KnowHealth: A Knowledge Graph based Question-Answer Platform for Elderly People

Chunshan Li[1], Shaoshi Hang[2], Xin Hu[3], Dianhui Chu[4], Hongzhen Zheng[5]
{lics@hit.edu.cn[1], hangss@hit.edu.cn [2], Hux@hit.edu.cn[3],
chudh@hit.edu.cn[4], zhenghz@hit.edu.cn[5]}

Affiliation, School of Computer Science and Technology
Harbin Institute of Technology (Weihai)[1,2,3,4,5]

Abstract. Nowadays, population aging has become a prominent problem all over the world, which brings new challenge to manage multi-type and multi-relation health data from elderly. Existing approaches store those complex data in traditional relational database, which lack support of relation retrieval and is hard to answer elderly's semantic query. In this paper, we design and implement a knowledge graph based question-answer platform (KnowHealth) to manage those health data, which can help the elderly know their health condition better, Specifically, we propose an ontology definition which can describes the entities and relations in aged disease domain. Based on the ontology, we crawl and extract health related information from different input sources. Then, entities and relation extraction methods can be used to construct a knowledge graph. Finally, we completed a historical behavior driven question-answering platform to serve query for elderly. By analyzing and extending intention of questions, answer can be retrieved and reasoned more accurately.

Keywords: Knowledge Graph, Aged Diseases Information Management, Knowledge Extraction, Question Answering Platform.

1 Introduction

With the development of healthcare data-aware and the booming of Web 3.0, a great deal of knowledge of the elderly health have accumulated on the Internet [1], such as the encyclopedia platform, the Website about health care and so on. Existing approaches [2], [3]to handle those data has two major limitations. (1) A large amount of health data exists on the Web and the data updates frequently [4]. Consequently, manually process the data is inapplicable. (2) Previous works [5], [6] store those data in a traditional relational database through a structured table. This storage mode can't manage the various relations among health data and it is detrimental to the data analysis. (3) There are few uniform definitions to describe entities and relations in health care domain with different types and attributes [7]. Therefore, it is an important challenge to design good storage and analysis approaches to manage relevant health care data and provide better data service in health care domain.

In recent years, knowledge graph [8], [9] has become a hot research trend due to its logical structure which has better knowledge managing ability than traditional storage methods. Knowledge graph can construct a huge network with concepts (nodes) and the relations (edges) among concepts. It can links scattered knowledge to form a massive

MOBIMEDIA 2019, June 29-30, Weihai, People's Republic of China

DOI 10.4108/eai.29-6-2019.2282866

knowledge network [10]. Knowledge graph are very useful in knowledge retrieval, question-answering, knowledge recommendation and other applications. Knowledge graph provides an ideal technical means to solve the problem of “knowledge island”, and it is helpful to realize the integration of knowledge resources and enhance the knowledge service ability.

The Web search communities first apply knowledge graph to enhance the query performance. Google improved the efficiency and accuracy of the search by employing knowledge graph technology in 2012. Baidu, Sogou, Microsoft and other giants had started to follow up the domain of knowledge graph, e.g. Baidu Zhixin, Sogou Zhilifang, and Microsoft Renlifang. Palomares et al. \cite{palomares2016wikipedia} built Wikipedia's knowledge graph with DeepDive. Lin et al. \cite{lin2015learning} studied and established of the relationship between entities by embedding the knowledge graph into the model. Ganggao Zhu et al. \cite{zhu2017sematch} proposed to take the similarity semantics into the knowledge graph to improve search and recommendation. Nevertheless, these knowledge graphs are all used in the open domain which are often not accurate enough facing professional issues, and they all have quite a few limitations.

Knowledge graph also has been used to address problems and mine the potential and valuable information in specific industries. Filtz [1] proposed the construction of a knowledge graph for legal knowledge, both lawyers and ordinary people can obtain legal knowledge from the knowledge graph. Yang et al. [14] suggested solving the technical issue of Microsoft products by establishing knowledge graph and using index-based random walk method. Zhu et al. [15] constructed the application which took the intelligent learning techniques of knowledge graph into geology. Yu et al. [16] presented a method of constructing knowledge graph for traditional Chinese medicine which can be convenient to retrieve data and recommend some advice. Although these knowledge graphs have played a role in data analysis for the domain, there are few cases in which the knowledge graph is applied to the health care for the elderly population.

