Intelligent question answering system for medical knowledge

Hongwu Zeng[a], Qiuyu Xu[*b]
[a]School of Medical Information, Chongqing Medical University, Chongqing 400016, China;
[b]Computer Laboratory of Experimental Teaching Management Center, Chongqing Medical University, Chongqing 400016, China

ABSTRACT

In recent years, intelligent question answering system has been widely used in various interactive services. Intelligent question answering system is to arrange the accumulated disordered corpus information in an orderly and scientific way, and establish a classification model based on knowledge to support the question answering of various forms. In this paper, vertical medical website data and electronic medical record data are fused to solve the mapping problem of colloquial medical vocabulary and medical professional vocabulary. PLSA (probabilistic latent semantic analysis) theme architecture model solves the hidden theme content crawling problem of web page crawler. The application of vector space model and knowledge graph in medical knowledge question answering system solves the problem of low efficiency in traditional interactive question answering system. After testing, it can accurately understand the user's intention, and the accuracy of question and answer is high.

Keywords: Artificial intelligence, clinical thinking, mining

1. INTRODUCTION

With the development of information technology, people have more and more means to obtain medical information. When ordinary people have uncomfortable symptoms, they will inquire the causes and treatment methods of the symptoms through the Internet at the first time[1]. However, due to the uneven network resources, users need to spend a lot of time to determine whether the search results are the answers they need. At the same time, even if they find the answers from the search engines, the results may not be correct or professional. Another kind of situation, when the patient goes to the hospital to see a doctor, do not know to hang that branch, but at this time guide the medical platform is overcrowded, too busy to take into account. The third situation is that medical college students mainly learn medical theory knowledge from freshman to junior year. At this stage, students cannot enter clinical practice and theory cannot be combined with practice[2].

Based on the above situation, this study uses crawler technology to crawl the relevant content of vertical medical websites such as Xunyao, Sanjiu Health website and Dingxiang Doctor, which includes diagnosis method, treatment plan and disease common sense[3]. At the same time, the relevant disease information is extracted from the hospital electronic medical record and the website crawl resources are integrated to form the relevant disease knowledge base. Entity, attribute and entity relationship were obtained from disease knowledge base by NLP technology, and disease knowledge map was constructed by Neo4j.The semantic recognition model is used to identify the input questions, and the input questions are converted into graph model queries, so that the most accurate medical answers can be quickly fed back to users.

2. SYSTEM ARCHITECTURE DESIGN

The architecture of medical knowledge intelligent question answering system based on knowledge graph is shown in Figure 1. The whole system is divided into three parts. The first part is data collation[4], the second part is knowledge graph construction, and the third part is intelligent question and answer. The data collation is divided into two parts, one part is data collation of vertical medical websites, medical websites mainly include Seek medical Advice, Sanjiu Health net and Dingxiang Doctor. The other part is hospital electronic medical record data. The electronic medical records of the hospital come from the Data Research Institute of Chongqing Medical University. The big data platform of the Data Research Institute includes the electronic medical records of the second affiliated Hospital of Chongqing Medical

* 101852@cqmu.edu.cn

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 125065B
 · 0277-786X · doi: 10.1117/12.2661853

University, stomatological Hospital, University Town hospital and some district and county hospitals in Chongqing. The second part is knowledge map construction. Neo4j is used to construct medical knowledge map based on medical entity, medical entity attribute and relationship between medical entity and medical entity obtained from the first part. The third part is the intelligent question answering part[5], which is divided into three steps: firstly, the natural language words input by the user are segmented, then the real intention of the user question is obtained by using the improved TF-IDF algorithm, and finally the answer is obtained by using the graph database query of Neo4j.

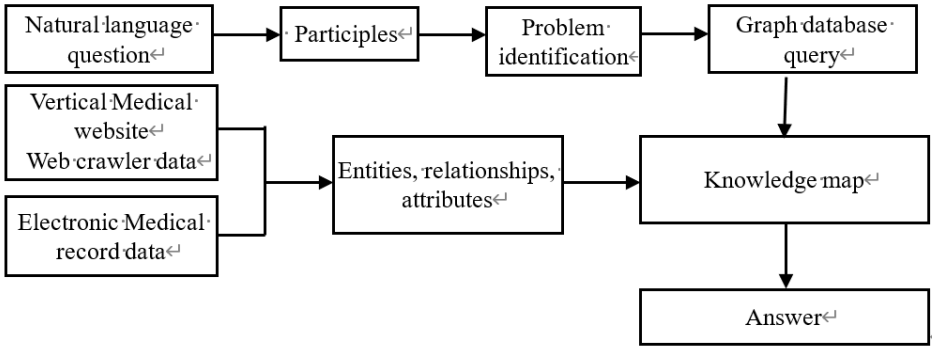

Figure 1. System architecture diagram.

