Diagnosis of Various Diseases Using Big Data Extraction From Question Answering Website

Bhalerao Akash.[1], Fulsundar Ajinkya.[2], Walunj Amol[3]
[1, 2, 3]Dept of Computer Engineering
[1, 2, 3]Jaihind Collage of Engineering, Kuran, Pune, India

Abstract- *The medical crowd sourced question answering (Q&A) websites are booming in recent years, and increasingly large amount of patients and doctors are involved. The valuable information from these medical crowd sourced Q&A websites can benefit patients, Doctors and the society. One key to unleash the power of these Q&A websites is to extract medical knowledge from the noisy question-answer pairs and filter out unrelated or even incorrect information. Facing the daunting scale of information generated on medical Q&A websites every day, It is unrealistic to fulfill this task via supervised method due to the expensive annotation cost. In this system, 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. 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, Ask A Doctor, which can automatically give patients suggestions to their questions.*

Keywords- Crowd sourced Question Answering, Medical Knowledge Extraction, and Truth Discovery.

I. INTRODUCTION

Recently, the Big Data challenge is motivated by a dramatic increase in our ability to extract and collect data from the physical world. One of the important property of Big Data is its wide Variety, i.e., data about the same object can be obtained from various sources. For example, customer information can be found from multiple databases in a company, a patient's medical records may be scattered in different hospitals, and a natural event may be observed and recorded by multiple laboratories. Due to recording or transmission errors, device malfunction, or malicious intent to manipulate the data, data sources usually contain noisy, outdated, missing or erroneous records, and thus multiple sources may provide conflicting information. In almost every industry, decisions based on untrustworthy information can cause serious damage.

For example, erroneous account information in a company database may cause financial losses; wrong diagnosis based on incorrect measurements of a patient may lead to serious consequences; and scientific discoveries may be guided to the wrong direction if they are derived from incorrect data. Therefore, it is critical to identify the most trustworthy answers from multiple sources of conflicting information. This is a non-trivial problem due to the following two major challenges.

To better cater to health seekers, a growing number of community-based healthcare services have turned up, including HealthTap2, HaoDF3 and WebMD4. They disseminate personalized health knowledge and connecting patients with doctors worldwide via question answering. These forums are very attractive to both professionals and health seekers. For professionals, they are able to increase their reputations among their colleagues and patients, strengthen their practical knowledge from interactions with other renowned doctors, as well as possibly attract more new patients. For patients, these systems provide nearly instant and trusted answers especially for complex and sophisticated problems.

In many cases, the community generated content, however, may not be directly usable due to the vocabulary gap. Users with diverse backgrounds do not necessarily share the same vocabulary. Take Health-Tap as an example, which is a question answering site for participants to ask and answer health-related questions. The questions are written by patients in narrative language. The same question may be described in substantially different ways by two individual health seekers. On the other side, the answers provided by the well-trained experts may contain acronyms with multiple possible meanings, and no standardized terms.

II. PROBLEM STATEMENT

To extract the data present over those websites and estimate the possible accuracy rate of those answers given to any question related to health. We also have to estimate the expertise of the doctors which will be useful for patients.

III. LITERATURE SURVEY

In this paper, we consider how to find true values from conflicting information when there are a large number of sources, among which some may copy from others. [1]

We introduce a framework for incorporating prior knowledge into any fact- finding algorithm, expressing both general "common-sense" reasoning and specific facts already known to the user as first-order logic and translating this into a tractable linear program .[2]

In this paper we propose a new problem called Veracity, i.e., conformity to truth, which studies how to find true facts from a large amount of conflicting information on many subjects that is provided by various web sites.[3]

This paper presents a scheme to label question answer(QA) pairs by jointly utilizing local mining and global learning approaches. Local mining attempts to label individual QA pair by independently extracting medical concepts from the QA pair itself and mapping them to authenticated terminologies. [4]

This paper presents a novel scheme to code the medical records by jointly utilizing local mining and global learning approaches, which are tightly linked and mutually reinforced.[5]

Due to recording or transmission errors, device malfunction, or malicious intent to manipulate the data, data sources usually contain noisy, outdated, missing or erroneous records, and thus multiple sources may provide conflicting information. In almost every industry, decisions based on untrustworthy information can cause serious damage. For example, erroneous account information in a company database may cause financial losses; wrong diagnosis based on incorrect measurements of a patient may lead to serious consequences; and scientific discoveries may be guided to the wrong direction if they are derived from incorrect data. Therefore, it is critical to identify the most trustworthy answers from multiple sources of conflicting information. This is a non-trivial problem due to the following two major challenges.[6]

IV. 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.

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. Based on these outputs, various real-world applications can be built.