Faced with the above issues, this paper designs a knowledge graph (KnowHealth) for managing the data in elderly's health care domain, and erects a senior question answer platform based on the knowledge graph. The main contributions of the platform are as follows:

(1) We automatically crawl and extract health related information from different input sources, and organize the data by cleaning noisy, splitting the data into words, extracting entity, mining the relationship of entities and other steps. We construct a knowledge graph (KnowHealth) to manage the processed concept and relations in the graph database.

(2) Talking with medical expert, we propose a ontology definition in health knowledge graph in which the entities are classified into 7 categories and the relations are classified into 14 categories.

(3) We design and implement a historical behavior driven question-answering platform. The QA platform can record the key elderly behaviors to analyze and dig potential service requirement of users.

The remaining of the paper is organized as follows. Section 2 reviews recent related works. The construction of knowledge graph is illustrated in Section 3. A question and answering platform are reported in Section 4 and the paper is concluded in Section 5.

2 Construction of Knowledge Graph

In this section, the main process of KnowHealth will be introduced. And then the detailed steps to construct KnowHealth and a historical behavior driven application will be presented, which consists of three major steps: query classification, knowledge retrieval and historical behavior driven reasoning.

Figure 1 shows the overview of our whole platform, the construction process of knowledge graph mainly includes acquisition of data, extraction of knowledge and storage of knowledge. Specifically, our system first sets up a local information library which can download text data from Web data or online encyclopedia according to the health domain and theme. At the same time, the system collates the heterogeneous data under the uniform semantic format. Then the system extracts the entities and relationships among entities under an ontology definition. Latter, the entities and relations are stored in a graph database to form the knowledge graph. The upper part of Figure 1 shows the question-answer platform based on knowledge graph. When the user want to query on the KnowHealth, our platform will identify the naming entity by the ontology library, classify the input text, and reason the knowledge in accordance with the question content and the historical behaviors of users. Finally, users can obtain the reasoning result from knowledge graph.

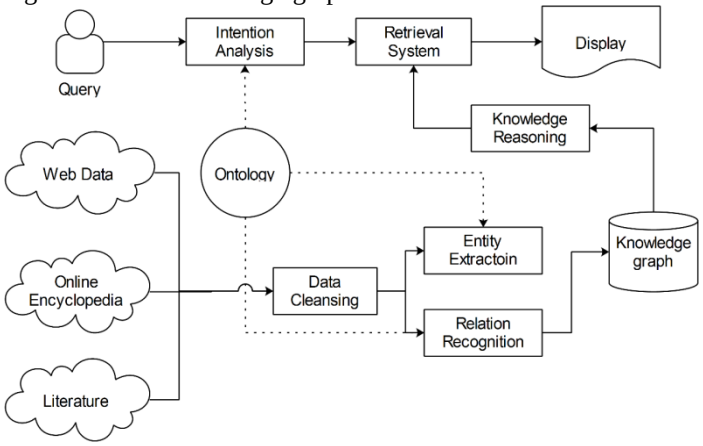

Fig. 1. The Work Process of KnowHealth.

2.1 Data collection and Ontology definition

Our platform extracts crisp elderly health information from the various encyclopedia sites (Baidu Baike, Wikipedia, etc.) and several websites of health care(chunyuyisheng.com, haodf.com, etc.). Because our KnowHealth mainly focus on elderly health issue, we restrict the crawler download text that related geriatric diseases. Those diseases are from Baidu Baike and then refined by medical experts. With process of data collection, we found that the content of user concern in health care Websites includes the basic information of disease (such as etiology, identification, etc.), the disease medication, diet contraindication, chief doctors who are adept at the disease and related hospitals. Hence, we can obtain the ontology representation for one object in Figure 2, in which one ontology object is made up of 3 parts: the attributes of object, the relations to other objects and the object's components. For example, the features of heart disease are cardiac enlargement, cardiac sounds, arrhythmia etc.. The components of heart disease include coronary heart disease, pulmonary heart disease etc.. The heart disease also have many relation with other object, e.g. ``checking'' relation with

angiocardiography. We then constructed 7 ontology categories, as shown in the Figure 3, the box means one ontology category, the arrow indicates the relationship among ontology categories.

