Fuzzy Systems and Data Mining X
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA241417

Knowledge Graph-Powered Question Answering System with Random Forest-Assisted Diagnosis for Elderly Healthcare

Yunlei Meng[a] and Rui Dai[a*]
[a] *School of Mathematics and Statistics, Beijing Technology and Business University, Beijing, China*

Abstract. As a knowledge representation tool, knowledge graph (KG) has been widely used. In this study, a question answering (Q&A) system for geriatric diseases based on knowledge graph was constructed to help the elderly obtain medical information. Initially, a total of 6,376 disease data items were collected and analyzed in order to identify the characteristics of these diseases. Then, the KG is constructed by Neo4j graph database. The establishment of Q&A system starts from semantic recognition. The Aho-Corasick (AC) automaton is utilized to filter user input questions. The Cypher language is employed for querying graph databases, and the obtained results are then imported into predefined templates for output. The accuracy of our system for different categories of questions is 87% and 94%, respectively. Finally, the random forest model is introduced to solve the problem of disease diagnosis. The feature variables were vectorized using TF-IDF model and the target variables were vectorized using one-hot model. In general, we introduce a novel Knowledge graph-driven Q&A system. Provide a new tool for health management of the elderly population. And the construction of Q&A system will promote the development of smart medicine and solves the health confusion of the elderly.

Keywords. medical knowledge graph, question answering system, random forest

1. Introduction

1.1. Background and Motivation

The Knowledge graph has received a lot of attention since it was proposed by Google in 2012. Knowledge graphs present knowledge in the form of triples that are easy to discover and reason about[1]. Knowledge graph-driven question answering systems appear in various vertical fields. For example, the knowledge map is constructed with the relationship between epidemic cases and personnel as the data source to visually and clearly present the complex information about the patients and their relationships[2]. In the field of insurance, entity extraction technology is used to build the insurance knowledge graph to realize the recommendation of life insurance knowledge and the query of life insurance product attributes[3].

[*] Corresponding Author: Rui Dai, Beijing Technology and Business University, Beijing, China; E-mail addresses: dr.dairui@hotmail.com

In addition, in the field of medicine, medical question answering systems utilize knowledge graph to store medical knowledge and provide accurate answers[4]. By integrating medical literature, clinical guidelines and expert knowledge, the medical knowledge graph containing various diseases, symptoms, drugs and other information is constructed. Provide basic data and knowledge support for question answering system. However, the current research mainly focuses on a certain type of diseases such as cardiovascular and cerebrovascular diseases[5]. Knowledge graph based on data from specific populations are still lacking. The elderly population has its particularity, even if some intelligent medical systems has been available, but there are still problems such as program, lack of personalization, real-time lag and so on[6][7]. Based on these problems, this study turns the face to knowledge graph-driven Q&A system to the elderly to provide them with real-time and accurate medical services.

1.2. Contributions

- Previous studies on knowledge graph were mostly based on a certain industry, but this study extends it to a certain group of people.
- This study sorted out an effective standard to classify diseases encountered by the elderly in real life.
- A large number of previous studies focused on the innovation of a certain method in the process of system construction, while this study completed the entire process from data to system construction.

1.3. Structure

Firstly, we collect data and analyze it with visual tools. Secondly, we construct the knowledge graph and build the question answering system based on it. Finally, the various parts of the system and their functions are as follows:

- Knowledge graph: Data is stored in Neo4j graph database and queried through Cypher.
- Semantic recognition module: Semantic recognition of user input questions, the specific process is AC automata word filtering[8].
- Classification model: The TF-IDF model and one-hot model were respectively used to vectorize the variables and then random forest classification model was trained.
- Answer retrieval and output module: Build Cypher statements to retrieve answers from KG and add them to the template for output.

2. Knowledge Graph

2.1. Data Acquisition and Analysis

The data comes from a project called QA Based on Medical Knowledge Graph on Github. The project data are uploaded by professional scientists and cover a wide range of disease information. The data of this project came from a disease-focused vertical medical website, which crawled data via python containing 44,000 knowledge entities and 300,000 entity relationships[9].

