IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 26, 2020, accepted January 25, 2021, date of publication January 28, 2021, date of current version February 5, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3055371

Research on Medical Question Answering System Based on Knowledge Graph

ZHIXUE JIANG[1], CHENGYING CHI[1], AND YUNYUN ZHAN[2]
[1]School of Computer and Software, University of Science and Technology LiaoNing, Anshan 114031, China
[2]College of Science and Health, Technological University Dublin, Dublin, D08 X622 Ireland

Corresponding author: Chengying Chi (chichengying@ustl.edu.cn)

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 61672138.

ABSTRACT To meet the high-efficiency question answering needs of existing patients and doctors, this system integrates medical professional knowledge, knowledge graphs, and question answering systems that conduct man-machine dialogue through natural language. This system locates the medical field, uses crawler technology to use vertical medical websites as data sources, and uses diseases as the core entity to construct a knowledge graph containing 44,000 knowledge entities of 7 types and 300,000 entities of 11 kinds. It is stored in the Neo4j graph database, using rule-based matching methods and string-matching algorithms to construct a domain lexicon to classify and query questions. This system has specific practical value in the medical field knowledge graph and question answering system.

INDEX TERMS Natural language processing, knowledge graph, question and answer system, medical knowledge.

I. INTRODUCTION

At present, there are many problems in China's manual medical interrogation system [1]–[4], mainly in: first, the allocation of medical resources in urban and rural areas is hugely unbalanced, some underdeveloped regions are short of professional medical personnel, and the quality of manual interrogation services needs to be improved urgently; Second, with the rapid growth and aging of China's population in recent years, the allocation of medical resources has gradually failed to meet the increasing medical needs of the people, and many experts are hard to find one vote. Third, Internet medical question-and-answer jobs are frequently chaotic. Irrelevant personnel without professional qualifications pretend to be experienced doctors for online consultation through part-time employment. Some minor diseases such as dizziness, fever, and stomachache that do not need to go to the hospital for diagnosis in person often become big problems when checked online. Based on the above issues, this paper proposes a knowledge graph based Q.A. system based on reliable medical data and studies the technical issues involved.

The knowledge graph is powerful for knowledge representation and management, which has obvious advantages in semantic retrieval [5]. With the in-depth study of artificial intelligence, knowledge graph has become a means to skillfully transform structured and unstructured knowledge, which has received extensive attention both in academia and industry. One of the most widely used, practically, direct vertical fields of knowledge graph is the medical field. China's application includes the knowledge graph of traditional Chinese medicine constructed by Shanghai Shuguang Hospital [6]. In the knowledge base-based question and answer field (KBQA), knowledge graph construction is the leading question and answer system. The construction technology of knowledge graph mainly includes top-down and bottom-up [7]. Top-down structure refers to extracting ontology and pattern information from high-quality data and adding it to the knowledge base with structured data sources such as Wikipedia. This construction method generally applies to the construction of an industry knowledge graph. For the industry, data content and data organization method are relatively easy to determine. The bottom-up construction uses specific technical means such as crawlers to collect and extract the required data from public resources, select the information with higher confidence, and add it to the knowledge base.

The mainstream implementation methods of a knowledge-based question answering system can divide into three

The associate editor coordinating the review of this manuscript and approving it for publication was Shadi Alawneh.

categories. First, based on Semantic Parsing: This kind of way transforms people's natural language into logical forms that can be processed by machines searches for and gives answers in the database. This method involves some linguistics and traditional NLP methods [8], requires many manual design rules, and has high accuracy but lacks generalization ability. Second, Information Extraction: This kind of approach extracts the feature information of relevant entities in sentences or knowledge bases and uses trained feature classifiers to sort the candidate answers and obtain the solutions. It is closely related to the traditional NLP method [9] and feature engineering, with strong generalization ability but relatively weak accuracy. Third, Vector Modeling: This kind of approach graphs the entities in the knowledge base and natural language questions to the same vector space and finds the answer by comparing these vectors' similarities. This method is a data-driven modeling method [10], which does not need too much preprocessing of data and is easy to implement. However, use keywords to represent knowledge content in isolation, ignoring the influence of context, which will affect the accuracy of the result [11].