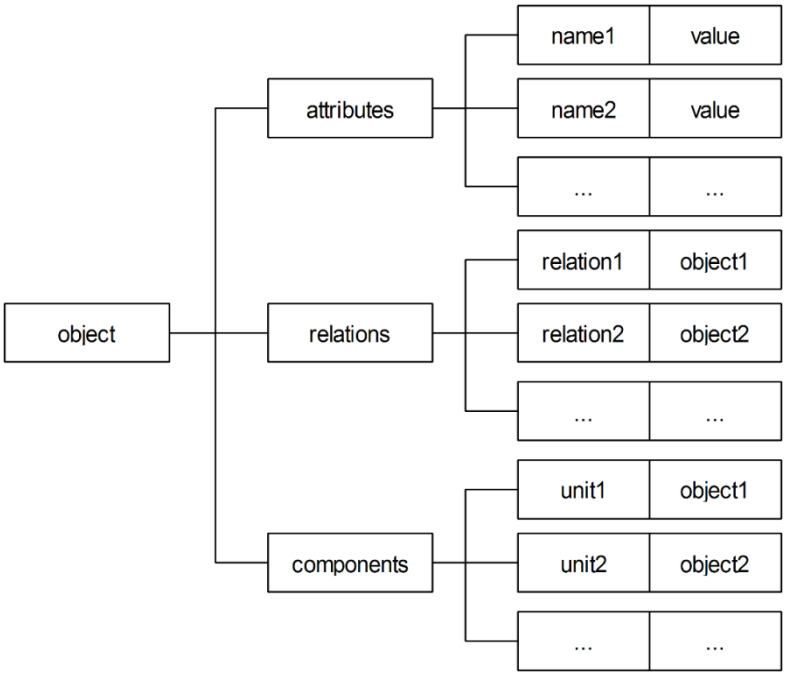

Fig. 2. The Ontology Representation in KnowHealth.

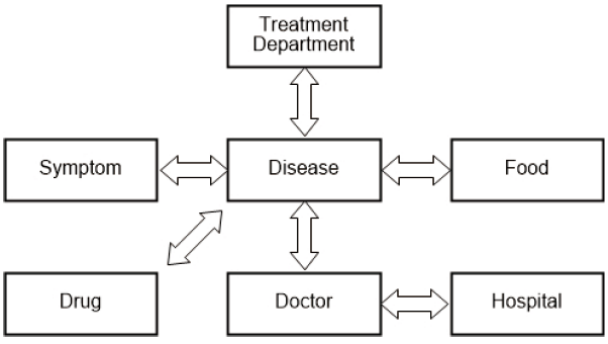

Fig. 3. The Ontology Category Relation in KnowHealth

2.2 Entity Extraction

The first significant stage in the KnowHealth work process is identifying all entities in whole text corpus. We use two approaches to extract health care related entities. (a) There

exist many entities on semi-structured webpage which can be downloaded on the online encyclopedia. Each of these pages can be parsed into an entity. After that those entities are stored in the local dictionary to control the extraction process on non-structural text corpus. (b) We also extract the entity on sentence-level, which means a sentence contains one or more entities. We employ hybrid method to parse sentences in health care related text sources, in which the local dictionary and NLP(Natural Language Processing) based method are combined to extract almost 1 million sentences. Specifically, we use NLP to identify all subjects and objects in a sentence. If one of them is included in the local dictionary, we assume other subjects or objects which aren't contained in dictionary are candidate for entities.

