International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 04 | Apr 2019 www.irjet.net p-ISSN: 2395-0072

CLINICAL MEDICAL KNOWLEDGE EXTRACTION USING CROWDSOURCING TECHNIQUES

D Aswini

M.Tech, Department of Information Technology, College of Engineering Guindy, Tamil Nadu, India.

Abstract - *The medical website plays a vital role in today's digital world and a lot of forum is available for answering the queries provided by the user. In these kinds of crowd sourced question answering websites, there are a lot of patients and doctors involved. The most important challenge to find the quality and trustworthiness of the answer provided by the doctors. Also, the queries posted by users can be noisy and ambiguous. The quality of the answers provided by the doctors may vary due to reasons such as doctor's expertise, their level of commitment and their purpose of answering queries. To extract useful knowledge, required and relevant information has to be distinguished from extraneous and erroneous information. To solve this problem, an opinion finding system is proposed for extracting medical knowledge. It can be accomplished by two processes namely, medical term extraction and trustworthiness calculation. The process of extracting medical term includes stemming technique to remove the superfluous words from the user query and identify the precise medical term like symptoms. The process of calculating trustworthiness of an answer includes truth discovery framework to estimate precise answer with the factors like expertise, ethnicity and commitment level of doctors. The chatbot is developed as an additional feature for the user in which it predicts the disease based on the symptoms given by the users.*

Key Words: Crowd-sourcing, Trustworthiness Degree, Stemming Technique, Truth Discovery Framework.

1. INTRODUCTION

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. A recent report shows that tens of millions of health-related queries are searched every day using the Baidu search engine. Globally the online health service now becomes a billion-dollar industry. Many health service websites are developed, such as **questiondoctors.com**. The latter website has millions of registered users and hundreds of thousands of registered doctors.

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 crowdsourced question answering websites provide crowd-to-crowd service, so the crowd-generated information grows tremendously. One way to utilize such information is to extract knowledge from the medical question answering websites. The extracted knowledge can facilitate the development of many online health services.

1.1 Crowdsourcing

Crowdsourcing is a process through which a task, problem or project is solved and completed through a group of unofficial and geographically dispersed participants. Crowd sourcing is a joint process development or problem-solving technique that requires help from a network of people or crowd. This network is usually connected via the Internet or through a specific website. It refers to the outsourcing of a particular task to a wide range of people (quite often an online community) to get a job done more quickly and/or more efficiently. It is based on the idea that more hands or heads or experiences are better than one.

1.2 Medical Service Website

Medical Search Medicine Network is a reliable one-stop Internet medical service platform and one of the early platforms for exploring and practicing Internet medical services. After more than ten years of intensive cultivation, the medical search network uses the Internet, cloud computing, big data, intelligent medical equipment and other advanced technologies to integrate patients, doctors, hospitals and pharmaceutical companies to form and form Internet medical care.

1.3 Objective of the Project

The truth discovery method automatically extracts medical knowledge from crowdsourced question answering websites without any supervision. The opinion finding system is proposed to find out the best possible answer to the question post by the user. An approach based on the unsupervised model, which regards identifying correct trustworthy answer. The real world medical application is demonstrated based on the proposed method. This chatbot application predicts the disease based on the symptoms of the user.

In summary, contributions in this project are:

a) This system proposes a truth discovery method to automatically extract best opinion and find the truthness of the each response from the doctors by using the trustworthiness degree of the each answer provide by the doctor without any supervision.

b) The medical blog is maintained to provide the medical knowledge and opinion of all the medical case which is given by the doctors and correct possible solution will be display. So the user who has similar case can gain the clinical knowledge from it.

c) The proposed method is further demonstrated by building the chatbot which can suggest the possible disease by raised the certain questions and symptoms mapping is used for the final result for medical diagnosis.

2. LITERATURE SURVEY

A literature survey shows the varied analyses and analysis created associated with the project and also the results already printed, taking into consideration the varied parameters of the project and also the extent of the project. It is the most important part of the report as it gives a direction in the area of research. The goal of the literature survey is to fully specify the technical details associated with the most projects in an exceedingly compact and unambiguous manner.

2.1 Crowdsourcing Question Answering

The quality evaluation task as a classification problem and solves the task via supervised learning methods. The extract 13 non-textual features from each question-answer pair and train a maximum entropy model to classify the unlabeled question-answer pairs into three quality measurement levels, i.e., Bad, Medium and Good [1].

The feature construction and selection for the quality evaluation task are further studied: The authors ask a crowd of people to annotate the quality of question-answer pairs in terms of 13 predefined criteria, such as the novelty of the answers and the helpfulness of the answers. By analyzing these annotations, they find that the prole of answerer is quite useful for the quality evaluation task, which indicates the necessity of considering answerer expertise. The above work simply concatenates different features. However, textual and non textual features usually have different representation and the correlations between them are non-linear [4].

