IMAS: An Intelligent Medical Analysis System Based on Concept Graph

Xiaoli Wang[1]([envelope]), Yixiang Xie[1], Siqi Xie[1], Zhifeng Bao[2], and Shuwen Su[1]

[1] School of Informatics, Xiamen University, China
xlwang@xmu.edu.cn; {yxx,sqx}@stu.xmu.edu.cn; sushuwen@xmu.edu.cn
[2] RMIT University, Australia
zhifeng.bao@rmit.edu.au

Abstract. We develop an intelligent medical analysis system, denoted by IMAS, to support medical practice based on concept graph. The system provides an efficient medical archives processing tool and supports effective medical data searching and analysis. We first collect clinical data and sensor data from several hospitals. Historical medical archives are processed to provide valuable prior knowledge for patients. Second, we employ a novel data modeling technique based on evolving graphs to effectively support medical practice. Graph indexing and searching algorithms are implemented to support efficient medical case searching. With our system, users can easily pinpoint valuable historical medical information they are interested in, retrieve closely relevant medical cases for further diagnosis, and navigate to our interactive Q&A platform for treatment information. To our best knowledge, this is the first full-fledged system to examine every phase in the smart healthcare system pipeline based on concept graph. We have implemented the system and the source codes are available at https://github.com/emmali808/ADDS.

Keywords: Concept Graph · Medical Case Search · Automatic Diagnosis

1   Introduction

Electronic information systems have become more and more popular in the smart healthcare industry. Many existing systems are designed to collect medical data as electronic records [4]. The data are the fundamental resources to support medical practice such as case searching and automatic diagnosis (see a comprehensive study in [3]). Existing systems generally support applications in a big data setting [3]. However, in developing countries, most hospitals do not have advanced information systems, and only a limited number of electronic data are collected. As far as we know, there is no mature medical retrieval or analysis system, which can have practical use in the Chinese medical industry.

In this paper, we develop an intelligent medical retrieval and analysis system denoted by IMAS, to support medical practice based on concept graph. The

system aims to provide an efficient medical archives processing tool and support effective medical data searching and analysis. In general, historical medical documents are created and stored in archives. Such data are a possible source of very valuable knowledge that provide prior medical information for patients. Thus, it is indeed important to first address the medical archives processing problem before we deploy medical practice in a big data setting. Several OCR based methods have been proposed to convert medical archives into electronic records (e.g., [5]). However, when applying these methods to our collected medical archives, many of them lack direct information on disease or diagnosis. To address the problem, we propose a novel concept graph based classification approach to automatically label medical archives.

With effective medical archives processing, we can collect numerous valuable historical medical information for patients. However, medical data contain massive information such as drugs, diseases and treatments. How to model the complex data for supporting medical practice is a very important and challenging problem. In our recent work [7] a novel modeling technique has been proposed to represent the data as sequence of evolving graphs. The graph model is shown to have better data expressivity for supporting efficient medical case searching and automatic diagnosis. We employ our constructed medical concept graph to model clinic data of patients as evolving graphs, and propose an efficient graph similarity search algorithm to support effective lazy learning for diagnosis prediction. In our system, we also implement a crowdsourcing-based expert Q&A platform to interactively improve the machine learning results.

2 System Overview

Figure 1 shows the architecture of our IMAS system, which contains six main components: medical concept graph construction, medical archives processing, data modeling, data searching, data analysis and Q&A platform.
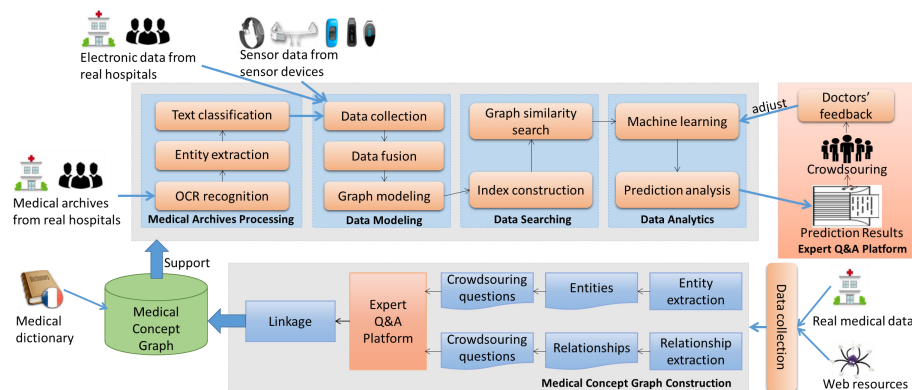

Fig. 1. System architecture (Several icon photos are from the Google search engine.)

We built a sematic-rich medical concept graph using medical dictionaries (e.g., ULMS[3]), web resources (e.g., Wikipedia[4]) and real clinical data. We extract six types of entities (Drug, Drug_Category, Disease, Disease_Category, Symptom and TestItem) and four types of relationships (HasSymptom, Diagnose, Treat and Subcategory-of). The details of the ontology can be seen in [7]. We employ the medical concept graph as the knowledge base in our system.

