li[1, a], Jun Pan[1, b, *]

[1]Zhejiang University of Science and Technology, Hangzhou, 510000, China

[a]951520286@qq.com, [b]panjun78@qq.com

Abstract

The field of Traditional Chinese Medicine (TCM) contains a large amount of knowledge data, but the organization of the data is very different from that required by modern information technology, and currently, the Chinese language domain lacks a knowledge graph related to the field of TCM. On the one hand, there are difficulties in constructing a high-quality knowledge graph in the field of Chinese medicine; on the other hand, automated QA requires a system with strong natural language understanding, and there is a lack of a knowledge base QA training corpus in the field of Chinese medicine, and there are challenges in automated QA tasks based on deep learning. This paper proposes a study on the construction of QA system based on TCM knowledge graph. Firstly, a training corpus of question sentences is generated and automatically annotated based on the existing knowledge base, secondly, a BiLSTM-CRF entity recognition model for question sentences and a BERT-TextCNN intention recognition model are used to achieve semantic parsing of the question sentences, and finally, the knowledge base answer query is completed by converting natural language question sentences to query statements. The aim of this thesis is to solve the difficulties in the application of knowledge graphs in the field of TCM through the above-mentioned research content, and to improve the natural language understanding capability of the automatic QA system based on knowledge graph through deep learning models, so as to finally build a QA system that can meet the needs of users and provide practical help in the popularization of knowledge in the field of TCM and the clinical application of TCM.

Keywords

Knowledge graph; QA system; Named entity recognition; Relation extraction; Deep learning.

1. INTRODUCTION

"Modernization of medicine" has become an important issue for both TCM and data science disciplines[1]. As an important part of the field of Chinese medicine, herbal medicine has also formed a complete and large system, but although the data of Chinese medicine is rich, it is at the same time very different from the form of data organization required by modern information technology. At present, most of the data and knowledge related to TCM have been stored in books or in the network in a semi-structured or unstructured form, and the lack of a more intuitive form of data storage as well as the difficulty of seeing the connection between certain kinds of TCM including composition, origin, efficacy, etc. in ordinary databases has largely limited and restricted the promotion and utilization of knowledge in the field of TCM[2].

QA systems are providing a new paradigm for human-computer interaction, correctly understanding the questions described by users in natural language and returning accurate

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202301_9(1).0009

answers based on the user's true intentions make automatic question and answer systems the new form of search engines[3]. The answers of traditional QA system mainly come from unstructured data such as web documents, encyclopedic knowledge, interactive communities, etc. With the concept of knowledge graph and the construction of domain knowledge graph, the research hotspot of automatic QA system is gradually shifted to the research and application direction of automatic QA system based on knowledge graph[4]. Knowledge graph has become a fundamental data service widely used in more and more industries to provide basic data support for various intelligent applications in the upper layer. Knowledge graph-based QA system is a form of QA system construction, and using knowledge graph as the data source of knowledge base, knowledge graph can play a crucial role in providing accurate QA services. At present, the technology of building knowledge graph and QA system based on knowledge base has been relatively mature, but there are still some problems in the actual application scenarios of both of them.

Most of the current open domain knowledge graphs are large-scale encyclopedic knowledge bases, and there is a lack of high-quality knowledge graphs in limited domains, especially in Chinese medicine.

The lack of high-quality labeled training corpus in the knowledge base of domain QA systems makes it difficult to apply QA systems based on deep learning models that can achieve good results to limited domains.

Based on the above discussion of knowledge graph and QA system and the related existing problems, this paper constructs a complete TCM knowledge graph, followed by the construction of a high-quality training corpus for QA system based on the knowledge graph database through data generation and automatic annotation, and designs and builds an QA system based on the TCM knowledge graph. The system is designed and built based on the TCM knowledge graph. The system provides systematic support for knowledge popularization and retrieval in the field of TCM.

2. FRAMEWORK OF QA SYSTEM

The framework of the TCM knowledge graph-based QA system focuses on the specific application of the information extraction methods proposed and used in this paper in four modules: data generation, question analysis, information retrieval and answer extraction.The overall framework is shown in Fig 1.