2.2 Truth Discovery

The main component in the MKE system belongs to the topic of truth discovery [4]. Truth discovery methods can automatically estimate source reliability (doctor expertise) from the data without any supervision and incorporate such estimated source reliability into the aggregation of noisy multi-source information. Most of truth discovery methods work on clean structured data. Recently, a person has begun to pay more attention on the noisy textual data [3].

2.3 Disease Inference System on the Basis of Health-Related Questions

The health is one of the increasing subjects used for assessing health condition among patients who suffer from specific ailment or diseases. It has been assumed that identification of the variables is able to mirror the ones overall health conditions [2]. The models describe the relationship between health variables using integrated model of inference system and linear regression. Linguistic data were collected by a guided interview and fed into the deep sparse inference system to yield health indices. The learning plan to surmise the conceivable maladies given the inquiries of well-being seekers. The proposed plan is embodied two key parts. The main part mines the discriminate therapeutic marks from crude elements. The second esteems the crude components and their marks as info hubs in one layer and concealed hubs in the consequent layer, individually. In the interim, it takes in the between relations between these two layers by means of pre-preparing with pseudo-marked information. Taking after that, the shrouded hubs serve as crude components for the more unique mark mining.

3. EXISTING SYSTEM

The young generations would also prefer to search the information readily available on the Web, or ask the doctors through the Internet. A recent report shows that tens of millions of health related queries are searched every day using Baidu search engine. The MKE system is build ahead a truth discovery framework, where it estimate trustworthiness of answers from the data without any supervision. The proposed truth discovery method is designed to tackle the new challenge in the medical knowledge extraction task, and the experimental results on real world dataset confirm its effectiveness.

4. PROPOSED SYSTEM

The proposed system, which extracts information from question-answer and summarizes into medical knowledge. After formally defining the task, introduce a basic truth discovery method and then propose solutions to address the unique challenges in medical knowledge extraction task.

4.1 System Design

An approach based on the unsupervised model, which regards identifying correct trustworthy answer which is illustrate Fig -1. To fulfill this aim, both questions and answers must be under deliberation. It is necessary to find the query and the domain in which the patient and the doctor is communicating. Separate login will be provided for patient, admin and doctors. They will be using their mailed and password. When a patient is verified he/she will be able to place their symptoms to the websites. This will be a vision to all doctors as well as patients. The answers will be posted by any doctors. Text analysis involves stemmer algorithm which removes the unnecessary words from the sentence and word analyzer are used to find medical words from the database. The assurance will be calculated using a trustworthy calculation. For quick medication advice, the user can also access the chatbot and get the medical diagnosis from the symptoms mentioned by them and disease is predicted by symptoms mapping from the trained database.

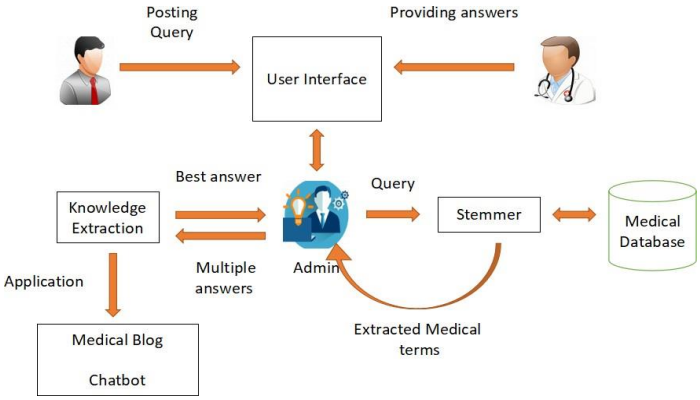

Fig -1: System Architecture for Medical Knowledge Extraction

5. METHODOLOGY

In this section, the technique details of the proposed MKE system, which extracts information from question-answer and summarizes into medical knowledge. After formally defining the task, a basic **truth discovery method**, and further demonstrate to the **chatbot application**.

5.1 Truth Discovery

To extract useful knowledge, it is important to distinguish relevant and correct information from unrelated or incorrect information. To solve this problem, a proposed scheme of 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 questions and answers websites. The MKE system is built upon a truth discovery framework, where estimate trustworthiness of answers and doctor expertise from the data without any supervision. The MKE system, which starts with the extraction of symptoms from the user queries and doctor response. These extracted entity is the key term to calculate the trustworthiness of the answers. Based on this calculation the highest trustworthy answer will be selected as the best opinion from the doctor and displayed in the blog. The Medical Knowledge Extraction is explained in detailed on the below following subsection.