2.3 Relation Recognition

Relation recognition is another important stage in construction of knowledge graph. KnowHealth currently supports 14 binary relations among entities. There are definition, symptoms, identifications, treatments, prevention, affects, alleviates, causes, complication, goodto, diagnoses, interacts, badto, sideeffect. In KnowHealth, we adopt a variant of association rule mining algorithm to detect the relation among entities. We consider the sentence-level co-occurrence as relation among entities. Then, We count the frequency of co-occurrences in whole corpus and the confidence and support of association rule mining can be used to estimate the relation quality. If two entities have a relation, both confidence and support are required to be larger than specified thresholds. Through parameter adjustment in manual testing, we set support=0.02, confidence=0.6. And the type of the relations can be inferred by ontology category definition. However, if the entity linked with the relation is one new entity and its ontology category is not clear, then we should employ a classifier to determine its ontology category. Moreover, we also need discriminate the type of some relation. For example, a new food is good or bad to a disease.

In order to achieve best performance of relation extraction, we implement and test several classifiers for recognition, including libsvm, c4.5 (Decision tree), NN (Neural Network) and Naive Bayes. We leverage Precision as the evaluating indicator. Let S denote the number of entities/relations acquired by classifier, M denote the real entities/relations set which is manually labeled as the ground truth label. Then Precision can be defined as $P=|S \cap M|/|S|$ and the comparison results can be show in table 1. The SVM classifier obtains highest accuracy and our platform will employ svm to extract the relations between the entities.

Table 1. Comparison of Classifiers on Entity/Relation Recognition.

Task	Classifier	Accuracy
Entity classificaition	libsvm	0.773
	ANN (Neural Network)	0.692
	c4.5 (Decision tree)	0.637
	NB (naive bayes)…	0.590
Relation classificaition	libsvm	0.595
	ANN (Neural Network)	0.556
	c4.5 (Decision tree)	0.527
	NB (naive bayes)…	0.489

2.4 Knowledge Fusion

Knowledge fusion refers to the process of finding the same entity that belongs to the real world for each entity in the knowledge base of heterogeneous data sources. This study prepares to compare the cosine similarity of entities in the field of elderly health. When it is difficult to identify the result accurately, we use the similarity of attribute information to further determine whether the entities in the heterogeneous knowledge source can be aligned. This experiment uses the Limes algorithm to fuse the entity relationships with 88446 entities and set up metric expressions. When the trigram similarity of the name of the entity X and the name of the entity Y is greater than or equals to 0.9, the entities will be aligned. When the trigram similarity of the name of the entity X and the name of the entity Y is less than 0.9 but greater than or equals to 0.6, the entities will be reviewed. Then, according to the subsequent entity attribute similarity, whether the entities need to be aligned is determined. If the trigram similarity of the name of the entity X and the name of the entity Y is less than 0.6, the entities will not be aligned.

Through the above steps, 3762 pairs of entities have been aligned, and there are 9275 pairs of entities that need to be reviewed. For the further identification of entities, let the attribute name set of the entity e_α be $\mathrm{Property}_\alpha = \{p_{\alpha1},\ p_{\alpha2}, \dots,\ p_{\alpha m}\}$ and its attribute value set be $\mathrm{Value}_\alpha = \{v_{\alpha1},\ v_{\alpha2}, \dots,\ v_{\alpha m}\}$. And let the attribute name set of the entity e_b be $\mathrm{Property}_b = \{p_{b1},\ p_{b2}, \dots,\ p_{bm}\}$ and its attribute value set be $\mathrm{Value}_b = \{v_{b1},\ v_{b2}, \dots,\ v_{bm}\}$. Then, the common attributes of the entity e_α and the entity e_b meets the formula $p_i \in \mathrm{Property}_a \cap \mathrm{Property}_b$, and the formula for calculating the attribute similarity is shown:

$$\mathrm{sim}(p_i) = \mathrm{lcs}(v_{\alpha x}, v_{by}) / \max(\mathrm{len}(v_{\alpha x}), \mathrm{len}(v_{bx})) \tag{1}$$