II. RESEARCH ON KNOWLEDGE GRAPH

The traditional knowledge graph construction method goes through three steps: Data Acquisition, Information Extraction, and Knowledge Fusion. (Fig. 1)

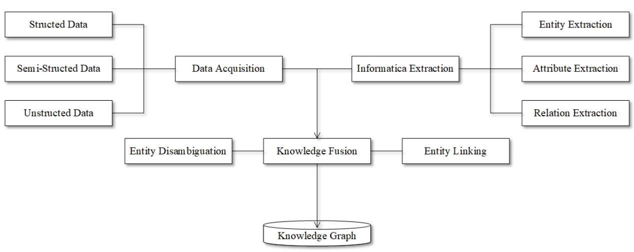

FIGURE 1. **Flow chart of constructing medical knowledge graph.**

A. DATA ACQUISITION

Data acquisition is the first step to establish a knowledge graph. At present, knowledge graph data sources can be divided into two types according to different application fields. One is professional data based on vertical areas. This part of the data is usually contained in the industry databases and stored in a structured way. It is a kind of non-public or semi-public data. The other is to capture the data disclosed on the network by grabbing, which usually appears in web pages and belongs to unstructured data, followed by data cleaning.

B. INFORMATION EXTRACTION

Information extraction is the second step in establishing knowledge graph. The critical problem is how to automatically extract information from data sources to obtain candidate knowledge units. The structured data can be used as input into the following question answering system only by simple preprocessing. However, semi-structured or unstructured data generally need natural language processing (NLP) technology to extract structured information, which is also the difficulty of information extraction [12]. The key technologies include Entity Extraction, Relation Extraction, and Attribute Extraction. The content of knowledge in the medical field is relatively straightforward, and the relationship is relatively short, so it is suitable for observation and classification directly through structured data.

Entity extraction, also known as Named Entity Recognition (NER), aims to establish "nodes" in knowledge graphs. The quality of entity extraction directly affects the efficiency and quality of subsequent knowledge acquisition and is a critical part of information extraction [13]. As the node of the graph, the entity can be obtained by sorting out the data, mainly including diagnostic items, departments, diseases, drugs, food, and symptoms (Tab. 1).

TABLE 1. **Entity types.**

Entity Type	Number of Entities	Examples
Checkup items	2,693	Hemoglobin
Department	50	Department of Respiratory Medicine
Disease	8,329	Pneumonia
Drug	3,685	Tinidazole tablets
Food	20,687	Black sesame porridge
Symptom	6,018	Ulcer pain

After entity extraction, a series of discrete named entities (nodes) are extracted from the text corpus. To obtain semantic information, it is necessary to extract the correlation (edge) between entities from the relevant corpus, thus connecting multiple entities or concepts to form a knowledge structure network. Data analysis mainly includes complications, subordinate departments, commonly used drugs, recommended diet, taboo diet, and items needed to check diseases (Tab. 2).

TABLE 2. **Entity relationship types.**

Entity Relationship Type	Number of Entities	Examples
Accompany_With	12,042	Dust mite allergy->allergic rhinitis
Belong_To	7,242	Allergic rhinitis otolaryngology
Common_Drugs	15,252	Allergic rhinitis->loratadine tablets
Recommend_Eating	32,452	Allergic rhinitis->honey
Avoid_Eating	23,322	Allergic rhinitis->almond
Required_Test	40,562	Allergic rhinitis->nasal mucosa and nasal secretions

The purpose of attribute extraction is to collect detailed information about specific entities from different information sources, thus completing entity attributes' characterization. Details can also be used as nodes of the graph. After observing the data again, it is concluded that they mainly include name, brief introduction, etiology, Yi population, treatment method, and treatment period (Tab. 3).

C. KNOWLEDGE FUSION

Knowledge fusion is the third step to build a knowledge graph. Due to the complex sources of knowledge extraction, the lack of hierarchy and logicality in the relationship between knowledge, redundancy of non-homologous

TABLE 3. Disease attribute types.

Entity Relationship Type	Number of Entities	Examples
Accompany_With	12,042	Dust mite allergy->allergic rhinitis
Belong_To	7,242	Allergic rhinitis otolaryngology
Common_Drugs	15,252	Allergic rhinitis->loratadine tablets
Recommend_Eating	32,452	Allergic rhinitis->honey
Avoid_Eating	23,322	Allergic rhinitis->almond
Required_Test	40,562	Allergic rhinitis->nasal mucosa and nasal secretions

knowledge, uneven knowledge expression, and other problems, a series of steps such as Entity Disambiguation and Entity Linking need to be involved in this process. In this system, the main application objects are entities, attributes, and relationships in the atlas triple.

Entity disambiguation aims to solve the phenomenon of polysemy. In the early disambiguation methods, Lesk [14] introduced an external dictionary into word sense disambiguation by comparing the word item's contextual language environment.

And the number of repetitive lexical items between the interpretation of the lexical item and the dictionary determines the lexical item's correct meaning. Still, this unsupervised method interprets the lexical item in articles and glossaries.