V. OBJECTIVES

- To conduct the medical knowledge extraction from the Question - Answering websites.
- To estimate the expertise of doctors without any supervision.
- Build a medical robot.
- To provide the guidance to the patients.
- To provide the accurate data.

VI. SYSTEM ARCHITECTURE

Following figure shows the architecture of the system :-

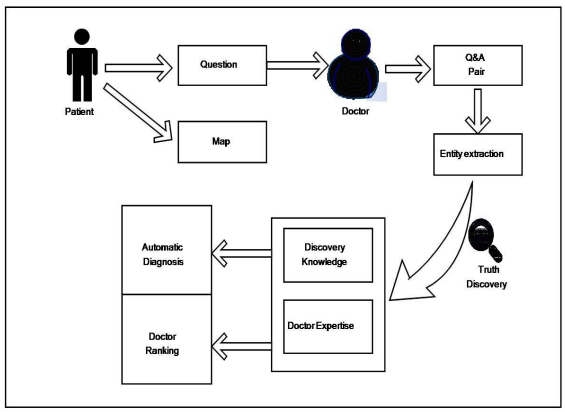

Fig.1 System Architecture

Proposed system mainly consists of three modules :-

- Patient
- Doctor
- Admin

Patients:

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.

Doctors:

A doctor is a person who answers questions on the medical Q&A websites. On the website from which we crawl the data, the "doctors" are real doctors, though it may not be this case for other websites.

Administrator:

As a website admin we would be liable for making sure the site's user interface is easy to understand and efficient. We would ensure that all websites are operating securely and at optimum speeds. We will likely be responsible for evaluating each website's analytics, such as user feedback and traffic.

VII. MATHEMATICAL MODEL

System description:

- Let S be the whole system,

S = I,P,O
I-input, P-procedure, O- Output

- I = UI, UP, CQA, MKE

I0 = UI (User id)
I1 = UP (User password)
I2 = CQA (Crowdsourced Question Answering.)
I3 = MKE (Medical Knowledge Extraction.)

- P = P0, P1, P2, P3...

P0 = Pre-processing: segment text to words.
P1 = Entity extraction: extract one type of entity.
P2 = Calculate the trustworthiness degree of each answer.
P3 = Initialize doctors expertise uniformly.

- O = (S)

S = Successfully algorithm implementation and proper input.

– Input: User Id, Password, Questions.

– Output: Detect the diseases, Find out the probability of questions, Increased the Doctors ratings.

– Functions : User login, User registration, User Information

– Success condition : Successfully algorithm implementation and proper input.

– Failure conditions: Hardware failure, Software failure. Huge database can lead to more time consumption to get the information.