In this formula, $\mathrm{lcs}(v_{\alpha x}, v_{by})$ represents the longest common subsequence of entity attribute values. According to the results of the similarity of each common attribute of entity e_α and entity e_b in the previous part, the calculation formula for the similarity of entity e_α and entity e_b is concluded:

$$\mathrm{property\_sim}(e_\alpha, e_b) = \left[\sum_{i=1}^{T} \mathrm{sim}(p_i)\right] / T \tag{2}$$

T is the size of the common attribute set of entity e_α and entity e_b. By determining the similarity of entity attributes, the extracted entities are aligned. In the end, because different data sources may generate different attribute values for the same attribute, attribute values need to be decided when entities are aligned. In this part, the decision method is on the basis of the reliability of data source and the attribute values that appear in multiple data sources.

As the entities and relations are acquired, our system uses the Neo4j graph database to store the data. Compared with the traditional semantic storage based on RDF, Neo4j graph database has the advantage of high scalability and high efficiency. Each node in Neo4j represents one entity that exists in KnowHealth, and each edge is the relationship between the entities, so that all kinds of information are connected together to form a network of knowledge. And it can facilitate knowledge reasoning and the excavation of potential knowledge. As shown in the Figure 4, the elderly who are suffering from “hypertension” tend to have head swelling, headache and other symptoms. In our graph database, we can search the relation of “good-to” and the ontology of “food”. The search result may be some fruit. Hence, the elderly should eat more fruit. Through the relationship between the nodes, it can be inferred that hypertension is easy to accompany the symptoms and diseases and so on.

