Medical Knowledge Extraction Using Truth Discovery Framework

[1]Gia Abraham & [2]Liston Deva Glindis

[1]Master Of Engineering, Dept. of CSE, Dhanalakshmi Srinivasan College Of Engineering, Coimbatore, India, Mail Id:- giakodath@gmail.com

[2]Assistant Professor, Dept. of CSE, Dhanalakshmi Srinivasan College Of Engineering, Coimbatore, India, Mail Id: - listong@dsce.ac.in

Abstract

With the booming new technology, the traditional health care system is undergoing an evolution. The medical crowd sourced question answering (Q&A) website is one among them. There are a lot of patients and doctors involved in these crowd sourced question and answering websites. The valuable information from these medical crowd sourced Q&A websites can benefit patients, doctors and the society. Facing the daunting scale of information generated on medical Q&A websites every day, it is unrealistic to full fill this task via supervised method due to the expensive annotation cost. In this concept, we propose a Medical Knowledge Extraction (MKE) system that can automatically provide high quality knowledge triples extracted from the noisy question-answer pairs, and at the same time, estimate expertise for the doctors who give answers on these Q&A websites. The MKE system is built upon a truth discovery framework, where we jointly estimate trustworthiness of answers and doctor expertise from the data without any supervision. We further tackle three unique challenges in the medical knowledge extraction task, namely representation of noisy input, multiple linked truths, and the long-tail phenomenon in the data. The MKE system is applied on real-world datasets crawled from xywy.com, one of the most popular medical crowd sourced Q&A websites. This system can automatically provide high quality knowledge information extracted from the noisy question-answer pairs.

Keywords: - Mine medical attributes, Signature Mining, Disease Inference

1. INTRODUCTION

With the booming new technology, the traditional health care system is undergoing an evolution. Besides visiting a doctor in person for the health concerns, the young generations would also prefer to search the information readily available on the Web, or ask the doctors through the Internet. As an emerging industry, this new type of health care service brings opportunities and challenges to the doctors, patients, and service providers. Compared to the traditional one-to-one service, the online medical crowd sourced question answering (Q&A) websites provide crowd-to-crowd. The medical crowd sourced question answering (Q&A) websites are booming in recent years, and increasingly large amount of patients and doctors are involved. One of the most important challenges of extracting knowledge from the medical crowd sourced Q&A websites is that the quality of question-answer pairs is not guaranteed. The questions asked by patients can be noisy and ambiguous. The answers' quality varies due to reasons such as doctors' expertise, their level of commitment, and their purpose of answering questions. To extract useful knowledge, it is important to distinguish relevant and correct information from unrelated or incorrect information. In the light of this challenge, one possible solution is to label the quality of question-answer pairs and then learn classification or regression models. Such models are then used to judge the quality of each question-answer pair. However, this approach may not work on the medical Q&A dataset. To build a training set, we need to hire experts to annotate the labels, but the cost can be prohibitive to annotate enough training examples. Besides, medical knowledge is highly domain specific, and because of this, multiple experts may be needed and this further boosts the cost. In addition, the privacy protection of patients adds to the difficulty of annotation. These issues motivate us to develop unsupervised learning methods that extract knowledge from noisy crowd sourced data on medical Q&A websites.

2. RELATED WORK

Existing System

In the existing, Information technologies are transforming the ways healthcare services are delivered, from patients' passively embracing their doctors' orders to patients' actively seeking online information that concerns their health. They are disseminating personalized health knowledge and connecting patients with doctors worldwide via question answering. Existing truth discovery methods can only work on structured data, but the knowledge extraction task deals with unstructured and noisy text data. Trustworthy answers to a health-related question are usually not unique. There may be multiple possible answers to the same question, and those answers are likely to be correlated. In this concept are have main drawback in finding the prescription from DB side and they reach only in mentioned data.

