Effective Medical Archives Processing Using Knowledge Graphs

Xiaoli Wang
Xiamen University, China
xlwang@xmu.edu.cn

Rongzhen Wang
Quanzhou Medical College, China
rongzhen0704@gmail.com

Zhifeng Bao
RMIT University, Australia
zhifeng.bao@rmit.edu.au

Jiayin Liang
Xiamen University, China
jyliang@stu.xmu.edu.cn

Wei Lu*
Renmin University of China
lu-wei@ruc.edu.cn

ABSTRACT

Medical archives processing is a very important task in a medical information system. It generally consists of three steps: medical archives recognition, feature extraction and text classification. In this paper, we focus on empowering the medical archives processing with knowledge graphs. We first build a semantic-rich medical knowledge graph. Then, we recognize texts from medical archives using several popular optical character recognition (OCR) engines, and extract keywords from texts using a knowledge graph based feature extraction algorithm. Third, we define a semantic measure based on knowledge graph to evaluate the similarity between medical texts, and perform the text classification task. This measure can value semantic relatedness between medical documents, to enhance the text classification. We use medical archives collected from real hospitals for validation. The results show that our algorithms can significantly outperform typical baselines that employs only term statistics.

CCS CONCEPTS

• **Applied computing** → **Health informatics**.

KEYWORDS

Medical archives processing; Medical information system; Knowledge graphs

ACM Reference Format:
Xiaoli Wang, Rongzhen Wang, Zhifeng Bao, Jiayin Liang, and Wei Lu. 2019. Effective Medical Archives Processing Using Knowledge Graphs. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3331184.3331350

*The corresponding author.

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331350

1 INTRODUCTION

Electronic information systems have become very popular in the healthcare industry. Existing systems have focused on supporting medical practice in a big data setting [6]. However, in developing countries, many hospitals do not have advanced information systems. Before we deploy big data analytics for healthcare, it is important to address the medical archives processing problem [10]. Historical medical archives often contain very valuable knowledge, providing prior medical information for patients. However, most of them are paper archives, making it difficult to apply those valuable knowledge to medical research. Therefore, several OCR based methods have been proposed to convert medical archives into electronic records (e.g., [10, 12]). In practice, it is more meaningful to assign the medical archives with class labels. Several studies employ typical classification techniques for label assignment (e.g., [5]), assuming that the recognized information by OCRs are legible, which unfortunately is not true in practice [10]. Recent works try to employ deep learning models for text classification [3, 7]. However, such methods may incur heavy training cost.

This paper presents a novel **K**nowledge **G**raph based **M**edical **D**ocument **M**ining **F**ramework called KG-MDMF with three main components, including medical knowledge graph construction, medical archives processing and text classification. First, we construct a semantic-rich medical knowledge graph using real clinical data, web resources and medical dictionaries. Second, we recognize texts from medical archives based on OCR engines, and extract keywords from the texts. The keywords are further used to classify medical documents in the third component. With the support of the knowledge graph, we define a semantic similarity between medical texts to support the text classification task. Our main contributions are as follows:

- We build a semantic-rich medical knowledge graph, using entities and relationships that are extracted from various resources.
- We propose a domain-specific word similarity measure based on the knowledge graph to select representative keywords for medical texts.
- We define a semantic similarity between medical texts using the knowledge graph, which is used to enhance typical text classification algorithms. Experiments also show encouraging results that the improved algorithms can outperform the baselines.

2 METHODS

2.1 Knowledge Graph Construction

Over the years, we focus on building a semantic-rich medical knowledge graph using web resources in the medical domain (e.g., Wikipedia[1]), medical dictionaries (e.g., ULMS[2]) and real clinical data. We first built a basic knowledge graph using entities and relationships from medical web resources and dictionaries. Then, we extracted entities and relationships from clinical data using word segmentation tools [11]. For the uncertain results, we submitted crowdsourcing questions to the expert Q&A system, and selected the correct answers using the majority vote algorithm. Those verified results would be integrated into the knowledge graph. Finally, we extracted six types of entities (Drug, Disease, Symptom, TestItem, Drug_Category and Disease_Category) and four types of relationships (HasSymptom, Diagnose, Treat and Subcategory-of). We semantically grouped entities of the same category together and formed a tree-like conceptual hierarchy. We have published partial conceptual hierarchies online in http://47.94.174.82:8080/ADDS/index.jsp.