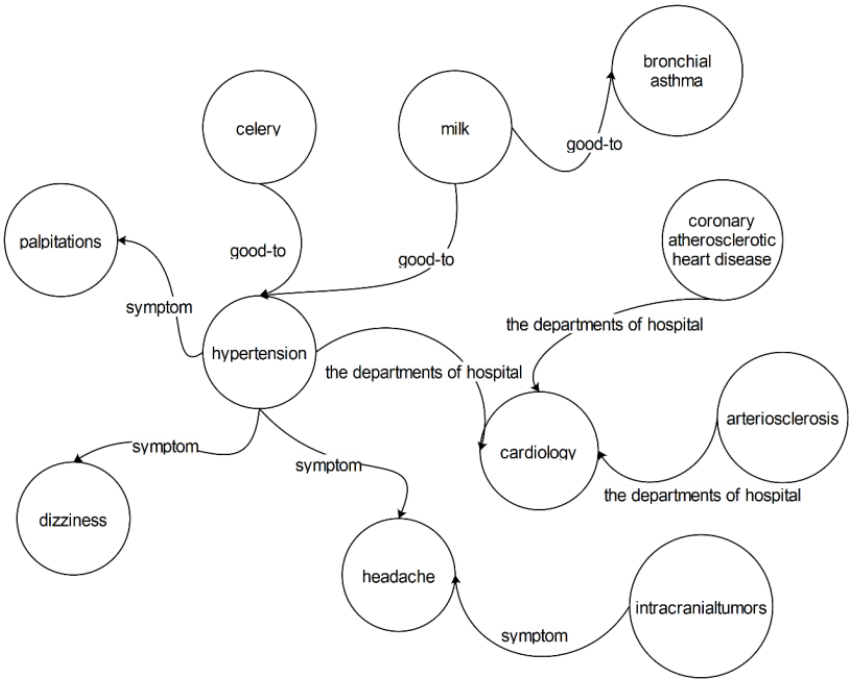

Fig. 4. An Example of hypertension defined in KnowHealth

3 A Question-answer Platform

When the knowledge graph has been constructed, we obtain an effective tool to explore the in-depth knowledge on health care domain which can provides (1) answer for health care relevant questions; (2)visual results to display relationship between the data. In this section, we establish an interesting application to search, browser health care knowledge for the elderly. As show in Figure 5, we design a historical behavior driven question-answering platform based on knowledge graph. The platform is divided into four hierarchical structures, including infrastructure layer, data processing layer, service layer and Web layer. In infrastructure layer, the QA platform provides the physical resources for KnowHealth, e.g. graph database ``Neo4j'' and Web server ``solr''. In data processing layer, the QA platform crawl the elderly data on the multiply Websites, clean the noisy date, extract the entities and relations and stored information in the Neo4j database of infrastructure layer. In service layer, the QA platform can provide the accurate and comprehensive knowledge, when elderly ask a health care related question. The Web layer is the interface between elderly and QA platform. The main implement step of the QA platform includes (1) determine the category of questions that the elderly asked in platform; (2) analyze the semantic objective of elderly's retrieval; (3) reason proper answer for elderly.

3.1 Elderly's Question Classification

The questions from the elderly always contained various types of sentences. In order to understand the elderly's intention, we should know the types of elderly input problem. For

example, the elderly input a question “what foods is good for diabetes?”. After the classification step, we can know the question is related to diet. So the answer will be extracted only focusing on the elderly’s diet, regardless of the search results from other categories as an answers’ candidate. Thus, the problem classification plays an important role in the whole question and answer system, and its performance will have a great impact on the final results. Previous works on problem classification mainly include rule-based methods and the machine-learning-based methods. However, the rule-based methods need a lot of manual participation. And this paper will employ classification method to handle this issue. Through consulting with experts, the problems in KnowHealth are divided into 12 categories. The table II shows the details of problem categories.

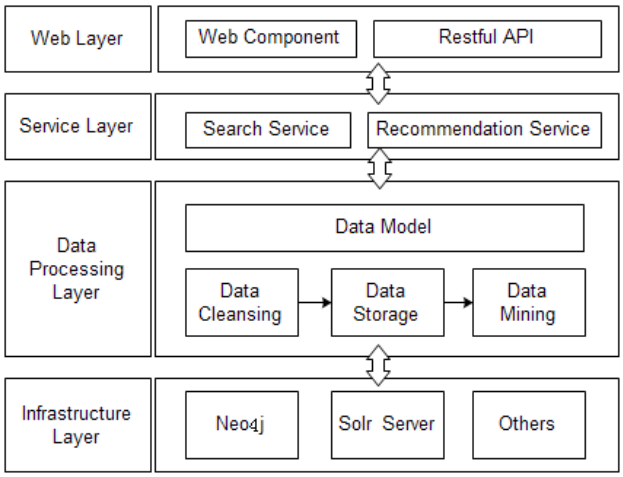

Fig. 5. The Historical Behavior Driven Question-Answering Platform

Table 2. 12 problem categories.

No	Type Of Question	Example
1	definition class	What is hypertension?
2	part class	What is the location of hypertension?
3	department class	What department of treatment should hypertension go to?
4	drug class	What drug should hypertension take?
5	diet class	What dietary contraindication does hypertension have?
6	symptom class	What are the common symptoms of hypertension?
7	pathogenesis class	What causes hypertension?
8	check class	What examinations should hypertension be done?
9	prevention class	How to prevent hypertension?
10	identification class	How to identify hypertension?
11	treatment class	How to treat hypertension?
12	complication class	What are the common complications of hypertension?

At first, we manually tagged a large number of elderly health questions taken from the Internet. Then, according to their types, we evenly selected 3600 different categories of questions. We first constructed eigenvectors using traditional one-hot encoding method, and then used the decision trees, naive Bayes, and SVM classifiers to classify the texts. Next, we used word embedding to map vocabularies into low-dimensional vectors, then used CNN to classify the texts. Finally we got the experiment results. The training data is passed through different classifiers based on the ten-fold-cross validation method. Accuracy for each class is calculated and is shown in Table III followed. From the data in the table, it can be seen that CNN performs better than the other three classifiers under most conditions. We apply the trained CNN classifier to the QA platform.

Table 2. 12 problem categories.