Proposed System

We propose a MKE (Medical Knowledge Extraction) system that can jointly conduct the medical knowledge extraction and doctor expertise estimation without any supervision. In medical crowd sourced Q&A websites, patients have various intentions when asking questions. For example, they may want to find out the possible diseases based on their symptoms, or particular side-effects of a drug. Doctors, who play essential roles in these Q&A websites, provide answers to these questions. For the same question, multiple doctors may give different answers due to their diverse expertise. In order to distil the trustworthy medical knowledge, we propose a truth discovery method to automatically estimate doctors' expertise, and conduct weighted aggregation based on the estimated doctor expertise. To apply the truth discovery framework, we first extract entities from texts and transform texts into entity-based representations. The new representations will then be fed into the proposed truth discovery method, which outputs the medical knowledge triples <question, diagnosis, trustworthiness degree> and the estimated doctor expertise.

Advantages

- Based on these outputs, various real-world applications can be built. For example, the extracted medical knowledge triples can be used to answer medical questions in Automatic Diagnosis and Medical Robot.
- Besides, the estimated doctor expertise can be applied in the tasks such as Doctor Ranking and Question Routing, which play important roles in crowd sourced Q&A websites.

Scope

The objective of this system is to extract knowledge triples <question, diagnosis, trustworthiness degree> from noisy question-answer pairs in medical crowd sourced Q&A websites. Meanwhile, for the doctors who give answers in these Q&A

websites, their expertise will be automatically estimated. A truth discovery method is proposed to automatically extract medical knowledge from noisy crowd sourced question answering websites without any supervision. This method provides a cost efficient and effective way to mine knowledge from crowd sourced question answering websites. The proposed truth discovery method is designed to tackle the new challenges in the medical knowledge extraction task, and the experimental results on real world dataset confirm its effectiveness. Last but not least, we demonstrate a real-world medical application built upon the proposed method. This application, Ask A Doctor, shows that the extracted knowledge can enable and facilitate many online healthcare applications.

Challenges

Although the medical knowledge extraction task can be formulated as a truth discovery problem, the basic truth discovery method overlooks some unique challenges of the task. Therefore, the basic truth discovery method have to be adapted to the medical knowledge extraction task. The first challenge we are facing is how to clean the noisy input. Existing truth discovery methods can only work on structured data, but the knowledge extraction task deals with unstructured and noisy text data. In order to achieve better performance, we need to derive better representations of the questions and answers. The second challenge of the medical knowledge extraction task is that there may be multiple trustworthy answers to a question, and these answers can be correlated with each other. This challenge violates an important assumptions of many truth discovery methods which is the single truth assumption. The assumption is that there exists one and only one trustworthy answer for each question. In many crowd sourcing applications, the long-tail phenomenon is observed. That is, for most doctors, they provide answers to a few questions, and only a small set of doctors provide answers to many questions. Or for most questions, the number of received answers is small, and only a small set of questions receive a large number of answers. Both types of long-tail distribution exist, but existing truth discovery work that handles long-tail phenomenon only considers the long-tail from the perspective of sources, i.e., most sources provide only a few answers.

3. IMPLEMENTATION

Health seeker needs analysis (Mine medical attributes)

This module takes user query and converts it in to a form which can be processed by the system. Medical term with the negative content has been filtered out. In medical field users do not share same vocabulary. We use META MAP tool to filter to detect medical attributes that are noun phrases in health domain, and then normalize them to standardized terminologies in the SNOMED CT.

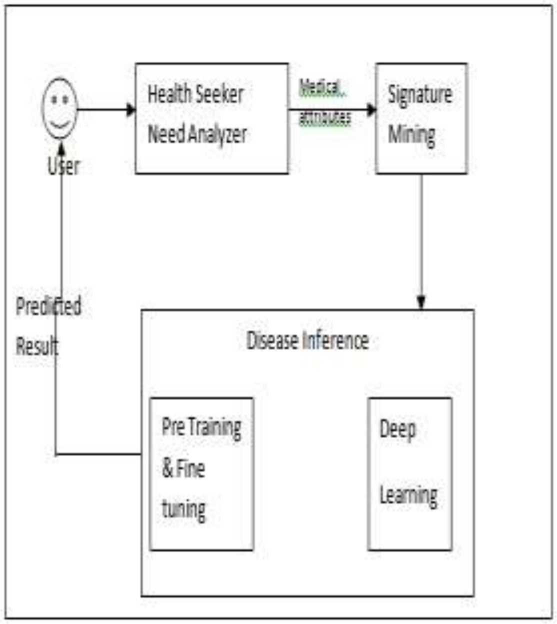
Signature Mining

This module mines the medical signatures. Signatures are inter dependent medical attributes and they are essential cues for disease. Latent signatures are represented by sub graphs embedded in a global graph. The nodes are normalized medical attributes and edges represent their co-occurrence relations.