Data processing phase includes data cleansing, error filtering, entity identification and screening (e.g., disease, symptom, department), and attribute selection (e.g., disease description, symptom characteristics). The first step is to remove errors or missing information, such as large chunks of text or illogical text. The second step is the completion of information, complete the missing information by the Internet. The third step is the screening of the core entity that is the screening of diseases in the elderly[10]. The elderly were defined by age (≥ 65 years) and disease characteristics (e.g., cardiovascular disease, eye disease). It is classified according to its high incidence, impact on specific populations, impact on quality of life, and feasibility of prevention and treatment[11]. Through data collection, screening and classification, a total of 6,376 diseases were identified. Finally, for the screening of attribute data, the system in this study is not to provide a large number of disorganized information like a search engine, but to screen more useful information for display. Such as symptoms, appropriate food, avoid food, drugs, inspection departments and other attributes[12].

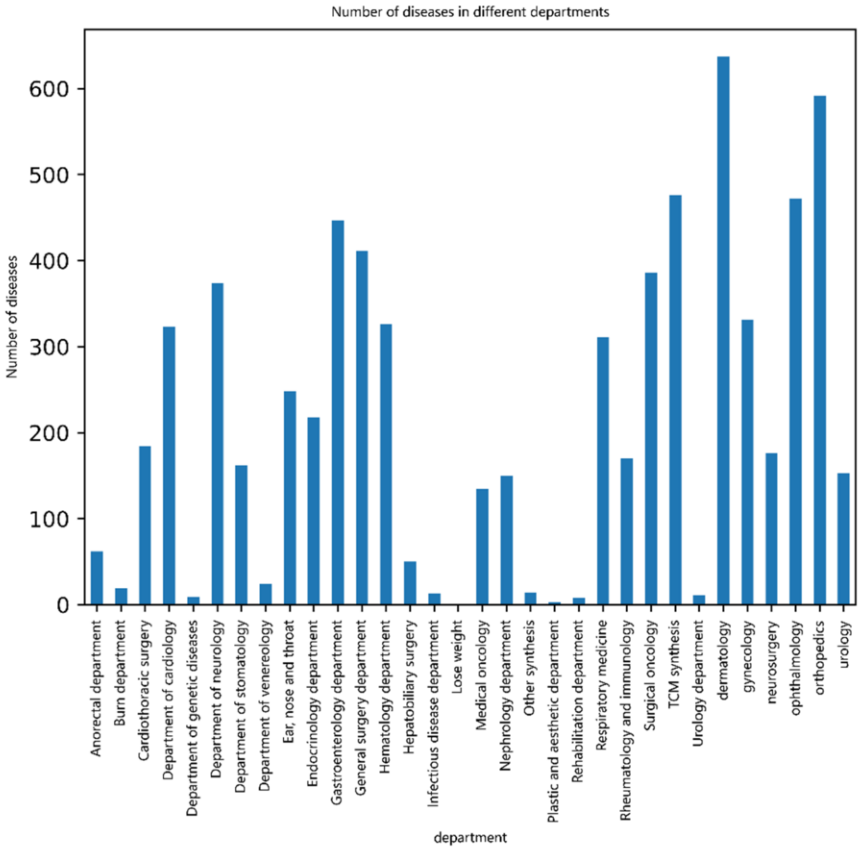

Figure 1. The figure shows the number of common diseases of the elderly in each department.

There is a lot of information in the disease data of the elderly. In order to optimize the allocation of medical resources and improve the efficiency of diagnosis and treatment, this study drew the distribution map of geriatric disease departments and analyzed their clustering in each department. Figure 1 shows the result. The diseases of the elderly are mainly concentrated in the departments of cardiology, respiratory medicine, neurology

and so on. Cluster analysis is helpful for medical institutions to optimize resource allocation, build professional medical teams, and improve the quality of diagnosis and treatment of senile diseases.

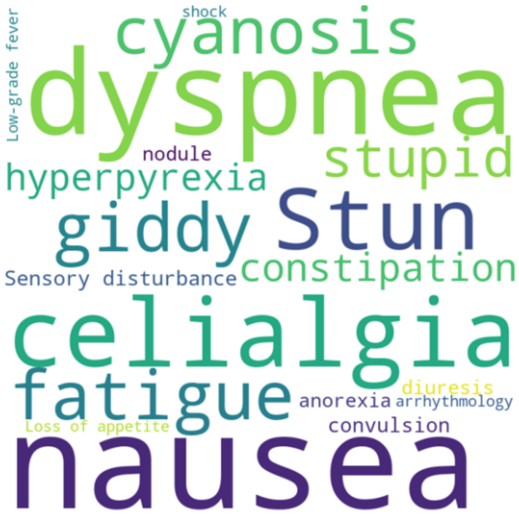

Figure 2. The figure shows the common symptoms of diseases in the elderly.