3. DATA SORTING

Data collation is to obtain triplet data for constructing knowledge graph by integrating vertical medical website data with electronic medical record data. As electronic medical records data for the doctor to the patient's oral information described as structured language is formed after related professional vocabulary[6], but users in the use of intelligent question answering system input for non-professional colloquial vocabulary more questions, so you need to crawl the data on the relevant websites and electronic medical record data integration to form a more practical value triples data, in order to facilitate subsequent figure database queries[7].

3.1 Topic architecture based on PLSA model

In medical knowledge, the topic of disease is fixed, so it is necessary to consider not only the repetitive relationship between the page documents of vertical medical websites and the topic of disease, but also the semantic relationship between the page documents and the topic of disease. If the theme architecture is a good reflection of the page's hidden theme, the results of climbing vertical medical sites will be more accurate[8].

A page of a vertical medical website must describe a disease as the topic, but also a small amount of other topics, so using PLSA model to build crawler theme is very beneficial. Since PLSA uses the pocket model, we can think of page documents as being independently interchangeable with each other, and words within the same page document as being independently interchangeable. This model is defined as follows:

(1) Suppose a disease has n subject words, denoted as, since the subject of the disease is fixed, the subject words are fixed.

(2) The vocabulary of n page documents is defined as $W = (d_1, d_2, ..., d_n)$, where each page document has a specific page theme.

(3) Assuming that we have a number for each word on each page, then in the PLSA model, the generation probability of each word in document DM in article M is:

$$P(\varphi|d_m) = \sum_{z=1}^{k} p(\varphi|z)p(z|d_m)$$
$$= \sum_{z=1}^{k} \delta_{z\varphi}\, \delta_{\varphi z}$$

The probability of all words and keywords in this article is:

$$P(\varphi|d_m) = \prod_{i=1}^{n} \sum_{z=1}^{k} \delta_{z\varphi}\, \delta_{\varphi z}$$

This algorithm can calculate the probability of page content and topic efficiently and quickly when given topic crawler, so as to crawl topic content more effectively. This topic is a specific and comprehensive description of a disease[9]. The comprehensive description here includes the general knowledge of the disease, diagnosis methods, treatment plans and other whole process to describe the whole diagnosis and treatment process of the disease.

The data in this paper is presented in JSON form after collation, with a total of 8807 pieces of data. Figure 2 shows the data representation after collation.

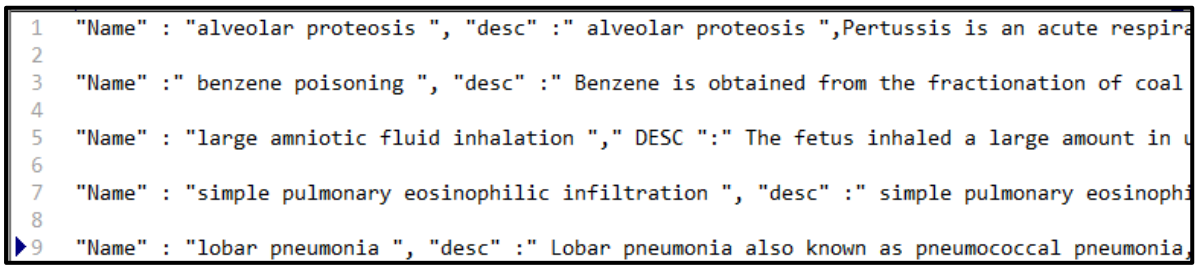

Figure 2. Sorted JSON data.

4. REALIZATION OF MEDICAL KNOWLEDGE ATLAS

The concept layer of medical knowledge graph is formed by obtaining various medical entities through sorting out the data of vertical medical website and electronic medical record. Entity relationships are then formed based on these medical entities. In this paper, neo4j graph database is used to construct knowledge maps. In Neo4j, entity storage forms are mainly expressed by nodes, and the relationship between nodes is expressed by edges.