2.2 Semantic Measures

We define a similarity score to measure the semantic closeness between two entities by considering their category chain in the corresponding conceptual hierarchy.

Definition 2.1 (Category Chain). Given an entity e which is mapped to its parent category C, the category chain C_e of entity e is a set consisting of all the categories that are in the path from the root category to category C.

Definition 2.2 (Conceptual Distance). Given two entities e_1 and e_2 with their parent categories C_1 and C_2 respectively, the conceptual distance $CD(e_1, e_2)$ between them is defined as the number of hops in a shortest path from category C_1 to category C_2.

Definition 2.3 (Conceptual Similarity). Given two entities e_1 and e_2 with their category chains C_{e_1} and C_{e_2}, the conceptual similarity $CS(e_1, e_2)$ between them is computed as $CS(e_1, e_2) = \frac{C_{e_1} \cap C_{e_2}}{d \times \max\{CD(e_1,e_2),1\}}$, where d is the depth of the conceptual hierarchy.

The conceptual similarity is used to evaluate two entities belonging to the same conceptual hierarchy. Otherwise, we have $CS(e_1, e_2)$=0. If $e_1 = e_2$ and $d = 0$ (i.e., we have only one entity in the hierarchy), we have $CS(e_1, e_2)$=1. We also define a similarity score to measure the semantic closeness between two entities in the knowledge graph.

Definition 2.4 (Semantic Distance). Given two entities e_1 and e_2 in the medical knowledge graph, the semantic distance $SD(e_1, e_2)$ between them is the number of hops in the shortest path connecting them.

The distance can be computed using the shortest path searching algorithm[3]. Then, we define the semantic similarity between two entities in the medical knowledge graph as inversely proportional to their distance, i.e., $SS(e_1, e_2) = \frac{1}{\max\{SD(e_1,e_2),1\}}$. If there is no path connecting these two entities, then the semantic distance is defined as infinite. That is, the semantic similarity between them is equal to 0.

Definition 2.5 (Entity Similarity). Given two entities e_1 and e_2 from medical texts, the similarity $ES(e_1, e_2)$ between them is $ES(e_1, e_2) = \alpha CS(e_1, e_2) + (1 - \alpha)SS(e_1, e_2)$, where α is the weight ($\alpha \in (0, 1]$).

Given two medical texts D_1 and D_2 with their word vectors W_{D_1} and W_{D_2}, we map each word in the vectors into an entity in the medical knowledge graph. Then, we have two entity sets U and V. Suppose $E \in U \times V$ is a set of edges weighted by the entity similarity $ES(e_i, e_j)$ between each pair of entities $e_i \in U$ and $e_j \in V$, we could form a perfect bipartite graph $G=(U, V, E)$ and define the semantic similarity as follows.

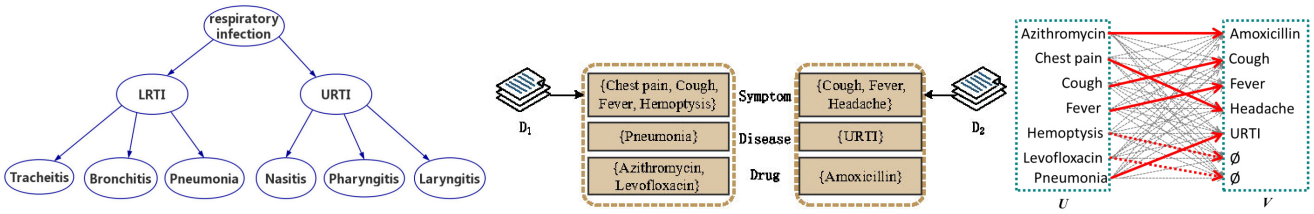
Definition 2.6 (Text Semantic Similarity). Given two medical documents D_1 and D_2 with their normalized entity sets U and V of the same cardinality and assuming that $P: U \to V$ is a bijection. The semantic similarity between them is

$$TSS(D_1, D_2) = \frac{\max\limits_{P} \sum\limits_{e_i \in U} ES(e_i, P(e_i))}{\max\{|U|, |V|\}}.$$