Disease Inference

Sparsely connected deep learning is introduced to infer the disease from medical attributes. The proposed deep learning consists of L layers. The first layer consists of the input raw features. The Lth layer represents the output disease type. The intermediate layers are hidden layers. The optimal model is trained by minimizing the cost function. The gradient descent is used to update the cost function parameters. These signatures are viewed as hidden nodes and placed in the first hidden layer. Before stacking up the next hidden layer, we treat the current architecture as a learning model with only one hidden layer, and learn the initialized W1 and b1 with pseudo-labelled samples. Following that, we view the hidden nodes in the first hidden layer as the raw features, and define the co-occurrence relations between two hidden. A graph is constructed with hidden nodes as vertices and their relations as edges. Again perform the dense graph detection with mined signatures as the node. Then process the third hidden layer.

System Architecture



4. POST IMPLEMENTATION

A Post-Implementation Review (PIR) is an assessment and review of the completed working solution. It will be performed after a period of live running; some time after the project is completed. There are three purposes for a Post-Implementation Review: To ascertain the degree of success from the project, in particular, the extent to which it met its objectives, delivered planned levels of benefit, and addressed the specific requirements as originally defined. To examine the efficiency of all elements of the working business solution to see if further improvements can be made to optimize the benefit delivered. To learn lessons from this project, lessons which can be used by the team members and by the organization to improve future project work and solutions.

5. SYSTEM EVALUATION

The system evaluation involves the hardware and software as a unit. The hardware selection is based on performance categories. The evaluation phase ranks vendor proposal and determines the one suited to the user's needs. It looks in to items such as price, availability and technical support. In the operation phase, the system performance must be monitored not only to determine whether or not they perform as planned, but also to determine if they should be modified to meet changes in the information needs of the business. In the evaluation phase, the first step adopted was to look at the criteria listed earlier and rank them in the order of importance. Three sources of information are used in evaluating hardware and software. They are benchmark program, experience of other users and product reference manuals.