	Na ïveBayes	**DecisionTree**	**SVM**	**CNN**
Accuracy	0.675	0.612	0.593	0.772

3.2 Answer Retrieval

The target of answer retrieval in QA system is that the system first analyzes user input problem (extension of problem), generate the keywords of retrieval. Then, those resulting keywords are retrieved in a collection of documents to get a set of related sentences. In this section, we established a knowledge graph based on online search engine. When the elderly search knowledge in KnowHealth, as can be seen in Figure 6, the system will classify his/her input in accordance with the trained classification model. Thereafter it will use NLP algorithm to identify the named entity with the ontology library and local dictionary. Later, the system retrieves the knowledge through the knowledge graph and returns the acquired sentences to the elderly. For example, when the elderly input “What are the symptoms of hypertension”, the system will determine which class the problem belongs to. And then it finds the “hypertension” entity according to the results after word segmentation with the local dictionary. Finally, the system searches in knowledge graph in conjunction with search patterns for symptom class and returns its possible symptoms to the users.

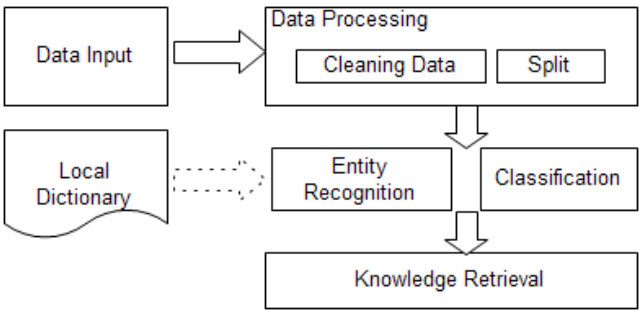

Fig. 6. the Answer retrieval process

3.3 Historical Behavior Driven Answer Reasoning

In order to understand the elderly’s intention better, we record the users’ history behavior data when they visit our platform. Analyzing these data, our platform can provide the most

useful answer. Algorithm 1 is used to discover potential intention of the elderly on the basis of their history behaviors. At the beginning of this algorithm, we identify entities related to health domain in input sentences and classify the sentences into predefined 12 categories. Then the algorithm uses those entities and categories as input, searches the knowledge graph and finds the answer that is relevant to the problem. When there exist multiple answers for one elderly's query, the algorithm will extract all entities from the elderly's history behaviors and those entities are mapped to the node in knowledge graph. With that, those nodes are traversed by levels in graph (we see the input entities as the first level and center nodes in knowledge graph). Furthermore, our algorithm adds weights to the searched node. The rules of weights are: (1) the distance from searched node to center node is farther, the weight is smaller; (2) the historical time of searched node is longer, the weight is smaller; (3) the degree of the searched node is larger, the weight is smaller. Finally, our algorithm can filter out the highest weight node in the candidate answers and return it as the final answer. At the same time, Algorithm 1 adds the entities of current input into the history behavior data for the next search. For example, an elderly has searched a question "which disease will lead to dizziness?", so the keywords of sentence "dizzy" will be recorded in the historical database. When the elderly search on our platform "which disease can coronary heart disease cause?", the probability of "hypertension" will be higher than others disease nodes. After this search, the "dizziness" and the weight of node "hypertension" will be added to historical database of this elderly user.

Algorithm 1 Historical Behavior Driven Answer Reasoning

```
Require: an input sentence s, a knowledge network N, user's history search data H
Ensure: a most related answer a;
 1: Split the s into word set W
 2: for each wi 2 W do
 3:   if wi is entity then
 4:     Search the N according to the corresponding nodeof the entity and class C
            and get result set r
 5:     if isMultiply(r) then
 6:       Extract entities e from the user's history search data H
 7:       Map entities e to the nodes n in knowledge network N
 8:       for each ni 2 N do
 9:         for each level lj 2 Nk do
10:           Traversed the node in level lj
11:           Update the weight
12:         end for
13:       end for
14:     end if
            Select the highest weight node nh
15:   end if
16: end for
      Return the result form nh
```

4. Conclusion

The knowledge graph has become more and more prevalent due to its unique knowledge storage structure in various industries. In this paper, we apply the knowledge graph to the health care domain. With the analysis of a large number of related issues in the literature, we construct ontology library on the domain of health care. Then, we extract entities and relations between the entities according to the ontology. We also design an application to show the ability of our KnowHealth. We apply the knowledge graph to the question-answer system to reason the answers of the elderly's query. The system can address the elderly's requirement, and provide better advice combined with the users' history behavior. In the future work, we prepare to further update and improve the knowledge graph by constantly iterating network of large amounts of information.

