KnowHealth: a Large Knowledge Graph for Elderly Health Information Processing

Zhuoran Zhen[1,a], Dianhui Chu[1,b], Chunshan Li[1,c], Yuanyuan Wang[2,d]

[1]School of Computer Science and Technology, Harbin Institute of Technology, Weihai, China

[2]Shenzhen University, Shenzhen, China, 518060

[a]misaki_526@163.com, [b]Chudh@hit.edu.cn, [c]lics@hit.edu.cn, [d]yywang@szu.edu.cn

Keywords: Knowledge Graph; Biomedical text mining; Relation extraction; Question answering platform

Abstract: Nowadays, population aging has become a prominent problem all over the world, which brings new challenge to manage huge health data. Existing approaches store those huge health knowledge data in traditional relational databases through a labor-intensive process and lack well-defined ontology to describe entities and relations in elderly health domain. Knowledge graph has better knowledge managing ability than traditional storage methods due to its logical structure. In this paper, we design a knowledge graph (KnowHealth) to manage the elderly's health data. Specifically, we automatically crawl and extract information related to health from different input sources, and construct a knowledge graph to manage them in the graph database. We give an ontology definition which describes the entities and relations with different types and attributes. In order to serve query for the elderly, we design and implement a historical behavior driven by a question-answer platform. Finally, we conduct a discussion on several key technologies for KnowHealth.

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, such as the encyclopedia platform and the Website about health care. Existing approaches [1] to handle those data have two major limitations. (1) A large amount of health data exists on the Web and the data is updated frequently [2]. Consequently, processing the data manually is inapplicable. (2) Previous works [3], [4] store those data in a traditional relational database through a structured table. This storage mode can't manage the various relations among health data, which 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 [5]. Therefore, it is important 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 has become a hot research trend due to its logical structure. Google improved the efficiency and accuracy of the search by employing knowledge graph technology in 2012. Baidu, Sogou, Microsoft and other giants started to follow up the domain of knowledge graph, e.g. Baidu Zhixin, Sogou Zhilifang, and Microsoft Renlifang. Nevertheless, these knowledge graphs are all used in the open domain which aren't often 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. Yang et al. [6] suggested solving the technical issue of Microsoft products by establishing knowledge graph and using index-based random walk method. Yu et al. [7] presented a method of constructing knowledge graph for traditional Chinese medicine which is 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.

Facing with the above issues, this paper designs a knowledge graph (KnowHealth) to manage the

data in elderly's health care domain, and constructs 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 entities, mining the relationship of entities and other steps. We construct a knowledge graph (KnowHealth) to manage the processed concepts and relations in the graph database. (2) Consulting with medical expert, we give an ontological definition of 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 by a 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-answer platform is presented in Section 4 and the paper is concluded in Section 5.

2. Related Words

During the past few years, the Internet has been a vast repository of health knowledge, but automatically extracting health knowledge on a large scale has proven to be a tough challenge. Many recent studies have focused on construct knowledge base, and many well-known broad domain and open knowledge managements appear, for example Google knowledge graph, the Never-Ending Language Learning (NELL) project and the OpenIE which used a variety of techniques to extract new knowledge from the Web. More recently, Pujara et al. [8] designed an online knowledge graph construction method, with which a scalable probabilistic model was built to combining statistical features with uncertain extractions and ontological constraints. Kapanipathi et al. [9] designed a Hierarchical Interest Graph which could leverage hierarchical relationships presented in knowledge-bases to infer users' interests. Yang et al. [10] proposed the public cultural knowledge platform, which provided users with rich cultural knowledge instead of simple data.

3. Construction of Knowledge Graph

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

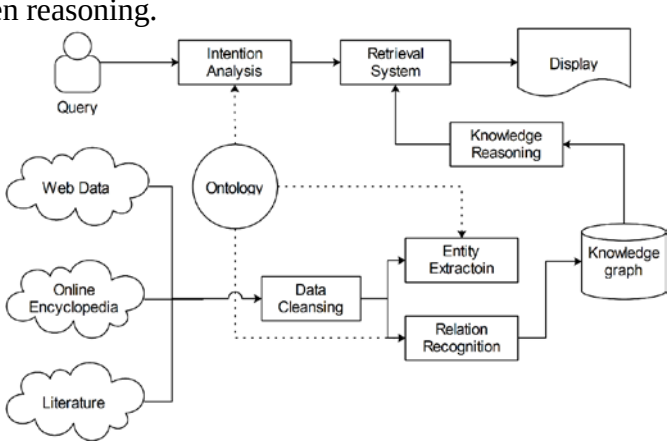

Fig. 1. The Work Process of KnowHealth

Fig. 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. Later, the entities and relations are stored

in a graph database to form the knowledge graph. The upper part of Fig. 1 shows an application of our knowledge graph. When users 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 results from knowledge graph.