6. EXPERIMENTAL RESULTS

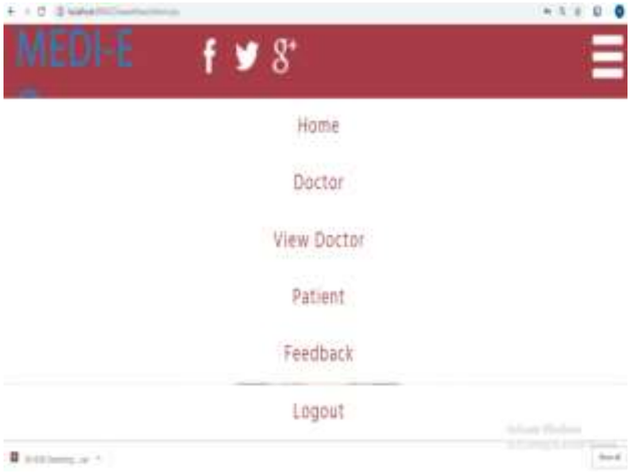


Fig:2 Application Menu

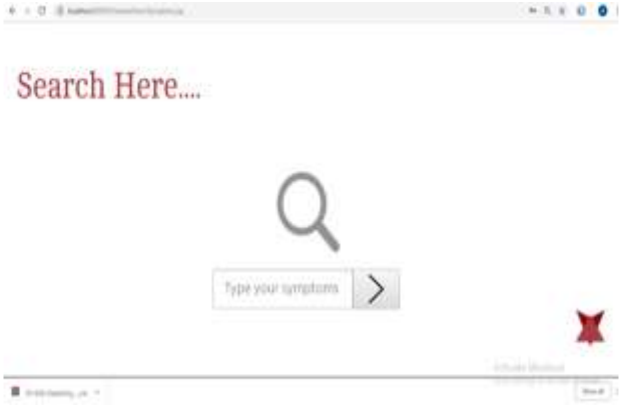
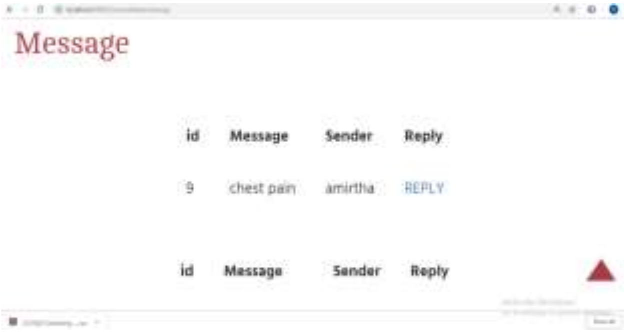


Fig:3 Search Page


Message

id	Message	Sender	Reply
9	chest pain	amirtha	REPLY

id	Message	Sender	Reply

Fig:4 Results Page

7. CONCLUSION

The medical crowd sourced Q&A websites provide valuable but noisy health related information. To extract high quality medical knowledge from the question-answer pairs, we propose a Medical Knowledge Extraction (MKE) system in this paper. The MKE system can extract knowledge triples <question, diagnosis, trustworthiness degree> and estimate doctors’ expertise simultaneously without any supervision. Three unique challenges in medical knowledge extraction tasks are recognized and tackled in the MKE system. We use entity-based representation to clean noisy text input and merge similar questions; A similarity function is applied to model the correlation between answers; And to handle the long-tail phenomenon on sources, a pseudo count is added so that we can estimate reasonable doctor expertise for each doctor. A set of experiments on real-world datasets crawled from xywy.com validate the effectiveness of the proposed MKE system in automatically extracting meaningful knowledge and estimating fine-grained doctor expertise from medical crowd sourced Q&A websites. We also show a real-world application, Ask A Doctor, to demonstrate the impact of the MKE system. Beyond this App, the MKE system has great potential to benefit more applications such as robot doctors and question routing in Q&A websites.

REFERENCES

[1] S. Fox and M. Duggan, “Health online 2013,” Pew Research Center, Survey, 2013.

[2] “Online health research eclipsing patient-doctor conversations,” Makovsky Health and Kelton, Survey,2013.

[3] T. C. Zhou, M. R. Lyu, and I. King, “A classification based approach to question routing in community question answering,” in *The International World Wide Web Conference*,2012.

[4] D. A. Davis, N. V. Chawla, N. Blumm, N. Christakis, and A.-L. Barabasi, “Predicting individual disease risk based on medical history,” in *The International Conference on Information and Knowledge Management*, 2008.

[5] L. Nie, M. Wang, Z. Zha, G. Li, and T.-S. Chua, “Multimedia answering: Enriching text qa with media information,” in *Proceedings of the International ACM SIGIR Conference*, 2011.

[6] L. Nie, M. Wang, Y. Gao, Z.-J. Zha, and T.-S. Chua, “Beyond text qa: Multimedia answer generation by harvesting web information,” *Multimedia, IEEE Transactions on*, 2013.

[7] L. Nie, Y.-L. Zhao, X. Wang, J. Shen, and T.-S. Chua, “Learning to recommend descriptive tags for questions in social forums,” *Acm Transactions on Information System*, 2014.

[8] D. Zhang and W. S. Lee, “Extracting key-substring-group features for text classification,” in *The ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2006.

[9] M. Gall´e, “The bag-of-repeats representation of documents,” in *Proceedings of the International ACM SIGIR Conference*, 2013.

[10] F. Wang, N. Lee, J. Hu, J. Sun, S. Ebadollahi, and A. Laine, “A framework for mining signatures from event sequences and its applications in healthcare data,” *IEEE Transactions on Pattern*

Analysis and Machine Intelligence, 2013.