Here, $|U|$ and $|V|$ are cardinalities of the entity sets. The computation of this similarity score can be formulated as the maximum weighted bipartite matching problem, which can be solved by employing the Hungarian algorithm [4]. If two entity sets are of different cardinalities, $\emptyset$ node is inserted into the bipartite graph for normalization. The similarity between any entity and $\emptyset$ is 0.

Example 2.7. In Figure 1, we have two medical texts D_1 and D_2 with normalized entity sets U={Azithromycin, Chest pain, ..., Pneumonia} and V={Amoxicillin, Cough, ..., $\emptyset$}. We compute the entity similarity for each pair of entities in $E \in U \times V$. We take two entities with e_1="Pneumonia" and e_2="URTI" as an example. Due to the conceptual hierarchy in Figure 1 (a), we compute their conceptual distance as $CD(e_1, e_2)$=1, as there is an 1 hop path between their parent categories ("LRTI→Respiratory infection"). We also have C_{e_1}={LRTI, Respiratory infection} and C_{e_2}={Respiratory infection}. Thus, $C_{e_1} \cap C_{e_2}$=1. We compute their conceptual similarity as $CS(e_1, e_2) = \frac{C_{e_1} \cap C_{e_2}}{d \times \max\{CD(e_1,e_2),1\}}$=1/2. We then calculate $SD(e_1, e_2)$=3, as there is a shortest path of 3 hops in the knowledge graph ("Pneumonia→LRTI→Respiratory infection→URTI"). Thus, $SS(e_1, e_2) = \frac{1}{\max\{SD(e_1,e_2),1\}}$=1/3. We have $ES(e_1, e_2) = \alpha CS(e_1, e_2) + (1-\alpha)SS(e_1, e_2)$=0.42. Here, we set α=0.5. Given similarity values for all pairs of entities, we can form a bipartite graph and find a maximum matching illustrated as red arrows in Figure 1 (c). Then, the text semantic similarity $TSS(D_1, D_2)$ is obtained by dividing the summation of entity similarity values in the maximum matching to $|U|$.

[1] http://wiki.dbpedia.org
[2] http://ulms.org.uk
[3] https://en.wikipedia.org/wiki/Shortest_path_problem



(a) An example of conceptual hierarchy (b) Extracted entities for two medical texts (c) The maximum matching result

Figure 1: An example for illustrating the proposed techniques

2.3 Medical Archives Processing

The second component of KG-MDMF has three modules: OCR recognition, preprocessing and keyword extraction. The OCR recognition module extracts texts from medical archives using several OCR based engines. With recognized medical texts, the preprocessing module filters out the words that do not contribute to the classification task. Given a medical text D, we split it into a set of candidate words W_D using Jieba [13] and remove stop words from W_D occurring frequently in all medical texts. In the keyword extraction module, we map each word in W_D to an entity of the medical knowledge graph. Then, we have an entity set E_D for each medical text D. We first build a graph model for D using the semantic relatedness among E_D based on the knowledge graph. For each pair of entities e_i and e_j in E_D, we compute their semantic similarity $SS(e_i, e_j)$. If the value is larger than 0, we create an edge between them and assign the value as the weight to the edge. Based on the graph model, we adopt the TextRank algorithm [8] to rank the nodes for extracting representative keywords.

2.4 Improved Classification Algorithms

2.4.1 The Improved KNN Algorithm. A medical document is assigned to a class by a majority vote of its k nearest neighbors. The training data are vectors in a multidimensional feature space (we use the keyword or entity as the feature), each with a class label (we use the disease as the class label). KNN [1] classifies a given medical document in two steps.

In the training step, it stores the entity vectors and class labels of the training medical documents. In the classification step, given a user-defined constant k, an unlabeled medical document called a test point is classified by being assigned the label which occurs most frequently among the k training medical documents nearest to the test point.