The matching degree requirement is too high to be suitable for the disambiguation of classic pieces. Given the rigor of medical question and answer knowledge, a certain number of entities have been designed in the knowledge definition part, so the clustering of entity categories is suitable for the improved K-Means algorithm [15]. The difficulty of artificial disambiguation can be significantly reduced using a clustering algorithm for the data to be added in the future.

Taking disease disambiguation as an example, the flow of the improved clustering algorithm is as follows:

Enter: n Documents to be categorized

Output: k Collection of documents for individuals with different diseases $D_1, D_2 \ldots D_k$

(1) initializing the integer part of which the number of clusters is $\sqrt{n}$;
(2) selecting an initial aggregation point of $\sqrt{n}$ according to formula (1), storing the aggregation point S in the set, and storing the index and the minimum distance in the collection S';
(3) Calculating the difference of the minimum distance between the two clustering points and storing it in the set S'';
(4) from finding the S'' point with the most massive distance difference, saving the previous aggregation points into the set S;
(5) Starting from this clustering center k, the K-means clustering algorithm is applied to obtain clustering results.

Through the above algorithm, we can automatically obtain k clustering center and get the final document set divided by disease individuals.

Entity linking refers to the operation of linking entity objects extracted from unstructured data (such as text) or semi-structured data (such as tables) to corresponding correct entity objects in the knowledge base. The basic idea is to select a group of candidate entity objects from the knowledge base according to the given entity objects and then link them to the correct entity objects through similarity calculation.

For calculating similarity, this paper adopts the calculation method of Jaccard similarity [16], which is suitable for extensive sparsity data and can compare the similarity and difference between finite sample sets. The larger the Jaccard coefficient, the higher the sample similarity. (Equations 1 and 2):

Given two sets AB, the Jaccard coefficient is defined as the ratio of the size of the intersection of A and B to the size of the union of A and B, as follows:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{1}$$

Defined $J(A, B)$ as 1 when all collections AB are empty.

Given the sum n of the two-dimensional vectors xy, the Jaccard coefficient is defined as follows:

$$J(A,B) = \frac{\Sigma_i min(x_i, y_i)}{\Sigma_i max(x_i, y_i)} \tag{2}$$

The phenomenon of multi-word and one meaning can be solved by entity linking through similarity calculation.

Finally, the Neo4J graph database can be used to visualize the knowledge graph after entity fusion is introduced in triple form. NEO4J uses graphs to represent data and their relationships. Its basic units are entities, relationships, and attributes, which can be intuitively seen as the relationships between entities in the knowledge graph. For data query, due to Cypher language's high retrieval efficiency and the use of index adjacency, fast and efficient target access can be realized, significantly improving the query speed and providing convenience for the next retrieval. Some visualization results are shown in Fig. 2.

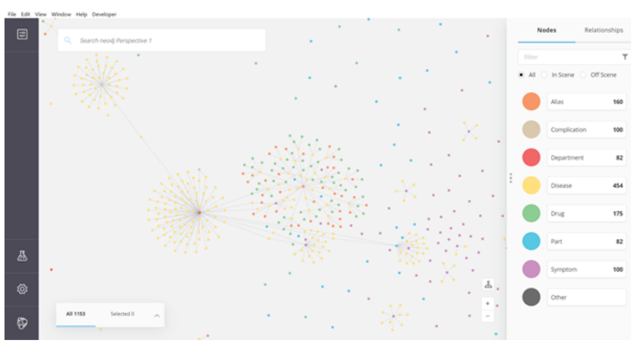

FIGURE 2. Representation of knowledge graph.

III. RESEARCH ON QUESTION ANSWERING SYSTEM

This QA system is based on semantic analysis and defines the feature word path, feature word, domain Actree, dictionary, and question word. The user's natural language questions are classified and analyzed to obtain the main entities and relationships. Finally, through intention prediction, Cypher

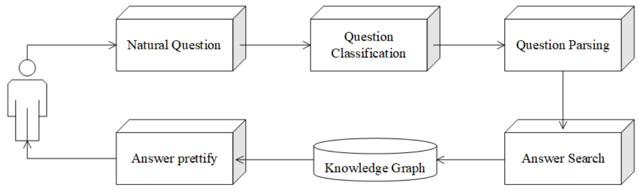

FIGURE 3. Construction flow chart of question answering system.

TABLE 4. Questions and answers supported by the question and answer system.