4.1 Node identification and edge formation

Traditionally, the text with substantial meaning in the text is called entity, and the node in Neo4j is the medical entity we are looking for. The relationships between entities are edges in Neo4J. All medical entities are labeled according to the part of speech labeling method. For example, the statement “Pulmonary candidiasis is a common pulmonary mycosis caused by candida infection (mainly white Candida) and is generally treated in the outpatient department of internal medicine or respiratory Medicine”, where “pulmonary candidiasis” is identified as Disease, “Internal Medicine” and “respiratory Medicine” are identified as Department. Another example is “acute respiratory infections caused by bordetella pertussis. It is characterized by paroxysmal spasmodic cough. The word “cough” in this sentence is marked as a Symptom entity.

In this study, crawler data and electronic medical record data are organized into JSON format, in which various attributes of diseases are presented in such key-value pairs. As we can see from Figure 1, the processed JSON data is row data with disease as the main body, and there are different key-value pairs in the row data, so it is easy to realize edge relationships programmatically.

By naming medical entities and creating edge relationships, the following knowledge graph can be formed. Figure 3 shows a screenshot of part of the disease knowledge graph, and Figure 4 shows a screenshot of edge relationships.

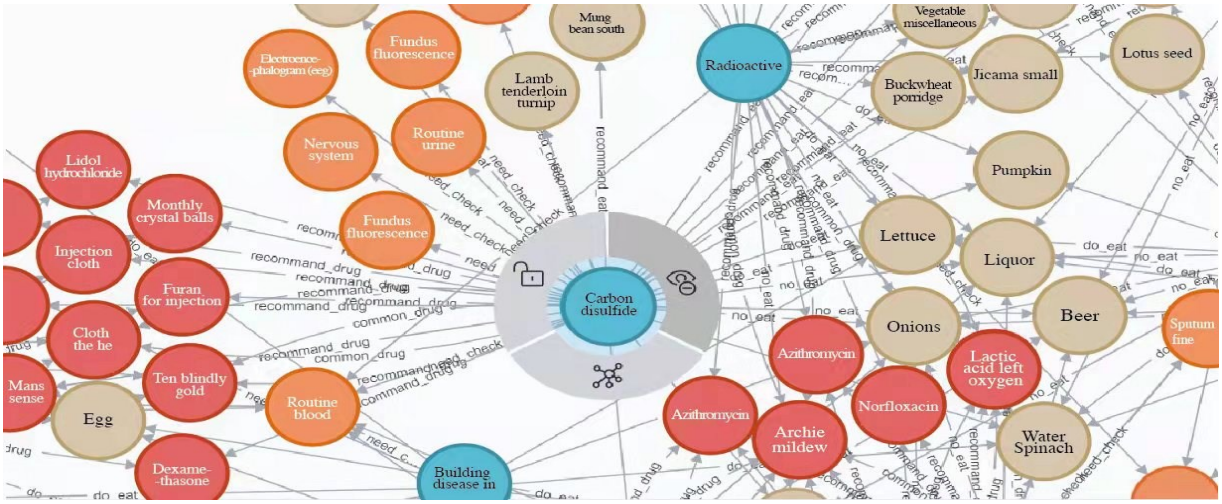

Figure 3. Screenshot of knowledge graph.

