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

Bhalerao Akash P [1], Fulsundar Ajinkya A [2], Walunj Amol S [3],
Prof. Jadhav. N. S.[4]

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.

Index Terms: 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.LITERATURE SURVEY

In this paper each medical record is coded with multiple terminologies with local mining, which are generated via mapping their embedded medical concepts to terminologies.

However, these mined terminologies may suffer from various problems. [2]

In this paper, we define medical concepts as medical domain-specific noun phrases, and medical terminologies as authenticated phrases by well-known organizations that are used to accurately describe the human body and associated components, conditions and processes in a science-based manner. Even though some health communities have recently suggested doctors to annotate their answers with medical concepts, we cannot ensure that they are medical terminologies. [5]

Many studies have been there on ranking web pages according to authority based on hyperlinks, such as Authority-Hub analysis , Page Rank , and more general link-based analysis . But does authority or popularity of web sites lead to accuracy of information? The answer is unfortunately no. For example, according to study the bookstores ranked on top by Google (Barnes & Noble and Powell’s books) contain many errors on book author information, and some small bookstores provide more accurate information.[6]

This system proposes the most basic approach is to take a vote: if multiple claims are mutually exclusive of each other, select the one asserted by the most sources. In our experiments, sources will be the authors of the document containing the claim, but other sources could be publishers/websites (when no authorship is given), an algorithm that outputs claims, etc. Although sometimes competitive, we found voting to be generally lackluster. [8]

In this paper, the Big Data challenge is motivated by a dramatic increase in our ability to extract and collect data from the physical world. One important property of Big Data is its wide variety, i.e., data about the same object can be obtained from various sources.[9]

In this paper, we have seen that how can adapt the existing data fusion techniques to more challenging area of automatically constructing large-scale knowledge bases. To build a knowledge base, we employ multiple knowledge extractors to extract (possibly conflicting) values from each data source for each data item; we then need to decide the degree of correctness of the extracted knowledge. [11]

This paper proposes practical data integration systems, it is common for the data sources being integrated to provide conflicting information about the same entity. Consequently, a major challenge for data integration is to derive the most complete and accurate integrated records from diverse and sometimes conflicting sources. [12]

III.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.

IV.OBJECTIVE

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

V.PROPOSED SYSTEM

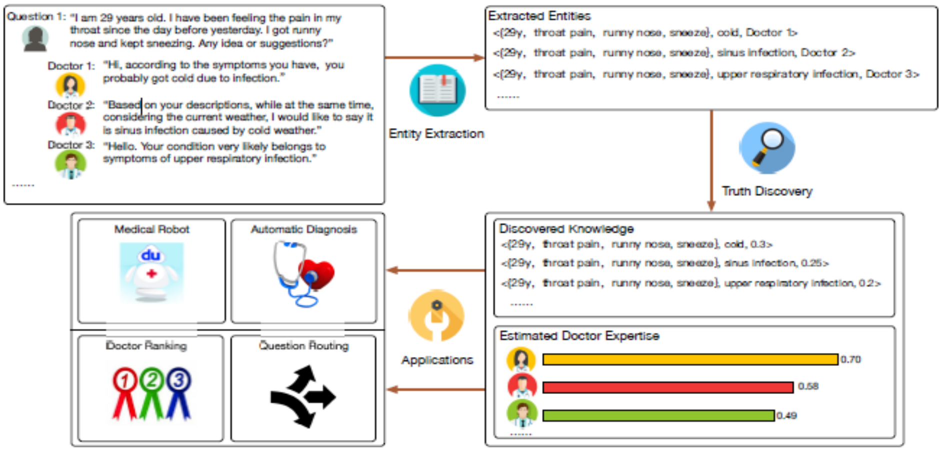
Question 1: "I am 29 years old. I have been feeling the pain in my throat since the day before yesterday. I got runny nose and kept sneezing. Any idea or suggestions?"
Doctor 1: "Hi, according to the symptoms you have, you probably got cold due to infection."
Doctor 2: "Based on your descriptions, while at the same time, considering the current weather, I would like to say it is sinus infection caused by cold weather."
Doctor 3: "Hello. Your condition very likely belongs to symptoms of upper respiratory infection."
......
Entity Extraction
Extracted Entities
<{29y, throat pain, runny nose, sneeze}, cold, Doctor 1>
<{29y, throat pain, runny nose, sneeze}, sinus infection, Doctor 2>
<{29y, throat pain, runny nose, sneeze}, upper respiratory infection, Doctor 3>
......
Truth Discovery
Discovered Knowledge
<{29y, throat pain, runny nose, sneeze}, cold, 0.3>
<{29y, throat pain, runny nose, sneeze}, sinus infection, 0.25>
<{29y, throat pain, runny nose, sneeze}, upper respiratory infection, 0.2>
......
Estimated Doctor Expertise
0.70
0.58
0.49
......
Applications
Medical Robot
du
Automatic Diagnosis
Doctor Ranking
1
2
3
Question Routing

Figure. System Architecture

Proposed system mainly consists of four 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.

VI. 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, 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

Fulsundar Ajinkya A:

Department of Computer Engineering, Jaihind College of Engineering, kuran, Maharastra, India.

Bhalerao Akash P:

Department of Computer Engineering, Jaihind College of Engineering, kuran, Maharastra, India.

Walunj Amol S:

Department of Computer Engineering, Jaihind College of Engineering, kuran, Maharastra, India.

Prof. Jadhav N.S.:

Prof Of Department of Computer Engineering, Jaihind College of Engineering, kuran, Maharastra, India.

Her area of interests is Computer programming languages, Computer Network, Data structure, Software Engineering.