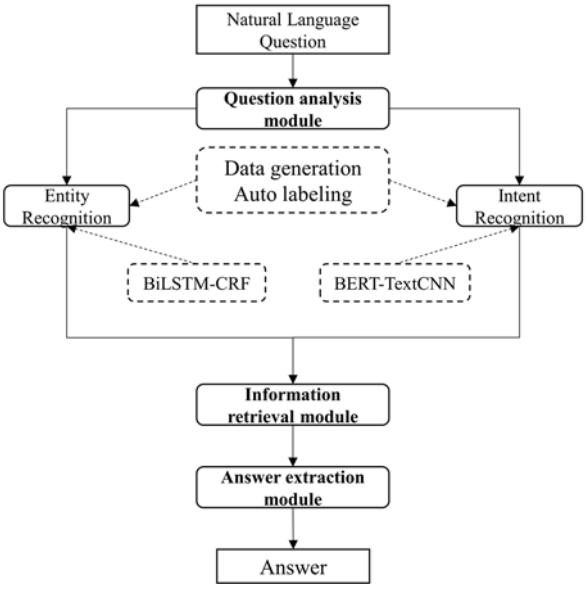

Figure 1. QA system construction framework

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202301_9(1).0009

2.1. Data generation module

Generate training data for deep learning models for entity recognition and intent recognition at scale. Firstly, using the herbal medicine related entities, relations and attributes of the herbal medicine knowledge graph constructed in this paper as the basis for data generation, a training corpus with entity labels is generated by formulating an interrogative template combined with entities and relations, and the statements with entity labels can be transformed into training data in the format required by the entity recognition model with the help of an automatic annotation tool before the entity recognition model is trained; and before training an intention recognition model, the same question can be labeled with an intention category label.

Question analysis module

The aim is to perform text pre-processing, entity recognition and intent recognition on user input question sentences. In this paper, entity recognition and intent recognition of question sentences are accomplished through deep learning methods. For the entity recognition task, the BiLSTM-CRF deep learning model is used; while for the intention recognition task, considering that the application scenario of this paper is the query of herbal medicine related knowledge and the user's question is mostly a simple question, i.e. the intention recognition is essentially a short text classification task, which is implemented by the BERT-TextCNN deep learning model. The training data used for model training in this module is the interrogative sentences with entity labels and intention classification labels constructed by the data generation module.

BiLSTM-CRF entity recognition model

With the continuous development of deep learning models as well as natural language processing techniques, deep learning algorithms are commonly used in information extraction and QA based on knowledge graph, so this paper uses a BiLSTM-CRF entity recognition model to identify herbal-related entities in natural language question sentences. The model takes the form of a word vector as input to the sentence, obtains contextual information through the BiLSTM layer, outputs the predicted label for each word and uses it as input to the CRF layer, and finally outputs the sequence of labels with the highest probability.

2.2. BERT-TextCNN intent recognition model

The intent recognition task involved in this paper is essentially a short text classification task, and the BERT-TextCNN model is used to accomplish the intention recognition task in this paper. The convolutional, pooling, fusion and fully-connected layers in the model are part of the TextCNN model. The model first obtains a vectorized representation of the training corpus after being trained by the BERT model, then selects the feature vectors suitable for text classification through filters of different sizes and pooling operations, converts the dimensionality of the feature vectors into the number of labels using the fully-connected layer, and finally outputs the predicted category labels through the softmax classifier The predicted category labels are output.

BERT (Bidirectional Encoder Representation from Transformer), a bi-directional encoder representation based on Transformer, is a pre-trained model proposed by Google AI Institute[5]. The most significant feature of the model is the pre-trained part, which shows that it can be directly inherited and used by later developers. The model was pre-trained on a large corpus of text (2.5 billion words of Wikipedia's unlabeled symbolic text corpus and 800 million words of book corpus) to gain a deeper understanding of how language works. For the input text, a word embedding vector, a segmented embedding vector and a positional encoding vector are directly summed as the model input, and the corresponding word vector output is obtained by a bi-directional Transformer encoder.

TextCNN (Text Convolutional Neural Network) is a variant of the convolutional neural network for textual tasks[6]. It extracts local features of different sizes from text sequences by

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202301_9(1).0009

setting different filter kernel sizes, so that the extracted feature vectors are diverse and representative. Compared with traditional CNN networks, TextCNN has no change in network structure: it mainly consists of a convolutional layer, a pooling layer and a fully-connected layer, but the difference lies in the different input data. In this paper, the BERT word embedding is used as the initial input to the model; subsequently, different n-gram features are extracted through the convolution layer as the input to the pooling layer; in the pooling layer, a number of one-dimensional vectors obtained after convolution are stitched together as the output of this layer after taking the maximum value. The fusion layer stitches the features obtained from the three pooling layers into a more representative text vector, and the features are finally passed through the fully-connected layer to obtain the final classification result.