Question type	Examples of questions
Dis_Sym	What are the symptoms of allergic rhinitis
Sym_Dis	What is the reason for sneezing frequently
Dis_Cau	Why does someone get allergic rhinitis
Dis_Acc	What does allergic rhinitis do have a complication
Dis_Dru	What medicine should allergic rhinitis take
Dru_Dis	What diseases can loratadine treat
Dis_Che	What does allergic rhinitis need to check
Che_Dis	What can HPV check to detect
Dis_Pre	How to Prevent Allergic Rhinitis
Dis_Tim	How long can allergic rhinitis be cured
Dis_Way	How should allergic rhinitis be treated
Dis_Pro	How much is that cure probability of allergic rhinitis
Dis_PPL	Who is easy to get allergic rhinitis

language is constructed to query in the knowledge graph been built in the previous step. Yet, the query results are combed and modified again and returned to the user as natural language. See Fig. 3 for the specific flow chart.

A. QUESTION CLASSIFICATION

Based on the characteristics of medical knowledge of diseases, users' questions are minimal within a specific range. Question answering system is more suitable for designing template types for keyword matching. Specifically, users' queries are classified, and corresponding dictionaries are constructed, and then algorithms are used for multi-mode string matching. Therefore, this paper has designed some question-and-answer types supported by the question-and-answer system (Tab. 4).

B. QUESTION ANALYSIS

Firstly, through the domain words and their domains contained in AC (Aho-Corasick) algorithm [17], the entity types involved in the question are collected. At present, it is an excellent method to extract disease or symptom words from natural question language input by users by using existing dictionaries. Zhou *et al.* [18] proposed a DFM to learn subject detection and predicate matching jointly. However, to ensure the user's efficient question-and-answer experience, the matching algorithm must quickly match all words in the glossary at the same time. Therefore, this paper uses a multi-pattern matching method to check the entity dictionary's words from the user's information input. Pattern matching is the essential operation of strings in data structures. It is used to find all the same substrings from a given series. It is widely used in information retrieval, intrusion detection, and other fields.

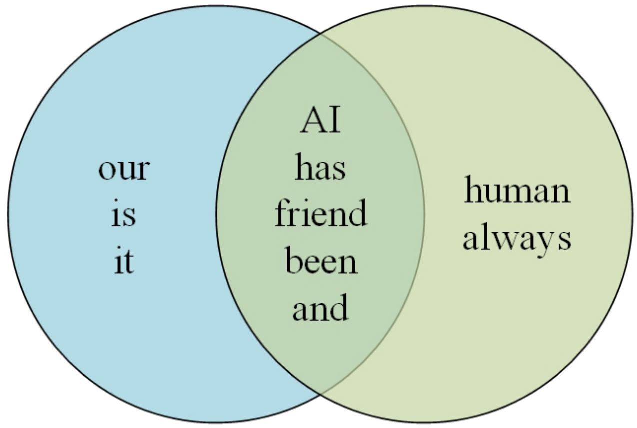

FIGURE 4. Intersection of sentences.

AC algorithm is a multi-pattern string matching algorithm that can match multiple strings from text and obtain relevant information such as its location and total number. The algorithm relies on constructing a finite state automaton to implement it. Ordinary automata cannot perform multi-pattern matching, while AC automata are realized by adding failover and transferring to suffixes of the successfully entered text. Assuming that the total length of the target string T is n and the set of pattern strings is $P = p_1, p_2, \ldots, p_m$, then the time complexity is $O(n)$, that is, within this complexity, all pattern strings can be matched within n without being affected by the size of m. The algorithm has an approximate linear time complexity when it is evenly spread, which is about the string's length plus the number of all matches. Therefore, the AC algorithm's time complexity is $O(N)$ in both the best and worst cases.

When full matching fails, the similarity calculation is used to find similar words. In this paper, the following similarity discrimination methods are adopted: First, Jaccard similarity [19], which is defined only as of the intersection size of two sentence words divided by the union size of two sentence words. For example:

Sentence 1: AI is our friend, and it has been friendly.

Sentence 2: AI and humans have always been friendly.

To calculate Jaccard similarity, this paper uses Lemmatization [20], a technique commonly used in English NLP, to replace words with roots that have the same sources. In the above example, "friends" and "friendly" have the same source and have the same root. We can draw the intersection and union of two sentence words, as shown in Fig. 4:

For the above two sentences, the Jaccard similarity is $5/(5+3+2) = 0.5$, i.e., The intersection of the words of the two sentences has five words, and the union has ten words.

Cosine similarity [21] is calculated by the angle between two vectors

$$Similarity = cos(\theta) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (3)$$

Evaluate the similarity between the two vectors. Since cosine similarity is calculated using vectors, we must first convert sentence text into corresponding vectors. There are many ways to convert sentences into vectors, the simplest of which is to use a bag of words [22]. In terms of text similarity, the two methods are not sensitive to text similarity. For example, if two words are synonyms, their vector directions are almost the same. If they are antonyms, their vector directions are practically opposite. Therefore, if the two words' semantics are more similar, their vector directions are often parallel, and the rules are the same.