Data on disease symptoms in the elderly are also worth studying. Figure 2 shows that fatigue and difficulty breathing are common symptoms of disease in the elderly and are essential for early recognition. These symptoms are associated with multiple underlying diseases, which complicate diagnosis and treatment. It is necessary to consider a variety of factors and develop an individualized treatment plan to improve the diagnostic accuracy.

2.2. Knowledge Graph Construction

After statistical analysis of the data, it is necessary to process the triplet data suitable for constructing KG. This includes steps such as cleaning, standardization, normalization, and conversion of data. Clean data to ensure quality and accuracy, standardize uniform data formats, and transform data into triples to build knowledge graph[13]. In the E-R-E(Entity-Relation-Entity) pattern, entities represent concepts and relationships represent associations between entities. For example, in the film knowledge graph, entities are films, actors, directors, etc., and relationships are participating in and directing, etc. Complex association networks are established to explore entity relationships. In E-A-V(Entity-Attribute-Value) mode, an entity is a concrete object, attributes describe its characteristics, and attribute values represent specific values[14]. Figure 3 shows the basic units of the knowledge graph. For example, in the medical knowledge graph, the entity is disease, drug, symptom, etc., the attribute is name, dose, description, etc., and the attribute value corresponds to the specific value. By associating entities with attributes, attribute information can be stored and retrieved.

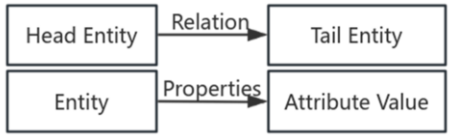

Figure 3. The figure shows the basic units of the knowledge graph.

The knowledge graph involved in the research is generally composed of interconnected entities and their properties[15]. A knowledge graph is a relational database that stores data in the form of graph. At present, most knowledge graphs involved in the research are made based on Neo4j graph database. Neo4j is a database management system based on graph database model, which stores and processes data in the form of graph and provides powerful graph database functions. Neo4j graph database is a powerful database system with many important features. First, it provides ACID (Atomicity, Consistency, Isolation, and Persistence) transaction support, ensuring data reliability and consistency. Neo4j's data model includes:

Node: indicates a node, including labels, attributes, attribute values, and ids.

Relation: Relation, including labels, attributes, attribute values, and ids.

Cypher is Neo4j's query language for graph data retrieval and manipulation, has been widely used because of its simple and intuitive syntax. This section uses Python's py2neo library and Neo4j graph database to build a medical KG.

The specific construction process is as follows: First, register the account of Neo4j graph database and connect with Python. Label the data after it is imported. Secondly, Cypher statement is used to construct the relationship between each label data. Finally, the database is retrieved, and the data are respectively popped into the knowledge graph model.

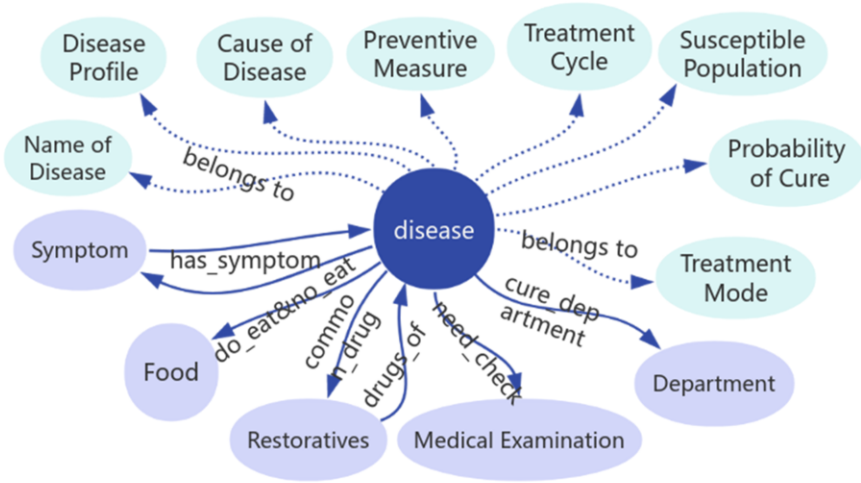

Figure 4. The figure shows all the entities, attributes, and relationships covered in this article. With the disease entity as the core, dotted arrows point to its attributes and solid arrows point to its associated entities.