2.3. Information retrieval module

Use the entities and question categories extracted in the interrogative analysis to create a mapping of questions to the database.

2.4. Answer extraction module

With these two modules, the natural language questions entered by the user are converted into Cypher query statements that return the Neo4j graph database query entities, relationships and attributes, and the final answer is returned by the query results combined with the developed answer template.

3. CONSTRUCTION OF TCM KNOWLEDGE GRAPH

Through data extraction and knowledge fusion, this paper constructs two kinds of knowledge graphs, herbal medicine graph and prescription knowledge graph, with a total of more than 200,000 entities and more than 300,000 triadic relationships.

Entities of the TCM knowledge graph correspond to independent objects in the real world ,Neo4j supports adding attributes to nodes and relationships, but it is not suitable to store the attributes of entities directly as node attributes in the knowledge graph. For example, in the column of "Basic information of Chinese herbal medicines" in the base data, each herb contains seven kinds of attributes, such as hanyu pinyin, original document, herbal base, nature and aptitude, etc., which should be expressed as seven kinds of attributes (entity, attribute) in the knowledge graph, and if stored as attributes of entity nodes, it is difficult to store the corresponding association information of the attribute. Therefore, this paper stores the attributes contained in the entity as nodes as well, and uses the attribute name as the relationship between the entity node and the attribute node, thus establishing a connection for the two nodes. In this way the data format of (entity, attribute) can be converted to a ternary form of (entity, attribute name, attribute value), e.g., (Gui Zhi, to the heart, lung and bladder meridians) can be converted to a ternary (Gui Zhi, to the meridians, to the heart, lung and bladder meridians).

In this paper, the attributes of the structured data "Basic information of Chinese herbal medicine" are transformed into a triad and a knowledge graph is constructed accordingly. In the knowledge map of Chinese herbal medicines, three types of attributes with weak meaning, namely, hanyu pinyin, original documents and original Latin plant, animal and mineral names, are removed from the knowledge map, and four types of attributes, namely, herb base origin, nature and taste, function and main treatment and geographical distribution, are retained. During the data processing stage, the attributes of the herbal entities were split at a fine-grained level, while retaining the original attributes, so that the associated knowledge of herbal medicines can be flexibly queried through each dimension. There are 12,997 Chinese herbal medicines in total, with blue nodes representing the names of the herbs and the remaining

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202301_9(1).0009

coloured nodes representing the attributes of the herbs, and the relationship names of the attribute categories.

The attributes of the structured data "Chinese medicine prescription" are also transformed into a triad format (prescription name, prescription attribute, attribute value) and a Chinese medicine prescription knowledge graph is constructed. In the prescription knowledge map, the prescription entity is the centre of the knowledge graph, and three types of attributes can be queried: prescription, prescription preparation and prescription main treatment. There are 34,333 prescriptions in the map (there are some duplicate data because the same prescription is recorded in different Chinese medical texts). The blue nodes are the prescription names and the remaining colour nodes indicate the attributes of the prescriptions, and the relationship names are the attribute category names.

4. IMPLEMENTATION OF QA SYSTEM BASED ON TCM KNOWLEDGE GRAPH

4.1. Question corpus generation

The effectiveness of deep learning models relies on large-scale training data, and the lack of large-scale domain data accumulation in restricted domains usually makes it difficult to apply deep learning models to specific domains. The methods used in the subsequent question and answer parsing tasks in this paper are all based on deep learning models, which require large scale training corpus to obtain good model results. Therefore, this paper proposes a data augmentation-based question corpus generation technique to generate high-quality, large-scale question corpus and annotation information from existing herbal-related entities and small-scale question corpus.

First, the entities, relationships and attributes in the TCM knowledge graph constructed in this paper are used as the knowledge base for data generation. For example, the entities include the Chinese herbal medicine entities Gui Zhi Tang, Ma Huang Tang and Shao Yao, the prescription entities Gui Zhi Tang, Ma Huang Tang and Chai Hu Tang, etc., the entity relationships include Chinese herbal medicine - evidence and symptoms, prescription - evidence and symptoms, etc., and the entity attributes include the distribution, nature and taste of Chinese herbal medicine and its function and main treatment, etc. (In this paper, the attributes are also treated as relationships in the graph database in the construction of the knowledge map).