TF-IDF [23] refers to the fact that a word or phrase frequently appears in one article but rarely in other items. It is believed that the word or phrase has good classification ability and is suitable for classification.

Word frequency refers to the frequency of a given the word appearing in a file. Its formula is as follows:

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{4}$$

TF-IDF assumes that high-frequency words should have a high weight. The inverse document frequency (IDF) is inversely proportional to the word frequency of words, i.e., The most common words given the smallest value, the more common words given smaller weight, and the less common words given more massive weight. The formula is as follows:

$$IDF_{i,j} = log\left(\frac{D}{1+j}\right) \tag{5}$$

After calculating TF and IDF values, respectively, multiplying them can obtain TF-IDF values, and its formula is as follows:

$$TF \bullet IDF_{i,j} = TF_{i,j} * IDF_{i,j} \tag{6}$$

After feature word extraction is completed, the system converts the user's question information into vector representation and inputs it into the intention classifier to identify the user's query intention. This paper is trapped in limited data sources. Through manually labeled data, keyword features, and word frequency-inverse document frequency features are constructed, and then the trained Naive Bayesian classification model is used to classify and identify intentions. Thus, the system has completed the task of question analysis.

C. QUERY PROCESSING

Due to the medical system's rigor, this system is not designed as a generative system but a retrieval system. Therefore, the information search module can generate search terms suitable for the database and play a role in processing information by searching the database. This system adopts two kinds of databases: the graph database and the relational database. Therefore, the information retrieval module will be composed of two sub-modules: the knowledge base retrieval sub-module and the question and answer retrieval sub-module. The question retrieval module judges the entity and intention identification passed in after the question analysis in the previous step and selects the sub-module to execute. Suppose there is the information of intent type in the incoming data. In that case, the knowledge base retrieval sub-module is completed. Query statement of a knowledge base is generated in sub-module, which is transferred to the knowledge base for a query. If there is no intention type information, a simple question-and-answer pairing query submodule is executed. In this sub-module, the problem classifier predicts the detailed types of problems and further narrows the retrieval range.

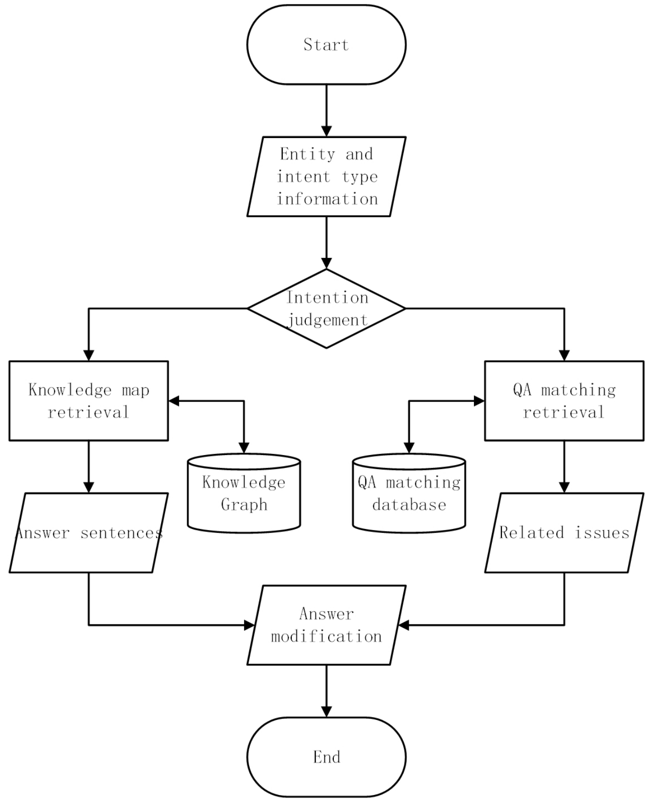

FIGURE 5. **Query processing module diagram.**

Through the classification results of the questions introduced in the previous step, the domain words and entity types in the questions are obtained. A dictionary is shaped like 'Entity Type': ['Domain Words'],... is constructed. To convert the results returned by question classification into Cypher language of NEO4J, call the corresponding reply template according to the corresponding question type, combine the query results with the answer template, and finally return the final answer. The corresponding flow chart is shown in Fig. 5:

After the question type is determined, the query instruction of the question-and-answer pairing database is generated. Then it is input into the question-and-answer pairing database to query the relevant question set. If the query result comes from the medical knowledge graph, the information retrieval module will generate the answer sentence and then transmit it to the user interaction interface. Suppose the query result comes from the question-and-answer matching database. In that case, the information retrieval module will directly call the candidate answer selection module to output the question-and-answer matching.