5.1.1 Authentication and Posting Questions

Authentication is a process in which the credentials provided are compared to those on file in a database of authorized user's information on a local operating system or within an authentication server. If the credentials match, the process is completed and the user is granted authorization for access. Separate login will be provided for patients, admin and doctors. After the users are verified they can post their queries in the question answering websites.

5.1.2 Symptom Extraction

The word-based alignment model to perform monolingual word alignment, which has been widely used in many tasks such as collocation extraction and tag suggestion in practice, every sentence is replicated to generate a parallel corpus. To improve high-quality knowledge extraction technique used a word aligned database consists of medically related words which is shown in Fig -2.

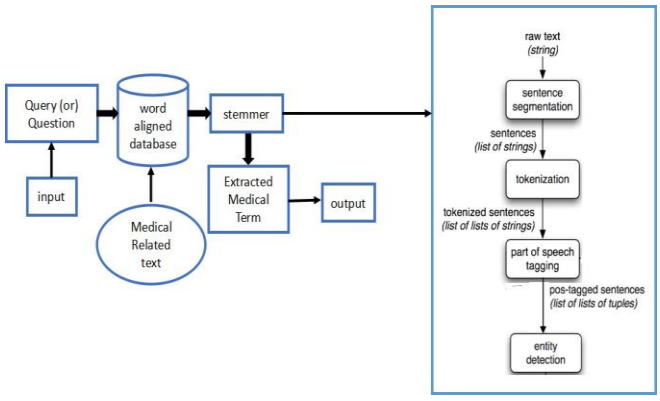

Fig -2: Flow Diagram of Stemming Process

The stemmer will remove all unwanted word. This extraction process is based on the stemming process. Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma. Stemming is important in Natural Language Understanding (NLU) and Natural Language Processing (NLP).

5.1.3 Trustworthy Calculation

Each doctor who answers the questions of a certain topic is associated with an expertise score that indicates his probability of providing trustworthy answers on this topic. As the system do not know the trustworthiness of answers.

$$T(x_q) = \sum_{d \varepsilon D} W_d . \P (x_q, x^d_q) \quad (1)$$

The above Equation 1 is used to calculate the trustworthiness of the answer. The notation of the equations is:

- $T(x_q)$ - trustworthiness degree.
- d - doctor.
- D - set of doctors.
- W_d - weight aggregate.
- x_q - answer for the Q^{th} question
- x^d_q - answer given by the D^{th} doctor for the Q^{th} question.
- ¶ - indication function.

The cosine similarity between two possible answers plays the role as a coefficient and automatically controls how much influence should be considered. Thus the trustworthiness of an answer is enhanced if it is supported by other similar answers; on the other hand, if an answer is not supported or even opposed by other answers, then the cosine similarity gives a negative value, so the answers trustworthiness is discounted. It helps find the other possible answer from the doctor response for the qth question, increase the weightage and combine it to the similar response which has high cosine value provide more accurate answer.

$$T(x_q) = \sum_{d \varepsilon D} W\, d . \P (x_q, x^d_q) + \sum_{y_q \neq x_q} Sim(Vx_q, Vy_q) \quad (2)$$

The Equation 2 notations are:

- $Sim(Vx_q, Vy_q)$ - cosine similarity between two vectors.
- y_q - another possible answer for the q^{th} question.

Evaluation of trustworthiness degree:

The trustworthiness degree of each doctor response. The Chart 1 shows the calculated trustworthiness degree for the sample set of data. The range of trustworthiness degree which is calculated by using Equation 1 and Equation 2.

In similar way the trustworthiness is calculated for other query and response. The highest trustworthy credit will be chosen as a best opinion and it will be displayed in the blog so that user can get the required medical knowledge for there queries.

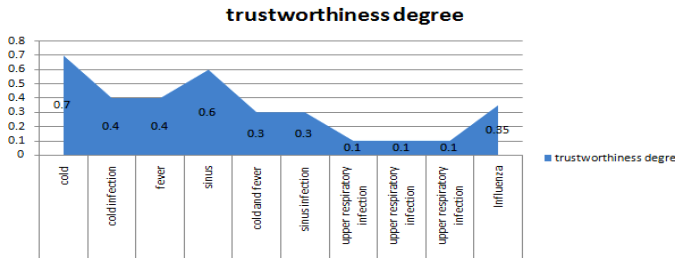

Chart -1: Trustworthiness Degree

Algorithm 1: Truth discovery

Input: Set of medical questions their corresponding answer an external entity dictionary with entity types and real-value vector representations of entities.
Output: Trustworthiness degree.