Secondly, a data generator is constructed by constructing question templates to generate a training corpus with entity annotation information. The annotation information contains two parts: entity recognition labels and intention recognition classification labels. In order to improve the generalization ability of the deep learning model, this paper generates an enhanced form of the original interrogative corpus through data augmentation techniques, which not only ensures that the basic semantic information contained in the original interrogative sentence remains unchanged, but also increases the variety of expressions of the utterance.

Finally, the original annotated corpus with entity labels is automatically transformed into a corpus that can be used for deep learning model training with the help of an automatic annotation tool, which eliminates the need for time-consuming manual annotation. For example, the aforementioned question "{{herbs:桂枝}}分布于什么地方？" can be transformed into BIO training data for the entity recognition model.

According to the data generation and automatic annotation method described above, this section generates more than 150,000 questions based on the knowledge base of Chinese herbal medicines in Chapter 2, including four categories of herbal medicine attributes, prescriptions, concoctions and main treatments of prescriptions. Among them, the questions on herbal attributes are based on the structured data "Basic Information Database of Chinese Herbs",

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202301_9(1).0009

which contains four types of herbal attributes, namely, herb base origin, nature and ascription, function and main treatment, and geographical distribution, and the specific question generation and data enhancement methods are as follows: taking the herbal entities as an example, the length of the data (i.e. the number of Chinese characters per entity) is between 1 and 8. The sample() method of pandas in Python and the RandomState() method of the numpy library were used to generate the source entities by randomly selecting 25% of each of the four attributes as the base, nature and aptitude, function and distribution (the meaning of random is that it can be ensured that the source entities are used in different ways according to the existing (the meaning of random is to ensure that different types of question sentences are generated from different entities according to the available entity data for model training, and after the selection, no more than 5% of the data is the same between two entities after text comparison, which means that they are effectively randomly selected), combined with the constructed question sentence template to generate a total of more than 10,000 question sentences starting from herbal entities, and then through data augmentation techniques, one question sentence is expanded into three, finally generating more than 30,000 question sentences of herbal attributes. Finally, for the "other" type of questions in the intent recognition, they are mainly used to identify and answer questions that are not related to the field of Chinese herbal medicine or whose answers do not exist in the database. Unrelated questions are added to the training dataset of the question and answer system constructed in this paper.

4.2. Semantic parsing based on deep learning

Pre-processing of questions

Before implementing the entity recognition and intent recognition tasks in the interrogative sentence analysis module, pre-processing operations such as word separation and deactivation of the user's input are required. In Chinese, although individual words are the basic unit of language, in most cases the semantics are expressed in the form of words. Therefore, word separation is an important part of the Chinese text processing task that cannot be ignored, followed by the conversion of sentences into word expressions, which constitutes the entire Chinese word separation task. In this paper, the Jieba word separation component is used to perform the word separation and lexical annotation operations. In addition, as most of the entities in the field of Chinese medicine, such as herbal medicines and Chinese herbal prescriptions, are not commonly used in everyday life, using only a generic lexicon may result in some words not being accurately segmented, so this paper also builds a custom lexicon in the field of Chinese herbal medicine to improve the precision of the segmentation.

Entity recognition of questions

In the QA system processing process, parsing the intention of the question sentence first requires completing the NER task to correctly identify the Chinese herbal medicine related entities involved in the question sentence. Implementing the word-based Chinese NER task requires first completing the task of word separation of the sentence, and whenever there is a word separation error, the model's judgement of entity boundaries will be affected, resulting in a wrong word separation , and the character-based approach performs better than the word-based approach on Chinese NER. In this paper, a BiLSTM-CRF entity recognition model is used to implement the interrogative entity recognition and extraction task. The data used in this paper is the entity-labelled interrogative corpus generated in Section 4.2. The entity-labelled interrogative corpus is converted into a training dataset for training the entity recognition model by an automatic annotation tool, using the BIO tagging system, where the prefix "B " denotes the beginning of a named entity, "I" denotes the remainder of a named entity, and "O" denotes the rest of a named entity.