TABLE 5. Experimental environment of question answering system.

	Parameters	Configuration
Hardware environment	CPU	Intel (R) Core (TM) i7-9750H CPU @ 2.60 GHz
	Total Physical Memory	31.9 GB
	Hard disk	2T
Software environment	Operating System	Windows 10 Pro
	Python	3.6. 5

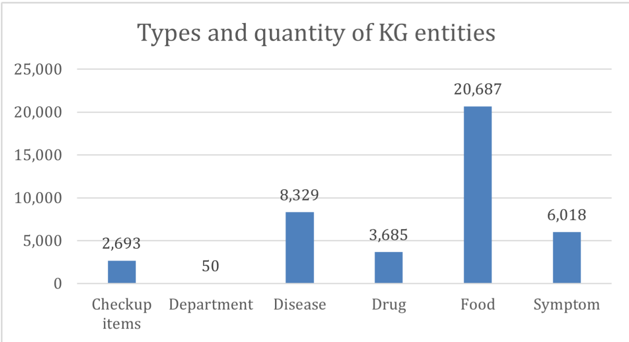

FIGURE 6. Types and quantities of atlas entities.

Some database problems are highly similar to the user input, but these problems cannot all be replaced by each other. Therefore, the module used for selecting candidate answers is mainly responsible for starting with the corresponding questions, checking the first several questions closest to the user's problems, and searching all the candidate answers to these questions. Many medical questions usually do not have a unique standard answer, so the medical question answering system needs to select the solution that best matches the user-input questions' semantics from these candidate answers. This module's essence is to use question and answer matching technology to calculate the semantic similarity score between the original problem and each candidate solution, select the answer with the highest score as the best answer, and finally directly return the user's response.

IV. EXPERIMENT

A. EXPERIMENTAL ENVIRONMENT

The algorithm and model proposed in this paper are all run in the same experimental environment, and the specific configuration is shown in Tab. 5.

B. EVALUATION AND ANALYSIS OF KNOWLEDGE GRAPH EXPERIMENT

1) DISPLAY OF GRAPH CONSTRUCTION RESULTS

This system locates the medical field, uses crawler technology to use vertical medical websites as data sources, takes diseases as core entities, and constructs a knowledge graph including 44,000 knowledge entities of 7 types and 300,000 entities of 11 classes, which are stored in Neo4J graph database. The results are as follows:

Fig. 6 shows the entity types in the knowledge graph and the corresponding quantitative relationships, It can be seen that diseases should dominate the medical knowledge graph, but in fact, the entities related to it are more corresponding food types, which reflects that in recent years, the medical question-and-answer field has paid more attention to the recommendation and taboo of food categories. The number of entities accounts for nearly half of the proportion in graph construction.

TABLE 6. Summary of experimental data for knowledge graph construction.

Data Set	K-means (max-min) F1 Value	Kmeans F1 Value	Lesk F1 Value
Seeking Doctors and Asking for Drugs Network	0.94	0.92	0.64
Clove Garden	0.95	0.91	0.56
39 Health Network	0.92	0.9	0.78

2) ANALYSIS OF ENTITY CLUSTERING METHOD

This experiment's data are collected from health medical networks such as 39 health networks and medical search and drug search networks, including 44,000 knowledge entities of 7 types and 300,000 relationship information of 11 types of entities. This experiment's data include medical information crawled from major health medical networks such as 39 health networks and drug search and drug search networks. This experiment's data include the related information of 44,000 knowledge entities of 7 types and 300,000 relationship information of 11 kinds. To construct a high-quality knowledge graph, When the data sources are relatively small, this paper focuses on optimizing and improving the knowledge fusion part of the knowledge graph construction algorithm. It proposes an improved k-means algorithm (k-means max-min) for entity clustering disambiguation to deal with almost the same text representation but different text meanings in a large number of texts. Based on comparing with other similar models, the most suitable K value is tested through experiments.

In this experiment, the total F1 value of precision rate and recall rate is used to measure the clustering disambiguation results of the entity fusion part in the knowledge graph using the improved K-means algorithm. Their formulas are as follows:

$$Precision = \frac{|Correct|}{|Predict|} \tag{7}$$

$$Recall = \frac{|Correct|}{|Labeled|} \tag{8}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{9}$$

In the above formula, Correct represents the intersection of the output result of the model and the correct work, Predict describes the prediction result of the model, and *Labebled* represents the right result. The balance between precision and recall is completed with $F1$ values.