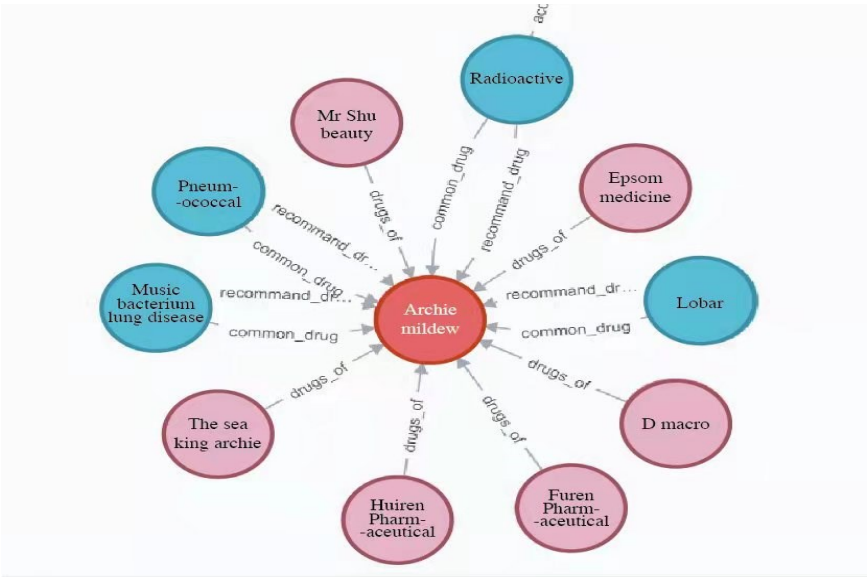

Figure 4. Screenshot of the edge relationship.

5. INTELLIGENT QUESTION AND ANSWER DESIGN

Intelligent question answering is to extract the key information from the user input statements, and then retrieve the key information in the graph database and return the retrieval results. Retrieval semantic matching is the most important step in intelligent question answering design. The purpose of this step is to match user input statements with questions in the question library[10]. In the medical intelligent question answering system of this paper, semantic matching method can also be used to retrieve the graph database and get the answers we need. Retrieval in the neo4j graph database requires named entity recognition for the questions entered by the user, which is consistent with the node real identification method in the previous step. After named entity recognition, it can locate in the knowledge graph, find the answer according to the edge relation, and return the result.

5.1 Semantic matching model based on vector space model

In the question answering system of this paper, after the nodes are determined, the relationship between nodes and edges can be formed into question sets. The answer is found by matching the input statement to the set of questions. Because different users express the same problem, the system needs to normalize different statements of the same problem, that is,

to mine the hidden information of user input statements for the convenience of graph data query. Based on this, TextRank algorithm is used to optimize the semantics of the input problem, and an improved spatial vector model is used to calculate the semantic similarity between user input statements and graph database nodes[11].

(1) Semantic optimization

Due to different users' different expressions of the same question, such as the question of what medicine to take for a cold, the intelligent question and answer input may be described as "what medicine to take for a cold and fever?" or "What medicine should I take for a cold and fever? I feel like I have a fever. I think I have a cold. What should I take for it?". In fact, these three groups of questions describe the same questions, so we need to first optimize the semantics of questions in the question answering system.

Semantic optimization first needs to compress the statement, its purpose is to get rid of irrelevant and meaningless words, to get rid of the noise in the statement. At present, the main methods of long difficult sentence compression are extraction and neural network. In this paper, TextRank algorithm is used to compress long and difficult sentences.

TextRank algorithm is a graph-based sorting algorithm for text. Its basic idea comes from Google's PageRank algorithm, through the text is divided into several component units (words, sentences) and the establishment of graph model, the use of voting mechanism to rank the important components of the text, only the use of single document itself information can achieve key words extraction, abstracting. Unlike models such as LDA and HMM, TextRank does not require multiple documents to be learned and trained in advance, and is widely used because of its simplicity and effectiveness[12].

(2) Semantic matching based on vector space model

Semantically optimized question input statements need to match as much as possible with nodes in the graph database to find the answer. This paper uses the improved vector space model to match semantics with node data, and its algorithm is as follows:

Definition 1 (defining word frequency), $Tf(i,j) = \frac{n_{i,j}}{\sum_k n_{k,j}}$, where n denotes the number of words in a question.

Definition 2 (definition of inverse word frequency), $Idf = log\frac{|D|}{|\{j:t_i \in d_j\}|}$, where D represents the total number of words.

Definition 3 (the calculation of the similarity), W(I) =Tf * Idf, where the higher the similarity, the closer the relation between the statement and the node, indicating that the node matches the input statement. If the value is less than 1, the node is not matched.

6. SYSTEM SIMULATION

The development language of the system is Python, and the integrated development environment is Eclipse 2019. The system development steps are as follows:

(1) Firstly, the theme crawler algorithm of PLSA model is used to crawler vertical medical websites. Combined with the electronic medical record data of the Medical Data Research Institute of Chongqing Medical University, 8807 medical record data were obtained after sorting, as shown in Figure 1.

(2) Neo4j was used to construct the knowledge map, as shown in Figures 2 and 3.

