International Journal of Contemporary Research in Computer Science and Technology (IJCRCST) ISSN: 2395-5325
Volume 3, Special Issue 3 (September '2017)
Proceeding of - International Conference on Recent Trends in Computational Life Science and Information Technology,
Conference held at Tiruppur Kumaran College for Women,Tirupur, Tamilnadu, India.

EXTRACTING MEDICAL KNOWLEDGE FROM QUERY RELATED WEBSITE-A SURVEY

V. Meena Gomathy,
M.Phil, Research Scholar,
Department of Computer Science,
Kongunadu Arts & Science College,
Coimbatore, Tamilnadu, India.

M.Lalithambigai,
Associate Professor,
Department of Computer Science,
Kongunadu Arts & Science College,
Coimbatore, Tamilnadu, India.

Abstract: The medical query related websites are developing in recent years and a large number of patients and doctors are involved. The valuable information from these medical query websites can benefit patients, doctors and the society. It has been a difficult process that to extract medical knowledge from the noisy question-answer pairs and filter out unrelated or even incorrect information. Facing the problem of getting information generated on the medical query websites every day, it is unrealistic to fulfill this task via supervised method due to expensive annotation cost. In this paper, it is to be surveyed a Medical Knowledge Extraction (MKE) System that automatically provides high quality knowledge extracted from the noisy question-answer pairs and also estimate doctor's expertise who gives answers on these query websites. The MKE system is a truth discovery framework to estimate trustworthiness of answers and doctor expertise from the data. This further handle three unique challenges in medical knowledge extraction tasks as: 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 "*icliniq.com*", one of the most popular medical query related websites. Both quantitative evaluation and case studies demonstrate that the proposed MKE system can successfully provide useful medical knowledge and accurate doctor expertise. We further demonstrate a real-world application: "*Care for You* ", which can automatically give patients suggestions to their questions.

Keywords: Crowdsourced Question Answering, Medical Knowledge Extraction, Truth Discovery Introduction

I.INTRODUCTION

As a developing 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 question answering websites provide crowd-to-crowd service for example "*icliniq.com*" alone receives thousands of new health related questions everyday. The information from these crowd sourced query related websites is valuable [2],[3], but how to take good use of such information is a big question. One way to utilize such information is to extract knowledge from the medical query related websites. The most important challenge of extracting knowledge from medical query related 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 doctor's expertise, their purpose of answering questions. To extract useful knowledge, it is important to distinguish relevant and correct information from unrelated or incorrect information.

First truth discovery methods are developed for structured data (i.e., database), but for query related websites the inputs are unstructured data (i.e., text). Second, the answers are not unique, multiple answers and the answers are correlated or not. To address this, there is a model to correlate through a similarity function defined on the word vectors of answers. Third, to observe severe long-tail process in Q & A data. It is difficult to estimate doctor expertise and trust worthy answers. Most doctors provide only few answers and many questions receive only a few answers. To tackle the problem, a pseudo count for each doctor in the doctor expertise can be added.

To evaluate the proposed medical knowledge extraction system, the information can be collected from "*icliniq.com*" a popular online health service. We compare the knowledge with the expert annotation and validate the doctor expertise with the relevant information. The System explains the usefulness of extracting knowledge in our real-world application.

In Summary, this paper deals with:

Provide a truth discovery method to automatically extract medical knowledge from noisy query –related websites. It provides a cost – efficient method.

Demonstrating a real-world medical application built upon the proposed system. This application, "Care for You" shows that the extracted knowledge can enable and facilitate many online healthcare applications.

II. EXISTING SYSTEM

The objective of the system is to find knowledge triples < question, diagnostic disease, truth discovery value> from several different question – answer pairs from query related websites. The doctors' expertise will be updated automatically. In query related websites, users (or) patients have various thoughts when asking questions, i.e., they want to know about possible disease using symptoms, the side-effects of a drug

etc. For these questions, different doctors provide different answers.
In order to find out the true knowledge, the system proposes a truth discovery method. To produce a truth discovery framework, first find out entities from texts and transforms into entity-based representations which results in the output as, <question, diagnostic disease, truth discovery value >
This output will result in the development of medical applications such as Automatic diagnosis, Medical Robot, Doctor Ranking and question routing.