Figure 4 shows the basic units of the knowledge graph in this study. Neo4j graph database supports the direct use of Cypher statements for data query and presentation. Figure 5 shows a visual example of the query results.

Figure 5. This figure shows the knowledge graph. The graph matches all nodes of the disease type named elderly diabetes and finds all nodes related to it. The right side of the database will provide the entity type, relationship type, and number of all retrieved contents. When you select an entity, the right section of the database displays the appropriate properties for that entity.

3. Q&A System

3.1. Q&A System Construction

The first step to construct a Q&A system is to judge the semantic meaning of the input[16]. In this study, the AC tree model is used to filter the problem, implemented through python's ahocorasick library, which is a string matching algorithm. The following details the specific process of building an AC automaton, there are two steps in simple terms:

1. Basic Trie structure: String all patterns together to form a Trie.
2. KMP: Construct failure Pointers for all nodes in the Trie tree. Then the AC automata can be used to implement the multi-pattern matching task. Figure 6 illustrates this process.

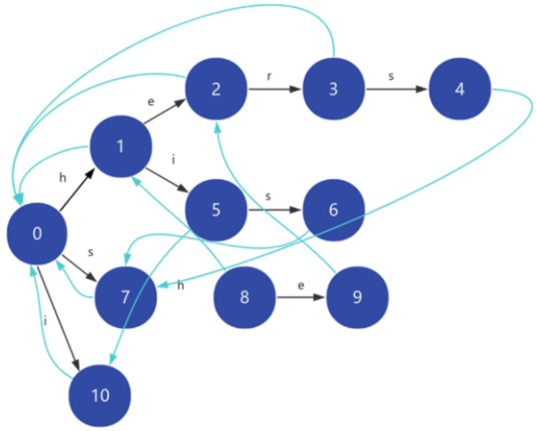

Figure 6. The figure shows how an AC automaton works. The following describes the pointing logic of failure Pointers, take node 6 as an example to illustrate the construction process of fail pointer: Find the parent node 5 of 6, fail[5]=10. But node 10 doesn't have an edge with the letter s; Continue to skip to the fail pointer of 10, where fail[10]=0. It is found that node 0 has an edge connected by the letter s, pointing to node 7; So fail[6]=7.

The filtering process is as follows: The first step is to construct a trie to filter the question word by word. The feature words required for trie are the data required for knowledge graph construction. Failure Pointers make filtering very efficient. The second step is the establishment of the question vocabulary, which requires manual input of various question words (such as what symptoms you have and what you can eat). The third step is to identify the keywords and question words in the question (some questions do not contain question words) and then pop them into the corresponding empty list. After successfully identifying the entity, intention and category in the question, type recognition is carried out[17]. For cases where the intention is not accurately identified, the disease description are returned if the entity is a disease, and the symptoms are returned if the entity is a symptom. Finally, after successfully identifying the type of the question, the corresponding Cypher statement is called and the query is executed. Table 1 shows an example of a cypher statement that performs answer retrieval after recognizing the question type.

Table 1. Example of a query statement.

Question Type	Sample Query Statement
Inquire about the cause of the illness	sql = ["MATCH (m: Disease) where m.name = '{0}' return m.name, m.cause".format(i) for i in entities]
Query the cure probability of the disease	sql = ["MATCH (m: Disease) where m.name = '{0}' return m.name, m.cured_prob".format(i) for i in entities]
Known food for disease	sql = ["MATCH (m: Disease)-[r:no_eat]->(n: Food) where n.name = '{0}' return m.name, r.name, n.name".format(i) for i in entities]

3.2. Random Forest

In the Q&A system based on KG, the random forest model is introduced to complete the classification task in order to improve the efficiency of disease diagnosis. Random forests are predicted by a combined decision tree model[18][19]. When we diagnose diseases in hospitals, in addition to using professional instruments for detection, doctors often pay attention to the patient's age, symptoms, past medical history and other factors. In this study, due to the limited data and considering the usefulness of variables, age and symptoms were selected as the characteristic variables for disease diagnosis. The age affects disease risk, and the symptoms are diagnostic clues. Because of the uniqueness of the disease data, a disease in most cases has only one professional name. Disease variables are vectorized by One-Hot Encoding, converting text data into vectors with the value of 1 for the location of the disease and 0 for the rest. The characteristic variables, age and symptoms, are disordered categorical variables (age is regarded as unordered here), so only the effective vectorization method for short text data is considered. TF-IDF model is the most widely used model to deal with this problem. The feature variables are vectorized using the TF-IDF model.