(3) The TextRank algorithm is used to optimize the semantics of the input statements, and the vector space model is used to match the semantics of the input statements. Finally, the answer is formed by graph retrieval in the knowledge graph.

The system simulation results are shown in Figure 5 below:

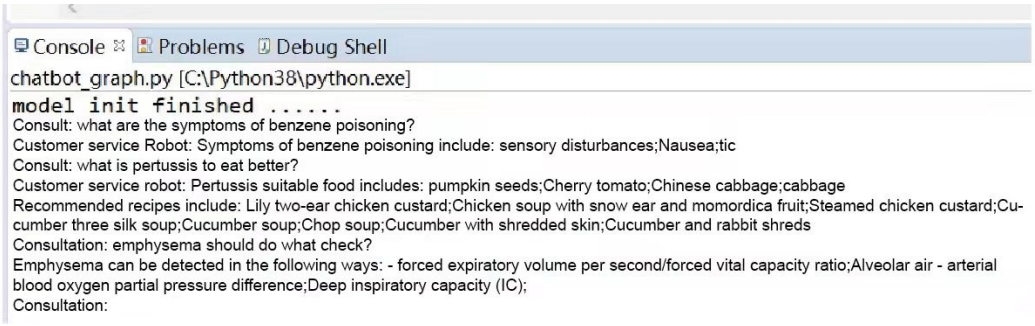

Figure 5. System simulation results.

7. CONCLUSION

In view of the fact that there is no efficient and reliable source of medical knowledge for ordinary people and no effective clinical practice for junior medical students, this paper uses PLSA model to crawl medical knowledge of related diseases from vertical medical websites, combines real electronic medical record resources, and uses NEO4j graph database to store data and form knowledge map of related diseases. TextRanK and vector space model are used to optimize user input questions and semantic matching.

Subsequent research will focus on the impact of other web pages, such as discussion forums, on the theme, so as to make crawling content more reasonable.

ACKNOWLEDGEMENTS

Key project of Intelligent Medicine research of Chongqing Medical University ZHYX202003, standardized training of resident doctors based on process management Development and application of training information platform CSTC2019JSCX-MSXMX0100.

REFERENCES

[1] Yi, L. T., Dong, C., Niu, Z. Y. and Liu, S. J., “Research and implementation of question answering system in intelligent medical field,” Information Record Materials 22(5), 232-234 (2021).
[2] Wu, L. and Wang, Y. B., “Research on topic crawler based on semantic similarity aggregation,” Journal of Communication University of China 25(1), 28-31 (2018).
[3] Wang, J. Z. and Qiu, T. X., “Topic focused web crawler based on improved tf-idf algorithm,” Journal of Computer Applications 35(10), 16-21 (2015).
[4] Miao, Y., Liu, X. Y., Jin, J. N. and Li, K. X., “Medical data crawling and analysis processing based on Python,” Computer Application Technology 4(16), 56-58 (2019).
[5] Xie, R. R., Xu, H., Zheng, S. W. and Ma, G., “Medical data crawling and analysis processing based on Python,” Computer Application Technology 38(6), 439-453 (2021).
[6] He, Y. and Zhang, N., “Design and implementation of intelligent medical question answering system,” China Medical Equipment 36(09), 100-108 (2011).
[7] Wang, J., Liang, H., Fan, W., Chen, G., Sun, F. and Lin, K., “Design and implementation of intelligent question answering system based on Chinese medical knowledge graph,” China Electronic Equipment 16(2), 54-58 (2021).
[8] Wei, Z., Zhang, S. and Wang, J., “Technical Implementation of knowledge graph question answering system,” Software Engineering 24(2), 38-43 (2021).
[9] WHO {Chronic diseases [EB/OL] WHO 2016-09-21 http://www.who.int/topics/chronic_ diseases/en/.
[10] Ren, S. Q. and Aung, K. M. M., “PPDS: Privacy preserved data sharing scheme for cloud storage,” International Journal of Advancements in Computing Technology 4(16), 493-499 (2012).
[11] Zeng, H. and Wang, J., “Clustering of electronic medical records based on association relationship,” Chinese Journal of Medical Library and Information 9(2), 73-79 (2018).
[12] Wang, J. and Zeng, H., “Design and implementation of intelligent medical system for chronic diseases American Journal of Computer Science and Technology 3(4), 86-91 (2020).