We first perform the recognition task of medical archives using OCR engines (e.g., [5]). Given a medical text D, we split it into a set of candidate words W_D, and map each word in W_D to an entity of the medical concept graph, to generate an entity set E_D. If D contains a "disease" entity, we use it as the class label. Otherwise, we input E_D into the text classification module for further label assignment. We define semantic measures between medical texts to improve typical text classification algorithms, such as K-Nearest Neighbor (KNN) [1] and Support Vector Machines (SVM) [2]. The proposed enhanced algorithms, and their effectiveness can be seen in our previous work [6].

With medical archives processing, we collect valuable historical data. To model the data, our previous work [7] employ concept graph to model clinic data as evolving graphs. Two vertices in the graph are connected with one edge only when there exists relationship between matching entities of them. In IMAS, we improve the modeling technique by using semantic information instead of direct relationships. Given a medical text D with its entity set E_D, we build a graph using the semantic relatedness among E_D. For each pair of entities in E_D, we compute their semantic similarity using Definition 1 cited from [6]. If the value is larger than a threshold, we connect an edge between them. In general, clinical data capture patients' visits to hospitals. If a visit at each time point is constructed as a graph, a sequence of evolving graphs can be formed as the final representation. Our system employs the graph mapping distance proposed in [7] to evaluate the similarity between two graphs. A three-level inverted index is built to support efficient medical case searching. In data analysis component, we use graph similarity search to efficiently support predication tasks. Obviously, a medical graph sequence can be simply predicted using a majority vote of its k nearest neighbors. More details on the graph similarity searching and prediction algorithms can be seen in [7]. In our IMRAS system, the predication results from the data analysis component will be sent to the expert Q&A platform. Real doctors can judge the prediction and give feedbacks. These feedbacks will be used to improve the accuracy and comprehensiveness of prediction results.

Definition 1 (Entity Semantic Similarity). *Given two entities e_1 and e_2 in concept graph, the semantic distance $SD(e_1, e_2)$ between them is the number of hops in a shortest path connecting them. Then, the semantic similarity between them is defined as inversely related to their distance, i.e., $SS(e_1, e_2) = \frac{1}{\max\{SD(e_1,e_2),1\}}$.*

---

[3] http://ulms.org.uk
[4] http://wiki.dbpedia.org

Fig. 2. The demonstration of our IMAS system

3   Demonstration

In our IMAS system, users first log in using their registration accounts. Then, users can search and navigate to any page by clicking the menus on the left. In Fig. 2, when clicking the "Upload Medical Records" menu, a user enters the page for uploading medical clinical records. The medical archives processing component extracts medical texts and do classification if needed. After processing, the user can click the "Machine Diagnosis" menu to open the diagnosis page. The processed medical texts are used as input to the data modeling component for constructing a profile graph based on concept graph, as shown in Step 2. In Step 3, the system performs both data searching and analysis to return predication results. The top results are returned with their profile graphs and their doctor notes are used to produce a diagnosis message. In this example, the user may have a bad cold by comparing against the result graphs. The user can then click the "Submit Questions" menu to do consultation by submitting questions to our expert Q&A platform or online communicating with the AI robot as shown in Step 4 and 5. Doctors can give their feedbacks by entering the "Q&A" page. Our system periodically collects doctors' answers for a certain question. Noted that doctors' feedbacks on automatic diagnosis results can be returned to adjust the prediction algorithms. The AI robot is designed to answer questions online when no doctor is available in our system. More details about our system can be checked in https://github.com/emmali808/ADDS.

Acknowledgments

References

1. Altman, N.S.: An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician **46**(3), 175–185 (1992)
2. Cortes, C., Vapnik, V.N.: Support-vector networks. Machine Learning **20**(3), 273–297 (1995)
3. Lee, C., Luo, Z., Ngiam, K.Y., Zhang, M., Zheng, K., Chen, G., Ooi, B.C., Wei, L.J.Y.: Big healthcare data analytics: Challenges and applications. In: Scalable Computing and Communications. pp. 11–41 (2017)
4. Scells, H., Locke, D., Zuccon, G.: An information retrieval experiment framework for domain specific applications. In: ACM SIGIR. pp. 1281–1284 (2018)
5. Thompson, P., Mcnaught, J., Ananiadou, S.: Customised ocr correction for historical medical text. In: Digital Heritage. pp. 35–42 (2016)
6. Wang, X., Wang, R., Bao, Z., Liang, J., Lu, W.: Effective medical archives processing using knowledge graphs. In: SIGIR. pp. 1141–1144 (2019)
7. Wang, X., Wang, Y., Gao, C., Lin, K., Li, Y.: Automatic diagnosis with efficient medical case searching based on evolving graphs. IEEE Access **6**, 53307–53318 (2018)