The typical KNN algorithm commonly used the cosine similarity as the distance metric between two term frequency vectors using tf-idf weights. Recent studies take into account of the semantic information (e.g., [5, 9]). We define a novel similarity measure $Sim(D_1, D_2)$ by considering both the data statistics and the semantic information:

$$Sim(D_1, D_2) = \beta WES(D_1, D_2) + (1 - \beta)TSS(D_1, D_2).$$

$WES(D_1, D_2)$ is a similarity considering both tf-idf and word embedding [9]. $TSS(D_1, D_2)$ is the text similarity from Definition 2.6 and β is the weight.

2.4.2 The Improved SVM Algorithm. SVM is a supervised learning algorithm, which can support binary classification. In medical practice, data are often not linearly separable in the feature space. To solve non-linear problems, studies based on the kernel trick are proposed, such as the linear kernel and the Gaussian kernel function. It has been shown that a Gaussian kernel performs a mapping into an infinite dimensional space, which can better handle the case when all classes are not linearly separable in the input space and often yields better performance than the linear kernel [2]. We use the SVM algorithm based on a semantic-based Gaussian kernel as a baseline [2]:

$$K(D_1, D_2) = exp(-\gamma||(D_1 - D_2)^T P^T P(D_1 - D_2)||^2).$$

Here, γ is an adjustable parameter of the kernel function. In general, its default value is set as $\frac{1}{N}$, where N is the number of features. P is any appropriately shaped matrix. The matrix P typically encodes pairwise term similarities by considering the concept similarity, which is inversely proportional to their shortest path distance in a conceptual graph.

In this paper, we employ the semantic-based kernel by considering the semantic relatedness from both the concept hierarchy and the medical knowledge graph. That is, we compute the pairwise term similarities in P using the entity similarity in Definition 2.5.

3 EVALUATION

We conduct experiments on a server with 32GB memory, running Centos 5.6. All the algorithms are implemented using Python.

3.1 Settings

An EMR dataset and a medical archives dataset (MRD) are collected from Chinese hospitals [11]. The EMR has 45,000 samples that are admissions from 2010 to 2018, containing 20 total class labels (we use disease code as the class label). Clinical data including demographics, lab tests, diagnoses and medications are collected. For each admission, there is one disease code. The MRD has 15,000 samples that are admissions from 2008 to 2014. For text recognition, we use two OCR tools: ABBYY[4] and Baidu AI[5]. 5,000 samples are recognized with explicit class labels. The well-labeled 50,000 samples from both datasets are used as the training samples,

[4] http://www.abbyychina.com
[5] http://ai.baidu.com/docs#/OCR-API/top

and the remaining unlabeled 10,000 documents are used as the testing samples. Table 1 shows the statistics and Table 2 shows the parameter settings.

Table 1: Dataset statistics

Datasets	No. of features	No. of samples
Training samples	200	50,000
Testing samples	192	10,000

Table 2: Parameter settings

Parameters	Setting
k for the KNN classification	20
α	0.5
β	0.5
γ	0.005

To evaluate the accuracy for the testing data, we conduct an empirical study with the help of an expert Q&A system. For example, given the classification result on one medical archive, we designed a simple yes or no question, which was distributed into the expert Q&A system for collecting answers from experts (i.e., doctors from real hospitals): Does the patient of this medical archive have the disease of the given class label? Experts can choose "yes" or "no" to answer the question. We statistically analyzed the collected answers and the given class label is considered to be correct only when all the answers are "yes". For 10,000 testing samples, we collected their answers and obtained 2,400 samples with validated class labels. We compute the average values of *Precision*, *Recall* and *F-measure* for all classes in 2,400 samples as the evaluation metrics.

3.2 Results on text classification

We used the KNN [1] with pure WES, SVM [2] with semantic-based Gaussian kernel and CNN [3] algorithms as baselines. Table 3 shows the results for text classification. The improved KNN algorithm called KG-KNN and the improved SVM algorithm called KG-SVM respectively outperform the KNN and SVM baseline algorithms. A surprising result is that the improved SVM algorithm even outperforms the costly CNN based method. By comparing and analyzing the results on each metric, we also draw a conclusion that the semantic relatedness can highly improve the recall of the classifier.