Some important terms used are :
Question : A question from a patient consists of a set of statements including patient name, age, symptoms to ask for a disease and drugs to follow.
Question topic: System contains already pre-defined questions in which each particular topic is concerned.
Doctor: Doctor is a person who answers the questions on query related websites.
Answer: An Answer is a diagnostic disease provided by a doctor for a particular question. Different doctors provide one to multiple answers and multiple doctors provide same answers too.
Claim : It is a tuple which contains question, a doctor ID and the corresponding answers from the doctor to the question.
Knowledge triple: It consists of question, diagnostic disease and truth discovery value.
Doctor expertise : Each doctor who answers the question is associated with a score, the doctor expertise can be estimated from data and weighted aggregation in derived.

Thus, the formal definition is as follows: If the set of medical questions QS and a set of doctors DS, Let ds denote answer to the qs-th question provided by ds-th doctor and the e_{ds} be the expertise score of the dsth doctor weighted aggregation is : { a_{ds} } qs Є Qs, to derive knowledge triples < question, diagnostic disease, truth discovery value > and final the doctor expertise.

III. METHODOLOGY

3.1TRUTH DISCOVERY METHOD

Truth discovery problem can be estimated using [4],[5],[6] which estimate trustworthiness answers and doctor expertise. Truth discovery methods take input tuples of < question, answer, doctor >. This method holds some principles as : if the doctor provides answers with high expertise, it is considered as truth discovery answers ; whereas if a doctor always provides truth discovery answers, then he/she is assigned as high expertise. Based on this, we can update the truth discovery answers and doctor expertise as follows :
The truth discovery value of a possible answer a_{qs} for qs-th question:

$$TD(a_{qs})=\sum_{ds\in DS} e_{ds}.\mathbb{I}(a_{qs},a_{qs}^{ds}) \quad (1)$$

Where $\mathbb{I}$-indicator function, $\mathbb{I}(a,b)$ =1 if a=b; otherwise $\mathbb{I}(a,b)=0$

Equation (1) is formulated based on the truth discovery.If e_{ds} is high, then truth discovery degree TD (a_{qs}) is also high. TD(a_{qs}) will be normalized if the sum of all answers truth discovery degree will be 1. Thus Td(a_{qs}) can be updated with the probability as :

$$e_{ds}=\log(1-\sum_{a\in V_{ds}} TD(a)/|v_{ds}|) \quad (2)$$

Where V_{ds}-answer provided by ds-th doctor.
Here the term $-\sum_{a\in V_{ds}} TD(a)/|v_{ds}|$ is the average degree of ds-th doctor's answer.

So $1-\sum_{a\in V_{ds}} TD(a)/|v_{ds}|$ as the probability of doctor providing wrong answers. From equation (2), it is clear that a doctor provide wrong answer will be given a lower expertise score.

Equation(1) calculates the truth discovery degree for each answer by conducting weighted voting where weights are doctor expertise scores.

Equation (2) updates expertise score for each doctor based on answers degree.

4. PROBLEMS AND SOLUTIONS

In representing the basic truth discovery method, there are some challenges to be faced. The relevant challenges and their appropriate solutions are as follows.

4.1 Clearing Noisy Input

The first problem handling is to clear the noisy input. The existing method deals only with structured data, but the proposed will work on unstructured text data. To get better performance, we have to convert the text into structured data.
Every process will be based on entity representation. Create a set of entities for eg: qs Є Qs for question text, which usually contains age, symptom, disease, drug etc., and the answer entity will normally be disease, drug, drug side-effect, etc., To execute the entity representation, we need to consider medical entity dictionary. If it contains the word in the question text, then the word will be placed into the entity set to get the answer text. If the doctor provides multiple answers, each answer is provided in separate entity set. The questions with similar meaning will be stated in the single entity set.