The numerical data is then trained and predicted using a random forest model, which consists of multiple decision trees that make decisions by voting or averaging the predicted results[20]. Table 2 illustrates these special visits and the target variables. Random forest is suitable for processing high-dimensional features and big data, and has good generalization ability and anti-overfitting ability, which is suitable for disease diagnosis. During the training of random forest model, the feature sparse matrix obtained

by TF-IDF model was taken as the feature variable, the disease vector was taken as the target variable, the proportion of test sets were 20% and 10% respectively, and the random number seed was 42.

Table 2. Examples of variables used by classification models.

Disease	Symptom_Features	Susceptible_Features
Chordoma	Sciatica	It occurs in the middle and old age of 50 ~ 60 years old
Squamous Cell Carcinoma	Epidermal Keratosis	It is most common in men over the age of 50

3.3. Results

In collecting real-life problems, we first categorize them. One is to diagnose the disease based on symptoms, and the other is to ask for various information about the disease[21].

Question Type 1 Example Question: What is the disease that causes sudden loss of memory and disorientation in the elderly?

Answer Sample Answer: Alzheimer's Disease

Question Type 2 Sample Question: What symptoms may indicate that you have heart disease?

Answer Sample Answer: Angina pectoris, shortness of breath, weakness, palpitations, etc

Once the question answer data is obtained, the question and answer are processed and identified through Python. When evaluating the accuracy of the question answering system, the physical match between the answer and the standard answer is taken as the criterion. Due to the text differences between the standard answer and the answer from Q&A system. The standard answer contains only the target word, while the system outputs the answer as a complete sentence. Only accuracy is selected. If the standard answer entity appears in the output answer, it is considered correct; Otherwise, it is an error. This approach is concise and intuitive, but it may overlook the nuances and grammatical correctness of the answers. Suitable for specific question answering tasks, other types of systems may require more evaluation indicators.

Table 3. Results of accuracy evaluation of question answering system.

Question Type	Total Number of Questions	Correct Answers	Accuracy Rate
Type 1	114	99	0.87
Type 2	115	108	0.94

The results of the evaluation in Table 3 show that the system responded well to 229 questions. Of the 114 questions asked about the disease, 87 percent were accurate; Of the 115 questions that asked about attributes, 94 percent were accurate. The system can provide users with accurate information, meet the query requirements of diseases and attributes, and show high reliability and usefulness. We also conducted research on incorrect questions. Research has found that a large portion of the errors from inaccuracies in the wording.

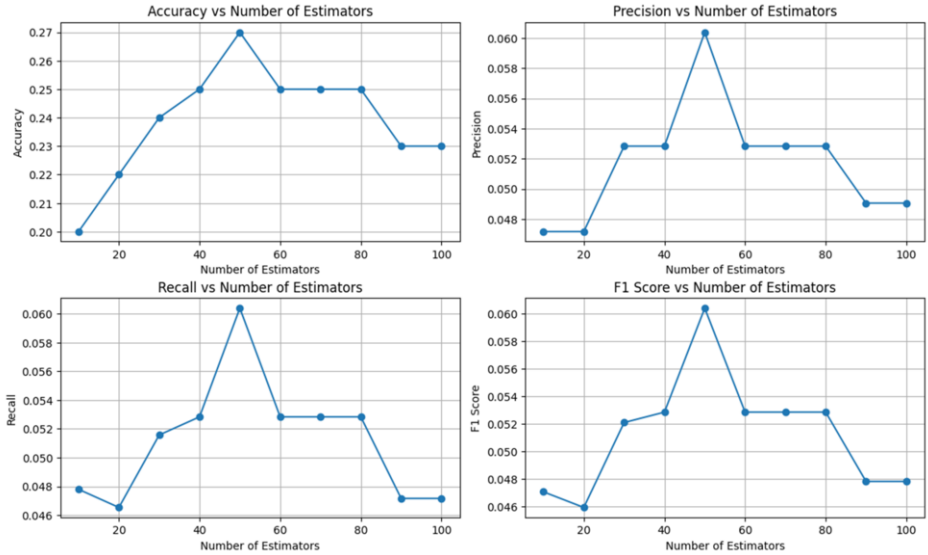

Figure 7. This figure shows the effect of the model with different number of decision trees.