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 focuses on elderly health issues, we limit the crawler to download text that is related to geriatric diseases. The knowledge of disease is crawled from Baidu Baike and then refined by medical experts. Through the process of data collection, we found that the content of users' concern in health care Websites includes the basic information of disease (such as etiology, etc.), the disease medication, diet contraindication, chief doctors who are adept at the disease and related hospitals. Then we constructed 7 ontology categories. As shown in Fig. 2, the box means one ontology category, and the arrow indicates the relationship among ontology categories.

1) Entity Extraction: The first significant stage in the working process of KnowHealth is identifying all entities in the whole text corpus. We use two approaches to extract entities related to health care. (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 controls 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 text sources related to health care, in which methods based on the local dictionary and NLP (Natural Language Processing) 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. Then, we check whether the verb in this sentence represents a predefined relationship or not. If the verb belongs to specific relation, we determine that the subject/object is a new entity.

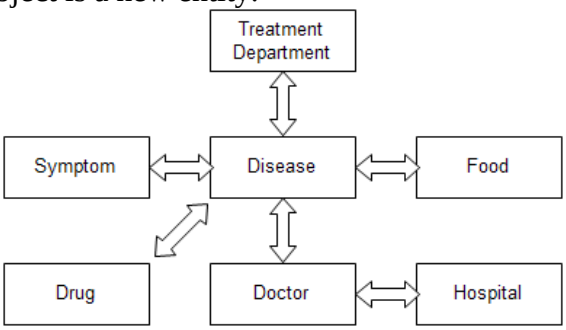

Fig. 2. The Ontology Category Relation in KnowHealth

2) Relation Recognition: Relation recognition is another important stage in the construction of knowledge graph. KnowHealth currently supports 14 binary relations among entities. There are definition, treatment, causes, symptoms, identifications, prevention, affects, alleviates, complication, good-to, diagnoses, interacts, bad-to and side-effect. 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 and 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, we should employ a classifier to determine its ontology category.

3) Data Storage: As the entities and relations are acquired, our system uses the Neo4j graph database to store the data. Each node in Neo4j represents one entity that exists in KnowHealth, and each edge is the relationship between the entities. As shown in the Fig. 3, the elderly who are

suffering from “hypertension” tend to have headache and other symptoms. Through the relationship between the nodes, it can be inferred that hypertension is easy to accompany the symptoms and diseases and so on.

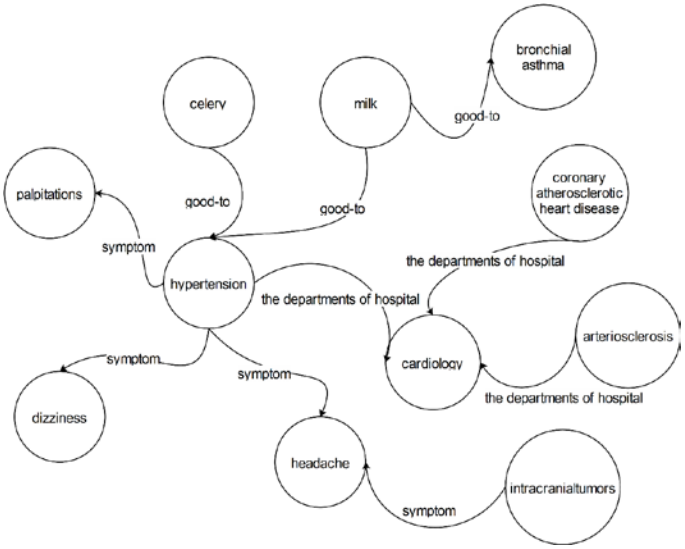

Fig. 3. An Example of hypertension defined in KnowHealth

4. An Application of Knowhealth

When the KnowHealth has been constructed, we obtain an effective tool to explore the in-depth knowledge in health care domain which can provide (1) answers for healthcare-relevant questions; (2) visual results to display relationship between the data. In this section, we design a question-answer platform driven by historical behaviors based on knowledge graph. The main implement steps of the QA platform include (1) determining the category of questions that the elderly asked in platform; (2) analyzing the semantic objective of the elderly’s retrieval; (3) reasoning proper answer for the elderly.