Classification of questions

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202301_9(1).0009

After the entity recognition model has correctly extracted the herbal related entities queried in the question, the system still needs to understand the intention of the user's question and classify the user input question to improve the efficiency of the system query. In this paper, the intent recognition of question sentences, i.e., the short text classification task, is implemented by a bert-textcnn deep learning model. The dataset used for training the relational classification model is also derived from a corpus of interrogative sentences with relational labels obtained from the interrogative sentence generation module. A text word vector is first generated by the BERT model, which completes the conversion from natural language to word embeddings and is used as input to the TextCNN text classification model, followed by convolutional, pooling, fusion and fully-connected layers to finally achieve relational classification of the intention of the interrogative sentences.

The convolution layer can use different sizes of convolutional kernels to extract local features in different windows of the question sentence e.g. size 2, 3, 4, 5 means that a window can contain 2, 3, 4, 5 words respectively, and defines the number of convolutional kernels e.g. 128 or 256. The output of a particular convolution is pooled by the pooling layer to obtain the maximum feature value in a certain dimension and output A feature vector carrying significant features representing the semantics of the sentence. The resulting vector is connected to a fully-connected layer and then a softmax classifier is used to implement a multi-classification task for herbal correlations.

4.3. Database answer search and generation

After entity recognition and intent recognition (relational mapping) of the question, the system needs to establish a mapping from the question to the graph database. The entities extracted from the question are described in Cypher query statements and returned to the Neo4j graph database for querying. The query results are generated according to the designed answer template and the final answer is given back to the user, who analyses the results to determine if they meet their needs.

For example, the user inputs the natural language question "桂枝常分布于哪里？" The two deep learning models, entity recognition and intent recognition, can extract the entities and question categories contained in the question and transform them into Cypher statements such as Match(a)-[:geographic distribution]-(b) where b.name='桂枝' return a.name to achieve a graph database query.

4.4. Implementation of knowledge graph based QA system

This section integrates the research methods proposed in the previous paper with the functions of the system that have been implemented, with the aim of integrating the system so that the knowledge map question and answer system can eventually be practically applied to people's daily lives and achieve knowledge popularization in the field of Chinese medicine, and with the expansion of the scale of the knowledge graph, the question and answer system can also be used to assist Chinese medicine practitioners in clinical diagnosis and improve the accuracy of clinical knowledge application. The system is based on the knowledge map of Chinese herbal medicine constructed in this paper as the database, so the types of questions are limited and the scope is restricted to the field of Chinese medicine herbal medicine. The question and answer system is ultimately provided by a web front-end interface to provide the user with a query function for knowledge related to herbal medicine.

System design analysis

From the user's point of view, the focus is on the convenience and comfort in the use of the Q&A system, mainly including whether the interface is brief and friendly, whether the operation is convenient and whether the effect is good, therefore, the system design should focus on these issues. The whole system design process is considered as follows: after the Q&A system gets the

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202301_9(1).0009

natural language questions input by the user, it will first perform the word separation process, return the entities contained in the question through the entity recognition module, perform intention recognition (relationship classification prediction) on the returned entities and question features, then convert the obtained results into Neo4j's Cypher query language and return to Neo4j for querying, and finally The answer is returned to the user in combination with a manually set answer template.

The QA system uses a web framework based on Django in the Python language to achieve data interaction between the web front-end and back-end functions, with the back-end server-side functions also written in Python. The first step in the implementation of the knowledge graph question and answer system is to build a knowledge graph. The key issue in the subsequent question and answer module is question analysis, and this paper uses a manual template based question and answer. The question and answer analysis task can be transformed into two sub-tasks of herbal named entity recognition and relationship classification, and finally the recognition result and question type are matched to the set manual template.

System implementation

The front-end interactive interface of the QA system contains functions such as entity recognition, herbal entity query, herbal relationship query and herbal mapping question and answer. The first is the entity recognition module, where the system can identify herbal-related entities from user input (this function contains only one task of entity recognition, so the user can enter any query content rather than content related to the question template); the second is the entity query and relationship query modules, both of which are knowledge graph display functions. In order to facilitate users' use and operation, the query and display functions of the knowledge graph are integrated into the question and answer system in this section; finally, there is the herbal graph question and answer module, in which the system receives question requests in the background after the user has entered a question, and carries out word separation, entity recognition, intent recognition and classification of the question, and after successful classification the answers are returned to the graph database based on the semantic information involved in the question, and finally the answers are returned to the front-end page for display.