The parameter setting of the random forest model has a great influence on the results. Figure 7 shows the effect of the model with different number of decision trees. We can see that when the number of decision trees is 50, the four indicators all present good results.

Table 4. Evaluation results of random forest model.

Test-Size	0.1	0.2
accuracy	0.62	0.59
precision	0.91	0.91
recall	0.62	0.59
f1	0.73	0.71

Table 4 show that the model has a low accuracy and recall in disease prediction, but a high precision. The F1 value is medium, indicating that the model has certain forecasting ability, but it still needs to be improved. We will continue to optimize the model parameters and increase the training sample size to improve the model performance. Figure 8 shows the interface of the Q&A system constructed in this study.

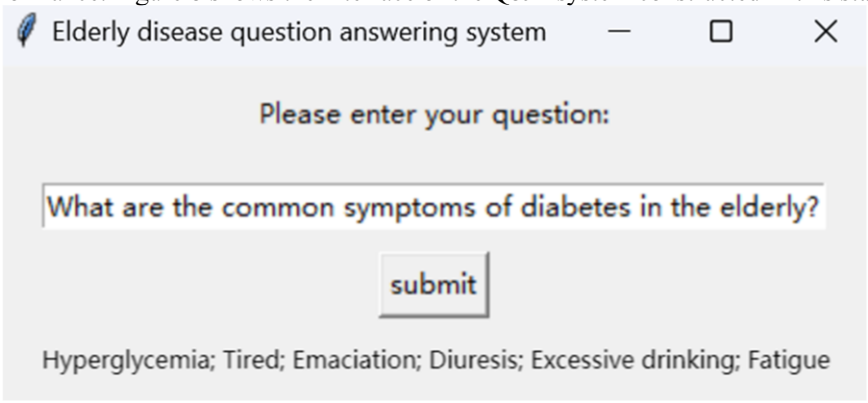

Figure 8. The figure shows the page of the question answering system. Enter the question in the input box and click Submit to get the answer.

4. Conclusion and Discussion

4.1. Summary

With the increase of the elderly population, medical resources are becoming increasingly scarce. We propose a knowledge graph-driven system to solve this problem. Based on Neo4j graph database, the Q&A system of elderly diseases is constructed. AC tree and Cypher query were used to achieve efficient question-answering, and random forest model was introduced to assist diagnosis. Finally, the results verify the good performance of the question answering system.

4.2. Limitation

Firstly, the data involved in this study is limited, and the data expansion module has not been built. Due to the limitation of equipment performance, the amount of data used in this study does not meet the standard of building a large-scale question answering system. Secondly, this study does not propose some effective algorithm improvement schemes. Finally, this study only completed a single round of questions and answers, and could not conduct multiple related questions and answers.

4.3. Future Work

In the future, we will continue to pay attention to the users in order to timely update the system. In addition, with the continuous updating of technology, we also expect to update the semantic recognition and vectorization tools used in this research to obtain higher accuracy. Finally, the issue of data privacy should also be paid attention to, and the data involved in this system should have a good protection to avoid information disclosure.

References

[1] Wu X, Duan J, Pan Y, Li M. Medical knowledge graph: Data sources, construction, reasoning, and applications. Big Data Mining and Analytics. 2023, 6(2): 201-217, doi: 10.26599/BDMA.2022. 9020021

[2] Ji S, Pan S, Cambria E, Marttinen P, Philip SY. A survey on knowledge graphs: Representation, acquisition, and applications. IEEE transactions on neural networks and learning systems; 2021 Apr 26; c2021. p. 494-514, doi: 10.1109/TNNLS.2021.3070843

[3] Zou X. A survey on application of knowledge graph. Journal of Physics: Conference Series. 2020, 1487(1): 012016, doi: 10.1088/1742-6596/1487/1/012016

[4] Ren Y, Shi Y, Zhang K, Chen Z, Yan Z. Medical treatment migration prediction based on GCN via medical insurance data. IEEE Journal of Biomedical and Health Informatics; 2020 Jul 10; c2020. p. 2516-2522, doi: 10.1109/JBHI.2020.3008493

[5] Li J, Liu J, Liu X, Yang F, Xu Y. A medical insurance fraud detection model with knowledge graph and machine learning. International Conference on Computer Application and Information Security (ICCAIS 2021); 2021; Wuhan, China: SPIE Press. c2022. p. 12260: 531-540, doi: 10.1117/12.2637418