4.1 Elderly’s Question Classification

In order to understand the elderly’s intention, we should know the types of elderly input problem. Through consulting with experts, the problems in KnowHealth are divided into 12 categories. The table I shows the details of problem categories.

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 use 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 Fig. 4. 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.

4.2 Answer Retrieval

When the elderly search knowledge in KnowHealth, the system will classify his 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. If 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.

TABLE I. 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?

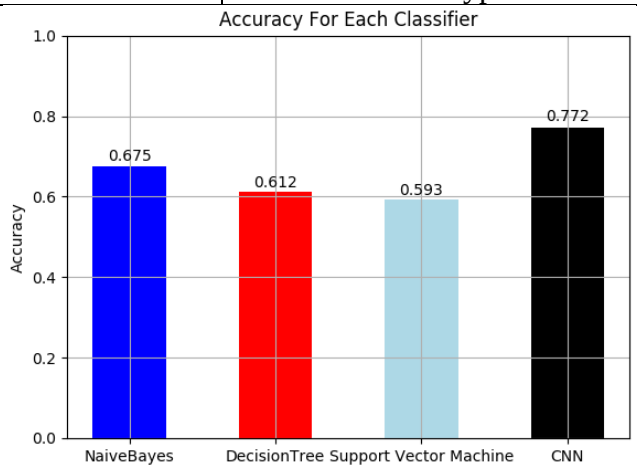

Fig. 4. Accuracy for each classifier

4.3 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. Algorithm 1 is used to discover potential intention of the elderly on the basis of their history behaviors. Firstly, we identify entities related to health domain in input sentences and classifies 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. 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. Finally, our algorithm can filter out the highest weight node in the candidate answers.

Algorithm 1 Historical Behavior Driven Answer Reasoning

Require: input sentence *s*, a knowledge network *N*, user's history search data *H*
Ensure: the most related answer *a*;
1: Split the *s* into word set *W*
2: **FOR** each w_i in *W* **DO**
3: **IF** w_i is entity **THEN**
4: Search the *N* according to the corresponding node of the entity and class *C* and get result set *r*
5: **IF** *isMultiply(r)* **THEN**
6: Extract entities *e* from *H*
7: Map entities *e* to the nodes *n* in *N*
8: **FOR** each n_i in *N* **DO**
9: **FOR** each level l_j in N_k **DO**
10: Traversed the node in level l_j
11: Update the weight
12: **END FOR**
13: **END FOR**
14: **END IF**
Select the highest weight node n_h
15: **END IF**
16: **END FOR**
Return the result form n_h

5. Conclusion

In this paper, we apply the knowledge graph to the health care domain and 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.

Acknowledgment

This work is support in part by the National Natural Science Foundation of China (No. 61772159), National Natural Science Foundation of Shandong Province (No. ZR201702150244), the Scientific Research Foundation of Harbin Institute of Technology (HIT.NSRIF.201703).

References

[1] P. Ping, K. Watson, J. Han, and A. Bui, "Individualized knowledge graph," Circulation research, vol. 120, no. 7, pp. 1078–1080, 2017.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] J. Pujara, B. London, L. Getoor, and W. W. Cohen, “Online inference for knowledge graph construction,” in Fifth International Workshop on Statistical Relational AI, 2015.

[9] P. Kapanipathi, P. Jain, C. Venkataramani, and A. Sheth, “User interests identification on twitter using a hierarchical knowledge base,” in European Semantic Web Conference. Springer, 2014, pp. 99–113.

[10] Y. Yang, G. Zhang, J. Wang, S. Ye, and J. Hu, “Public cultural knowledge graph platform,” in Semantic Computing (ICSC), 2017 IEEE 11th International Conference on. IEEE, 2017, pp. 322–327.