Acknowledgements: This work was supported in part by the National Natural Science Foundation of China(No.61772159), National Natural Science Foundation of shandong province(No.ZR201702150244) and University Co-construction Project.

References

[1] E. Filtz, "Building and processing a knowledge-graph for legal data," in European Semantic Web Conference. Springer, 2017, pp. 184–194.
[2] P. Ping, K. Watson, J. Han, and A. Bui, "Individualized knowledge graph," Circulation research, vol. 120, no. 7, pp. 1078–1080, 2017.
[3] H. Paulheim, "Knowledge graph refinement: A survey of approaches and evaluation methods," Semantic web, vol. 8, no. 3, pp. 489–508, 2017.
[4] M. Rotmensch, Y. Halpern, A. Tlimat, S. Horng, and D. Sontag, "Learning a health knowledge graph from electronic medical records," Scientific Reports, vol. 7, 2017.
[5] K. Sankar, K. Jia, and R. L. Jernigan, "Knowledge-based entropies improve the identification of native protein structures," Proceedings of the National Academy of Sciences, p. 201613331, 2017.
[6] P. Ernst, A. Siu, and G. Weikum, "Knowlife: a versatile approach for constructing a large knowledge graph for biomedical sciences," BMC bioinformatics, vol. 16, no. 1, p. 157, 2015.
[7] X. Zhao, Z. Xing, M. A. Kabir, N. Sawada, J. Li, and S.-W. Lin, "Hdskg: Harvesting domain specific knowledge graph from content of webpages," in Software Analysis, Evolution and Reengineering (SANER), 2017 IEEE 24th International Conference on. IEEE, 2017, pp. 56–67.
[8] D. Kim, L. Xie, and C. S. Ong, "Probabilistic knowledge graph construction: Compositional and incremental approaches," in Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. ACM, 2016, pp. 2257–2262.
[9] D. Collarana, M. Galkin, I. Traverso-Rib ́on, C. Lange, M.-E. Vidal, and S. Auer, "Semantic data integration for knowledge graph construction at query time," in Semantic Computing (ICSC), 2017 IEEE 11th International Conference on. IEEE, 2017, pp. 109–116.
[10] S. Choudhury, K. Agarwal, S. Purohit, B. Zhang, M. Pirrung, W. Smith, and M. Thomas, "Nous: Construction and querying of dynamic knowledge graphs," in Data Engineering (ICDE), 2017 IEEE 33rd International Conference on. IEEE, 2017, pp. 1563–1565.
[11] T. Palomares, Y. Ahres, J. Kangaspunta, and C. R ́e, "Wikipedia knowledge graph with deepdive." in Wiki@ ICWSM, 2016.
[12] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, "Learning entity and relation embeddings for knowledge graph completion." in AAAI, 2015, pp. 2181–2187.
[13] G. Zhu and C. A. Iglesias, "Sematch: Semantic similarity framework for knowledge graphs," Knowledge-Based Systems, 2017.

[14] S. Yang, L. Zou, Z. Wang, J. Yan, and J.-R. Wen, “Efficiently answering technical questions-a knowledge graph approach.” in AAAI, 2017, pp. 3111–3118.
[15] Y. Zhu, W. Zhou, Y. Xu, J. Liu, and Y. Tan, “Intelligent learning for knowledge graph towards geological data,” Scientific Programming, vol.2017, 2017.
[16] T. Yu, J. Li, Q. Yu, Y. Tian, X. Shun, L. Xu, L. Zhu, and H. Gao, “Knowledge graph for tcm health preservation: Design, construction, and applications,” Artificial Intelligence in Medicine, vol. 77, pp. 48–52, 2017.