[6] Bekoulis G, Deleu J, Demeester T, Develder C. Joint entity recognition and relation extraction as a multi-head selection problem. Expert Systems with Applications. 2018, 114: 34-45, doi: 10.1016/j.eswa.2018. 07.032

[7] Chandak P, Huang K, Zitnik M. Building a knowledge graph to enable precision medicine. Scientific Data. 2023, 10(1): 67, doi: 10.1038/s41597-023-01960-3

[8] Yang R, Ye Q, Cheng C, Zhang S, Lan Y, Zou J. Decision-making system for the diagnosis of syndrome based on traditional Chinese medicine knowledge graph. Evidence-Based Complementary and Alternative Medicine. 2022, 2022, doi: 10.1155/2022/8693937

[9] Wu X, Zhu X, Wu GQ, Ding W. Data mining with big data. IEEE transactions on knowledge and data engineering; 2013 Jun 26; c2013. p. 97-107, doi: 10.1109/TKDE.2013.109
[10] He B, Guan Y, Dai R. Classifying medical relations in clinical text via convolutional neural networks. Artificial intelligence in medicine. 2019, 93: 43-49, doi: 10.1016/j.artmed.2018.05.001
[11] Petkovic D, Altman R, Wong M, Vigil A. Improving the explainability of Random Forest classifier–user centered approach. In Pacific symposium on biocomputing 2018: proceedings of the pacific symposium. 204-215, doi: 10.1142/9789813235533-0019
[12] Guo Q, Zhuang F, Qin C, Zhu H, Xie X, Xiong H, He Q. A survey on knowledge graph-based recommender systems. IEEE Transactions on Knowledge and Data Engineering; 2020 Oct 07; c2020. p. 34(8): 3549-3568, doi: 10.1109/TKDE.2020.3028705
[13] Li L, Wang P, Yan J, Wang Y, Li S, Jiang J, Sun Z, Tang B, Chang T, Wang S, Liu Y. Real-world data medical knowledge graph: construction and applications. Artificial intelligence in medicine. 2020, 103: 101817, doi: 10.1016/j.artmed.2020.101817
[14] Gong F, Wang M, Wang H, Wang S, Liu M. SMR: medical knowledge graph embedding for safe medicine recommendation. Big Data Research. 2021, 23: 100174, doi: 10.1016/j.bdr.2020.100174
[15] Zhao C, Jiang J, Xu Z, Guan Y. A study of EMR-based medical knowledge network and its applications. Computer methods and programs in biomedicine. 2017, 143: 13-23, doi: 10.1016/j.cmpb. 2017.02.016
[16] Yin Y, Zhang L, Wang Y, Wang M, Zhang Q, Li G. Question answering system based on knowledge graph in traditional Chinese medicine diagnosis and treatment of viral hepatitis. BioMed research international. 2022, 2022, doi: 10.1155/2022/7139904
[17] Chai X. Diagnosis method of thyroid disease combining knowledge graph and deep learning. IEEE Access; 2020 Aug 14; c2020. p. 149787-149795, doi: 10.1109/ACCESS.2020.3016676
[18] Lei Z, Sun Y, Nanehkaran YA, Yang S, Islam MS, Lei H, Zhang D. A novel data-driven robust framework based on machine learning and knowledge graph for disease classification. Future Generation Computer Systems. 2020, 102: 534-548, doi: 10.1016/j.future.2019.08.030
[19] Teng F, Yang W, Chen L, Huang LF, Xu Q. Explainable prediction of medical codes with knowledge graphs. Frontiers in bioengineering and biotechnology. 2020, 8: 867, doi: 10.3389/fbioe.2020.00867
[20] Li Z, Geng P, Cao S, Hu B. Few-shot knowledge graph completion based on data enhancement. International Conference on Bioinformatics and Biomedicine (BIBM); 2022 Dec 06-08; Las Vegas, NV, USA: IEEE Press; p. 1607-1611, doi: 10.1109/BIBM55620.2022.9995024
[21] Tao X, Pham T, Zhang J, Yong J, Goh WP, Zhang W, Cai Y. Mining health knowledge graph for health risk prediction. World Wide Web. 2020, 23: 2341-2362, doi: 10.1007/s11280-020-00810-1