The experimental data are summarized in Tab. 6.

It can be seen that the Lesk algorithm fluctuates significantly between different types of data sets because the method before improvement is limited by the size of external dictionaries and the fields to which questions and answers belong. In the process of disambiguation, each document's different characteristics lead to the mismatch of grammatical rules, so the effect of text clustering is also volatile. In contrast,

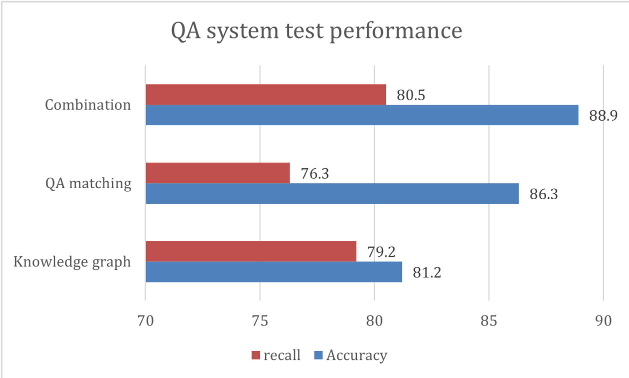

FIGURE 7. Test performance of question answering system.

the improved K-means algorithm determines the clustering point with the farthest convergence (minimum similarity) from a macro perspective before the K-means algorithm is executed when selecting the K value, and at the same time determines the K value from the previous data analysis. It can be seen from the experiment that the results obtained in this way are more consistent with the actual situation. It shows that the method proposed in this project has an apparent excellent effect in entity clustering disambiguation.

C. EXPERIMENTAL EVALUATION AND ANALYSIS OF QUESTION ANSWERING SYSTEM

In the actual use of the QA system, The inquiry of common diseases, drugs, symptoms, etc., is more efficient than the website's direct examination. On the one hand, users will not be interfered with by other irrelevant information due to the question-and-answer system's specificity. On the other hand, users' diagnosis results will be relatively accurate due to the original high-quality data and knowledge fusion steps.

At present, the question answering system's performance test is also based on accuracy and recall rate. The more questions a question answering system can answer (those that cannot be answered guide users to conduct further intention inquiries), the higher the system's recall rate. The more correct questions the question answering system responds to, the higher the accuracy. For the field of medical question and answer, people often focus on accuracy. That is, they would instead choose not to answer then give wrong answers.

Due to the QA system evaluation's subjectivity, even if a certain amount of test data is collected, human errors in the labeling process cannot be avoided. The system needs as many people as possible to take part in the test and then takes an average value as the final result. The procedure takes the approval of some answers collected on the network as a particular reference, further modifies the test data, and obtains the following products:

As can be seen from Fig. 7, when the knowledge graph is used for question-and-answer retrieval alone or question-and-answer pairing is used for recovery, the recall rate of knowledge graph question-and-answer is higher, but its accuracy rate is lower than that of question-and-answer pairing retrieval. The specific reason is that question-and-answer pairing can give more accurate response answers when the system reencounters the same problem by collecting the existing question-and-answer pairs on the network more accurately. However, due to its limited scale, it cannot achieve higher problem coverage. However, the knowledge graph just makes up for its deficiency. When the corresponding relationship between medical entities and related entities is analyzed, the graph pattern query can be carried out more accurately through intention classification. Therefore, we can combine the two to achieve higher accuracy and recall rate. However, it can be seen that the final accuracy still does not reach the ideal height because it is closely related to the quality of the data source.

V. CONCLUSION

This system establishes a high-quality medical question-and-answer knowledge graph based on relevant professional knowledge in medical question-and-answer research through steps such as information extraction and knowledge fusion. Based on this knowledge graph, problem domain words and interrogative words are added to perfect matching rules. This system aims to explore the improvement of the medical question-and-answer system's efficiency and accuracy, but there are still some defects. For example, the construction and retrieval speed of the knowledge graph still need to be improved, and the multi-round dialogue and complex knowledge reasoning need to be improved.

REFERENCES

[1] Y. Lin and S. Li, "Influencing factors and improvement of unbalanced allocation of medical resources in urban and rural areas," *Econ. Trends*, vol. 5, no. 9, pp. 57–68, 2016.

[2] W. Di, "Vigorously carry out special rectification action for medical disorders," *J. Traditional Chin. Med. Manage.*, vol. 10, no. 7, p. 29, 2019.

[3] Z. Xu and Q. He, "Analysis of the current chaos of over-medical treatment in China and its causes," *J. Beijing Inst. Electron. Sci. Technol.*, vol. 25, pp. 82–88, Jan. 2017.