1. User authentication and posting questions in medical forum.
2. Extracted the symptoms form user queries and doctor response using stemming process.
3. An extracted symptom is used for calculating the trustworthiness degree.
4. Input tuple construction: form age, entities from question text, entity from answer text, doctor ID as input to truth discovery.
5. Calculate the trustworthiness degree of each answer using Equation 1 and Equation 2.
6. Estimate the trustworthiness degree of a possible answer for the q-th question.
7. Until stopping criterion is satisfied.
8. Return discovered knowledge triples question, diagnosis and trustworthiness degree.

5.1.4 Medical Blog

The proposed methodology of truth discovery uses the trustworthy value which is calculated using graph and it provides the highly possible solution to the user. So that the patient will be out of confusion as well as to get a good quality answers for their questions. These data will be maintained and available for the common user to get the medical knowledge if they have the similar medical case.

5.2 CHATBOT

The Chatbot is built based on the predefined case stored in the dictionary. When the user start the chat in bot emulator. The bot begins with the set of question like name, age, gender of user. The bot raise the question based on the symptoms which is specified by the user. This process helps

the bot to predict and diagnosis the disease. The user can also search for the disease which is unknown to them. The usage of this chatbot application is to predict the disease based on the symptom which is provided by the patients. The disease is identified by using symptoms matching technique. In general, patients will not get in touch with physicians or nurses or any medical professional with every one of their health questions but will turn to chatbot. This makes possible for the patients to get the instant suggestions about the disease.

6. CONCLUSION AND FUTURE WORK

The MKE system is developed to provide the medical knowledge from crowd sourcing data. Compared to the traditional one-to-one service, the online medical crowdsourced question answering websites provide crowd-to-crowd service, so the crowd-generated information grows tremendously. For example, questiondoctor.com alone receives thousands of new health-related questions every day. There is no doubt that the information from these crowdsourced question answering websites is valuable, but how to make good use of such information is a big question. One way to utilize such information is to extract knowledge from the medical question answering websites. The extracted knowledge can facilitate the development of many online health services. For example, the knowledge can contribute to the construction of a robot doctor who can automatically generate answers to new health-related questions. The medical crowdsourced question answering websites provide valuable but noisy health-related information. To extract high-quality medical knowledge from the question-answer pairs, a MKE system is proposed. The summarisation of the proposed system is that the system model starts up with the user registration to give access to the medical user. The user can query for the medical issue and get the key to extract the query result. The system predicts the results based on the MKE process which is calculated based on the knowledge triple and chatbot for the disease diagnosis. The future work of this system is that the system has to enhance so that the voice input data can also be accepted. By this, all type of user can use the system. The chatbot can be enhanced using Artificial Intelligence (AI) techniques and it can automatized for providing the diagnosis results and question routing can be added so that the user can get more accurate results.

REFERENCES

[1] Chirag Shah and Jefferey Pomerantz. “Evaluating and predicting answer quality in community QA”. In the Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval, pp.411-418, 2010.

[2] Eugene Agichtein, Aristides Gionis, Carlos Castillo, Debora Donato and Gilad Mishne. “Finding high - quality content in social media”. In the Proceedings of the ACM International Conference on Web Search and Data Mining, pp.183 - 194, 2008.

[3] Jeff Pasternack and Dan Roth. “Making better informed trust decisions with generalized fact-finding”. In the Proceedings of the International Joint Conference on Artifical Intelligence, pp.2324-2329, 2011.

[4] Jiwoon Jeon, W. Bruce Croft, Joon Ho Lee and Soyeon Park. “A framework to predict the quality of answers with non-textual features”. In the Proceedings of ACM SIGIR International Conference on Research and Development in Information Retrieval, pp.228-235, 2006.

[5] Paresh Karande, Manoj Khairavkar, Vinayak Lokhande, Khemchand Mahajan, Rupali Umbare and Reeta Kamble. “Disease Inference System on the basis of Health-Related Questions by using Sparse Deep Learning”. In the Transactions of IOSR Journal of Computer Engineering, Vol.3, pp.91-97, 2016.

[6] X. Yin, J. Han and P. S. Yu. “Truth discovery with multiple conflicting information providers on the web”. In the Transactions of IEEE Transactions on Knowledge and Data Engineering, Vol.20, pp.796-808, 2008.

[7] Yaliang Li, Chaochun Liu, Nan Du, Wei Fan, Qi Li, Jing Gao, Chenwei Zhang and Hao Wu. “Extracting Medical Knowledge from Crowdsourced Question Answering Website”. In the Transactions of IEEE Big Data, 2016.