4.2 Multiple Answers Truth Extraction

The second problem is that there are many truth discovery answers for single question and they can be correlated with each other. For example, a patient describes his symptoms as "Cough , throat pain and fever " and asks for disease. The doctor says the answer as the disease he might have : " Common cold or Dengue". Both are possible and have many common symptoms. They are not independent answers. This can be formulated using many truth discovery methods [5]-[7] that contains single truth assumption such that there is one and only one truth answer for each question.
To find the correlation between multiple possible answers, the system use the neural word embedding method [8]-[10]. Each word is represented as real – value word vector. Here the vector representation of words can be formulated without syntax analysis or any manual labeling. Using this concept,

we can easily measure the similarity of words. If two words have similar meanings, then similarity vector will be high.
The correlation of words can be used to improve the answers trust worthiness. If "common cold" and "flu" are both considered as truth answers, then they are highly correlated. So the Equation (1) can be changed based on idea of " implication" [5],[11].

We can formulate the cosine similarity between answers into equation (1) as :

$$TD(a_{qs})=\sum_{ds\in DS} e_{ds}.\mathbb{I}(a_{qs},a_{qs}^{ds})+\sum a'_{qs} sim(v_{aqs},v_{a'qs})TD(a'_{qs}) \quad (3)$$

Where sim (v,v') is cosine similarity between two vectors.

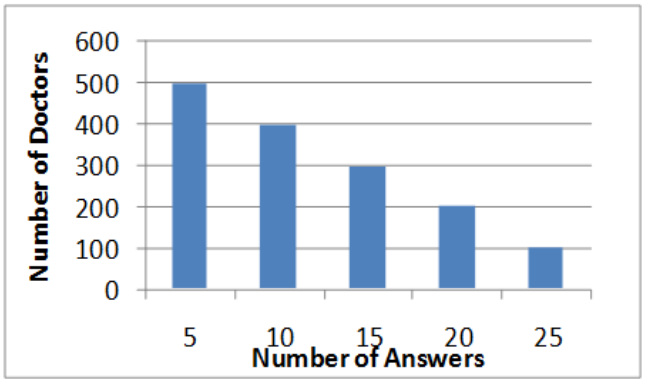
a'_{qs} another possible answer to the qs-th question. Thus the truth answer is mentioned if it is supported by other similar answers; else, if the answer is not supported or opposed by other answers, then cosine similarity gives negative value then the truth answer value is discounted.

4.3 long Tail Process

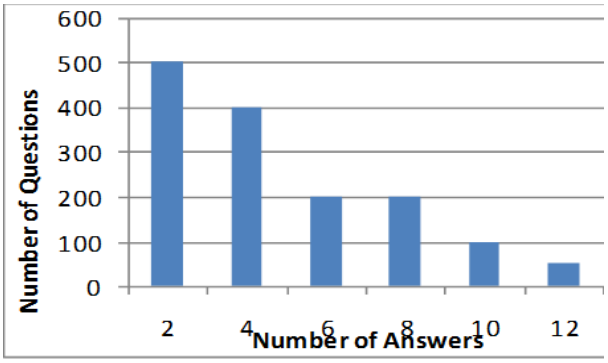
In medical query websites, some doctors can give answers to few questions and also some others give answers to many questions. And also the answers received will be small or large number. This long-tail process will be considered in truth discovery problem that was handled in [12], figure 2 clearly explains the long-tail process. Without sufficient answers for the questions, we can't accurately evaluate the doctors' expertise.
To handle the problem held by long-tail Process, Equation (2) can be modified based on [24]. The weights of these sources that give few answers will be discounted. Using this, add a pseudo count c_{pseudo} for each source,