[4] X. Hu, "Analysis and thinking on medical advertising chaos," *Doctor Online*, vol. 8, no. 12, pp. 44–45, 2018.

[5] Y. Kaiqi, D. Yang, C. Daoyuan, Z. Bing, and L. Kai, "Construction technology and research progress of medical knowledge graph," *Comput. Appl. Res.*, vol. 35, no. 7, pp. 1929–1936, 2018.

[6] R. Tong *et al.*, "Construction and application of knowledge graph of traditional Chinese medicine," *J. Med. Inform.*, vol. 37, no. 4, pp. 8–13, 2016.

[7] H. Mengwei *et al.*, "Review of knowledge graphping research and its application in medical field," *Comput. Res. Develop.*, vol. 55, no. 12, p. 2587, 2018.

[8] J. Berant, A. Chou, R. Frostig, and P. Liang, "Semantic parsing on freebase from question-answer pairs," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2013, pp. 1533–1544.

[9] X. Yao and B. Van Durme, "Information extraction over structured data: Question answering with freebase," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, 2014, pp. 956–966.

[10] A. Bordes, S. Chopra, and J. Weston, "Question answering with subgraph embeddings," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 615–620.

[11] B. Liu and H. Chen, "Discussion on vector space model information retrieval technology," *Intell. J.*, vol. 25, no. 7, pp. 92–93, 2016.

[12] Q. Wang, J. Liu, Y. Luo, B. Wang, and C.-Y. Lin, "Knowledge base completion via coupled path ranking," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2016, pp. 1308–1318.

[13] T. S. Lee, S. M. Shin, and S. S. Kang, "Named entity recognition for patent documents based on conditional random fields," *KIPS Trans. Softw. Data Eng.*, vol. 5, no. 9, pp. 419–424, Sep. 2016.
[14] M. Lesk, "Automatic sense disambiguation using machine readable dictionaries: How to tell a pine code from an ice cream cone," in *Proc. 5th Annu. Int. Conf. Syst. Document. (SIGDOC)*, 1986, pp. 24–26.
[15] X. Yang and Y. Wang, "Design and implementation of a personal name disambiguation system based on improved K-means algorithm," *in Proc. 7th Nat. Conf. Web Inf. Syst. Their Appl., 5th Nat. Conf. Semantic Web Ontol., 4th Nat. Conf. Electron. Government Technol. Appl. China Comput. Fed.*, vol. 17, 2012, pp. 10–12.
[16] S. Niwattanakul, J. Singthongchai, E. Naenudorn, and S. Wanapu, "Using of Jaccard coefficients for keywords simularity," in *Proc. Int. Multicultural Eng. Comput. Sci.*, 2013, vol. 1, no. 6, pp. 380–384.
[17] X. Wang and D. Pao, "Memory-based architecture for multicharacter Aho–Corasick string matching," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 26, no. 1, pp. 143–154, Jan. 2018.
[18] G. Zhou, Z. Xie, Z. Yu, and J. X. Huang, "DFM: A parameter-shared deep fused model for knowledge base question answering," *Inf. Sci.*, vol. 547, pp. 103–118, Feb. 2021.
[19] S. Bag, S. K. Kumar, and M. K. Tiwari, "An efficient recommendation generation using relevant Jaccard similarity," *Inf. Sci.*, vol. 483, pp. 53–64, May 2019.
[20] T. Bergmanis and S. Goldwater, "Context-sensitive neural lamentation with elements," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Lang., Hum. Lang. Technol.*, vol. 1, 2018, pp. 1391–1400.
[21] M. Abdel-Basset, M. Mohamed, M. Elhoseny, L. H. Son, F. Chiclana, and A. E.-N. H. Zaied, "Cosine similarity measures of bipolar neutrosophic set for diagnosis of bipolar disorder diseases," *Artif. Intell. Med.*, vol. 101, Nov. 2019, Art. no. 101735.
[22] R. Zhao and K. Mao, "Fuzzy bag-of-words model for document representation," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 2, pp. 794–804, Apr. 2018.
[23] S. Qaiser and R. Ali, "Text mining: Use of TF-IDF to examine the relevance of words to documents," *Int. J. Comput. Appl.*, vol. 181, no. 1, pp. 25–29, Jul. 2018.

ZHIXUE JIANG was born in Taiyuan, Shanxi, China, in 1996. He received the B.S. degree from the University of Science and Technology LiaoNing, where he is currently pursuing the M.S. degree. His mentor's main research interests include search engine, automatic abstract, question and answer system, and so on.

CHENGYING CHI, Author's photograph and biography not available at the time of publication.

YUNYUN ZHAN, Author's photograph and biography not available at the time of publication.