4 CONCLUSION

This paper presents a medical archives processing framework called KG-MDMF with the support of knowledge graphs. The experimental results have proved that the KG-MDMF can outperform the typical text classification algorithms on various metrics. We surprisingly found that the improved SVM algorithm could perform better than the costly deep learning algorithm.

Table 3: The results on classification

Algorithms	*Precision*	*Recall*	*F-measure*
KNN	0.57	0.48	0.51
KG-KNN	0.58	0.55	0.56
CNN	0.72	0.71	0.71
SVM	0.71	0.6	0.65
KG-SVM	**0.78**	**0.75**	**0.76**

ACKNOWLEDGMENTS

This work was supported by NSFC (61702432), the Fundamental Research Funds for Central Universities of China (20720180070) and the International Cooperation Projects of Fujian in China (2018I0016). This work was partially supported by the Research Funds administered by the Digital Fujian, at the Big Data Institute for Urban Public Safety. Rongzhen Wang was partially supported by Research Funds of Fujian Province for Young Teachers (JAS161068). Zhifeng Bao was partially supported by ARC (DP170102726, DP180102050) and NSFC (61728204, 91646204), and is a recipient of Google Faculty Award. Wei Lu was partially supported by Beijing Municipal Science and Technology Project (Z171100005117002) and NSFC (U1711261).

REFERENCES

[1] N. S Altman. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician* 46, 3 (1992), 175–185.
[2] Claudio Carpineto, Carla Michini, and Raffaele Nicolussi. 2009. A Concept Lattice-Based Kernel for SVM Text Classification. ICFCA, 237–250.
[3] M Hughes, I. Li, S Kotoulas, and T Suzumura. 2017. Medical Text Classification Using Convolutional Neural Networks. *HTI* 235 (2017), 246.
[4] Harold W. Kuhn. 2010. The Hungarian Method for the Assignment Problem. *Naval Research Logistics* 52, 1 (2010), 7C21.
[5] Kleanthi Lakiotaki, Angelos Hliaoutakis, Serafim Koutsos, and Euripides G. M. Petrakis. 2013. Towards personalized medical document classification by leveraging umls semantic network. In *Health Information Science*. 93–104.
[6] Chonho Lee, Zhaojing Luo, Kee Yuan Ngiam, Meihui Zhang, Kaiping Zheng, Gang Chen, Beng Chin Ooi, and Luen James Yip Wei. 2017. Big Healthcare Data Analytics: Challenges and Applications. In *Scalable Computing and Communications*. 11–41.
[7] Cheng Lin Liu, Gernot A. Fink, Venu Govindaraju, and Lianwen Jin. 2018. Special issue on deep learning for document analysis and recognition. *IJDAR* 21, 3 (2018), 159–160.
[8] Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Text. *Emnlp* (2004), 404–411.
[9] Yixuan Tang, Weilong Huang, Qi Liu, Anthony K H Tung, Xiaoli Wang, Jisong Yang, and Beibei Zhang. 2017. QALink: Enriching Text Documents with Relevant Q&A Site Contents. In *ACM CIKM*. 1359–1368.
[10] Paul Thompson, John Mcnaught, and Sophia Ananiadou. 2016. Customised OCR correction for historical medical text. In *Digital Heritage*. 35–42.
[11] Xiaoli Wang, Yuan Wang, Chuchu Gao, Kunhui Lin, and Yadi Li. 2018. Automatic Diagnosis With Efficient Medical Case Searching Based on Evolving Graphs. *IEEE Access* 6 (2018), 53307–53318.
[12] Li Xiang, Hu Gang, Xiaofei Teng, and Guotong Xie. 2015. Building Structured Personal Health Records from Photographs of Printed Medical Records. In *AMIA Annu Symp Proc*. 833–842.
[13] H P Zhang, H K Yu, D Y Xiong, and Q Liu. 2003. HHMM-based Chinese lexical analyzer ICTCLAS. In *Sighan Workshop on Chinese Language Processing*. 758–759.