$$e_{ds}=\log(1-\sum_{a\in v_{ds}} TD(a)/|v_{ds}|)+c_{pseudo} \quad (4)$$

If a doctor provides only a few answers, then C_{pseudo} will be $|v_{ds}|+c_{pseudo}$.So doctor score is low.
If he/she provides many answers, then $|v_{ds}|$ will be $|v_{ds}|+c_{pseudo}$ & his score will close to original estimation.
If he/she provides many answers, then $|v_{ds}|$ will be $|v_{ds}|+c_{pseudo}$ & his score will close to original estimation.
If he/she provides many answers, then $|v_{ds}|$ will be $|v_{ds}|+c_{pseudo}$ & his score will close to original estimation.



(a)Distribution of number of Answers per Doctor



(b)Distribution of number of Answers per Question

Algorithm: MKE System

Input: Set of questions QS and their answers $\{ a^{ds}_{qs}\}$ qs Є QS, ds Є DS, with an entity of (entity type, real-value);
Output: Find knowledge triples < question, diagnostic disease, truth discovery value > and doctors expertise $e_{ds;}$

1. Pre-processing : Separate entire text into words;
2. Create entity , for example : symptom in one entity from question text, disease in one entity from answer text;
3. Input creation : < { age, question entities }, answer entity, doctor ID>;
4. Initialize doctors expertise
5. repeat
6. Calculate equation (3) to find truth discovery value;
7. Estimate doctor expertise e_{ds} using equation (4);
8. until stop
9. Return founded knowledge triples < question, diagnostic disease, truth discovery value> and calculated doctor expertise (e_{ds})

V. EXPERIMENTAL RESULTS

5.1 Data Collection

All the datasets in this paper are collected from the medical query related website, "*icliniq.com*". Here the patients can ask their doubts related to health issues and the doctors can give suggestions for their queries. We collected the datasets for specific six topics and the number of questions and doctors involved are listed as Table 1.

Topic	No. of Questions	No. of Doctors Who answers
1	10,228	650
2	15,960	686
3	6,543	899
4	4,049	200
5	2,983	622
6	2,476	450

Table 1 : Statistics of Datasets

5.2 Evaluation of Doctors' expertise

Here we experimentally define the quantitative evaluation of doctor expertise. The "icliniq.com" website manages the profile for each doctor. Based on the registration and the replies by the patients' satisfaction, these website allocate the score for the doctor. The external information cannot clearly identify the doctors' expertise. The proposed system will be more powerful than the level score allotted by "icliniq.com" because it infers fine-grained topic for expertise doctor. Figure 3 shows the estimated doctor expertise on different topics .

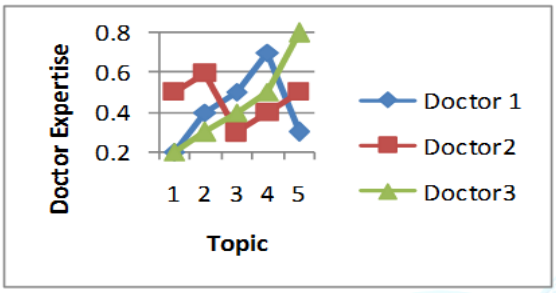


Figure 3 clearly shows that the doctors' expertise varies by topics. This confirms the necessity of fine-grained doctors' expertise estimation.

5.3 Case Study

There are different cases in various types of question intention. The question intention contains symptom – disease case, disease – drug case, disease – test case. In these cases because the symptoms and diseases are common, so that every viewer can easily understand the process. Table 3 shows Symptom – disease case. The patient with the age of 40 years have told his symptoms as headache and stuffed nose. The doctor suggested that his disease will be Chest cold with the probability of 0.235, Common cold with the probability of 0.288 or Flu with the probability of 0.247.