For example, if you enter the herb '白桂' on the herbal entity query function page and click on the query, the front-end will return the content to the back-end and convert it into a Cypher query statement:" MATCH (n:中草药) n.name = '白桂' return n.name". And then return to the Neo4j database to get the content, and finally the herbal entity "白桂" related nodes and relationships back to the front-end page to show the user, the query result is shown in Fig 2.

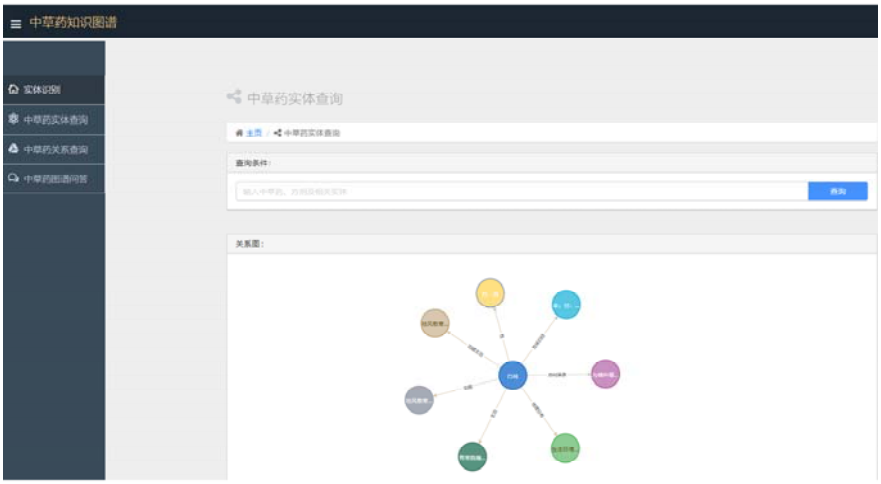

Figure 2. QA system query result

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202301_9(1).0009

Similarly, for the TCM knowledge graph QA function, the question "中药桂枝的功能主治有哪些？" is entered in the query interface. The front-end returns the natural language question to the back-end, and after the functions of word separation, entity recognition and intent recognition, it identifies that the question contains the entities "cinnamon stick" and "functional main treatment", and the question is classified as a query about the properties of Chinese herbal medicine, and the system returns the final answer from the Neo4j database The system obtains the relevant knowledge from the Neo4j database and then returns the final answer to the user in combination with the answer template, the query result is shown in Fig 3.

Figure 3. query result

5. CONCLUSION

In this paper, we use the knowledge graph as the database base, design the framework for building the automatic question and answer system, study the algorithm model required to realize each function of the QA system, and build a perfect QA system. This paper mainly carries out and completes the following research works. In order to address the problem of knowledge popularization in the field of Chinese medicine, an QA system based on knowledge graph is designed and constructed. Finally, an interactive interface is designed and implemented for user input and query.

ACKNOWLEDGMENTS

This work is supported in part Zhejiang Public Welfare Technology Application Research Project of China (Grant: LGN21F020003).

REFERENCES

[1] XIAO Peigen, XIAO Xiaohe. The 21st Century and the Modernization of Chinese Medicine[J]. China Journal of Chinese Materia Medica,2000(02):3-6.

[2] DENG Hongyong, XU Ji, ZHANG Yang, YUAN Min, SHI Yi. Analysis of the Current Status of Data Mining Research in Chinese Medicine[J]. Chinese Journal of Information on Traditional Chinese Medicine, 2012, 19(10):21-23.

[3] ZENG Shuai, WANG Shuai, YUAN Yong, NI Xiaochun, OUYUAN Yongji, Research Progress on Automated Q&A for Knowledge Automation[J]. Acta Automatica Sinica,2017,43(09):1491-1508.DOI:10.16383/j.aas.2017.c160667.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202301_9(1).0009

[4] YUAN Kaiqi, DENG Yang, CHEN Daoyuan, ZHANG Bing, LEI Kai. Advances in medical knowledge graph construction techniques and research[J]. Application Research of Computers, 2018, 35(07): 1929-1936.

[5] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[6] Zhang Y, Wallace B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification[J]. arXiv preprint arXiv:1510.03820, 2015.