Result Analysis

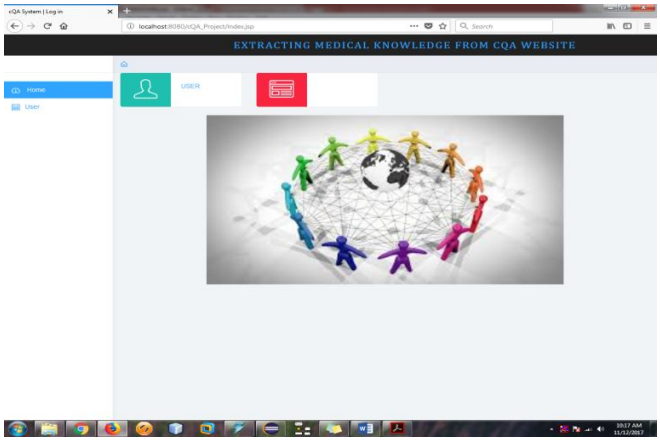

Fig 2. Homepage

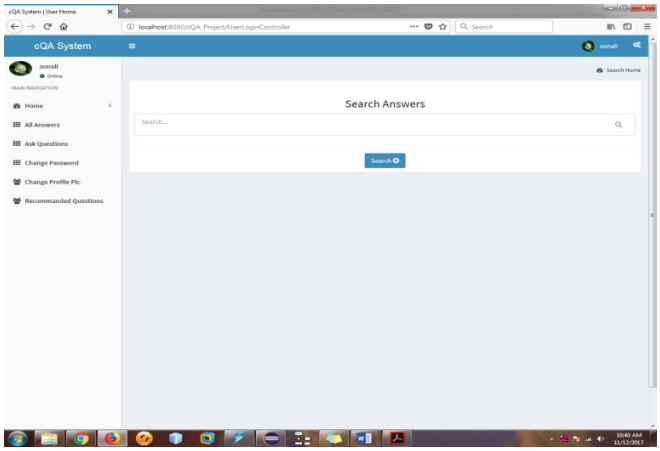

Fig 3. User homepage

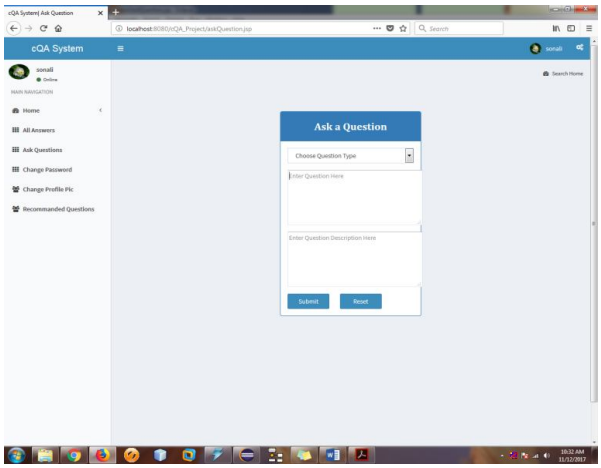

Fig 4. User Ask Question Form:

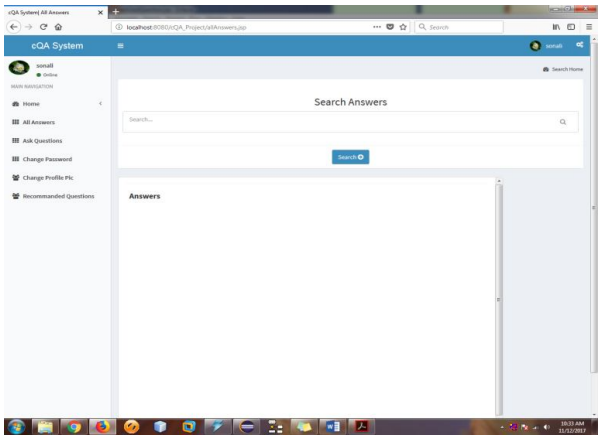

Fig 5. User All Answers Form:

VIII. CONCLUSIONS AND FUTURE WORK

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, Medical Knowledge Extraction (MKE) system is proposed in this paper. Free advertisement of the doctors so it will beneficial to them for gaining popularity. Medical robot itself give the answer automatically based on its analysis.

REFERENCES

[1] X. Yin, J. Han, and P. S. Yu, "Truth discovery with multiple conflicting information providers on the web," in SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'07), 2007.

[2] X. L. Dong, L. Berti-Equille, and D. Srivastava, "Integrating conflicting data: The role of source dependence," The Proceedings of the VLDB Endowment (PVLDB),2009.

[3] 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.

[4] L. Nie, M. Akbari, T. Li, and T.-S. Chua, "A joint local-global approach for medical terminology assignment," in SIGIR 2014

[5] L. Nie, Y.-L. Zhao, M. Akbari, J. Shen, and T.-S. Chua, "Bridging the vocabulary gap between health seekers and healthcare knowledge," IEEE Transactions on Knowledge and Data Engineering,2015.

[6] Yaliang Li, Chaochun Liu, Jing Gao, Qi Li, Nan Du,Wei Fan Extracting Medical Knowledge from Crowdsourced Question Answering Website 2016 IEEE