Symptom	Disease	Truth-discovery value
40 years old, Headache, Stuffed nose	Chest cold	0.235
	Common cold	0.288
	Flu	0.247

Table 3 : Symptom – Disease case

Disease	Drug	Truth Discovery value
25 years old gastritis	Omeprazole	0.456
	domperidone	0.722
	Cimetidine	0.112

Table 4 : Disease – Drug case

Disease	Clinical Test	Truth Discovery value
60 years old pulmonary heart disease	Chest x-ray	0.255
	Function test	0.117
	ECG examination	0.118

Table 5 : Disease – Test case

Age	Drugs to take
1-4	Pediatric Paracetamol
4-10	Pediatric Paracetamol
10-20	Amoxicillin
20-40	Amoxicillin, azithromycin
40-60	Amoxicillin,antibiotics
Above 60	Antibiotics, antiviral drug

Table 6: Age – Drug case for common cold

Table 6 Shows the drugs to take to cure common cold for patients with different ages. For children up to 10 years, the dosage is mild and safer. For age 10-40, the recommended drug is Amoxicillin. For above 60 years, the drug to take is antibiotics and antiviral drug. This shows that the patient's age is necessary for the process.

VI. CONCLUSIONS

The medical query related websites gives valuable health information. To gain knowledge from noisy input, we use medical knowledge Extraction (MKE) System. The MKE System evaluates knowledge triples < question, diagnostic disease, truth discovery value> and also estimates doctors' expertise. In this system, facing three challenges and evaluate solution for them for clearing the noisy input, the system use entity based representation. To evaluate multiple answers truth extraction using similarity function. To overcome long-tail process, using Pseudo count method.

VII.FUTURE ENHANCEMENT

In the existing system "Ask a Doctor" application have been implemented. For easy user access, the proposed system "Care for You" application has been introduced. In this application, if the patient gives the symptom, it automatically evaluates the disease the patients have, drugs to take for the treatment. This is one application planned to create using knowledge extraction and doctors expertise score. It has a great potential to benefit various real-world applications. It is planned to build more applications based on the MKE system in the future.

VIII. REFERENCES

[1] www.icliniq.com
[2] L. Nie, Y.-L. Zhao, M. Akbari, J. Shen, and T.-S. Chua, "Bridging the vocabulary gap between health seekers and healthcare knowledge,"
[3] L. Nie, M. Wang, L. Zhang, S. Yan, B. Zhang, and T.-S. Chua, "Disease inference from health-related questions via sparse deep learning," IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 8, pp. 2107–2119, 2015.
[4] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han, "A survey on truth discovery," arXiv preprint arXiv:1505.02463, 2015.
[5] X. Yin, J. Han, and P. S. Yu, "Truth discovery with multiple conflicting information providers on the web," IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 6, pp. 796–808, 2008.

[6] Q. Li, Y. Li, J. Gao, B. Zhao,W. Fan, and J. Han, "Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation," in Proc. of the ACM SIGMOD International Conference on Management of Data (SIGMOD'14), 2014, pp. 1187–1198.

[7] J. Pasternack and D. Roth, "Knowing what to believe (when you already know something)," in Proc. of the International Conference on Computational Linguistics (COLING'10), 2010, pp. 877–885.

[8] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Advances in Neural Information Processing Systems (NIPS'13), 2013, pp. 3111–3119.

[9] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in Proc. of the International Conference on Machine Learning (ICML'08), 2008, pp. 160–167.

[10] A. Mnih and G. E. Hinton, "A scalable hierarchical distributed language model," in Advances in Neural Information Processing Systems (NIPS'09), 2009, pp. 1081–1088.

[11] X. L. Dong, L. Berti-Equille, and D. Srivastava, "Integrating conflicting data: The role of source dependence," The Proceedings of the VLDB Endowment (PVLDB), vol. 2, no. 1, pp. 550–561, 2009.

[12] Q. Li, Y. Li, J. Gao, L. Su, B. Zhao, D. Murat,W. Fan, and J. Han, "A confidence-aware approach for truth discovery on long-tail data," The Proceedings of the VLDB Endowment (PVLDB), vol. 8, no. 4, pp.425–436, 2015.

