ISSN (Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 1, January 2018

A Survey on Knowledge Extraction from Health Related Data Using Machine Learning Approach

Swati Ramesh Kalore, Prof . V. S. Mahalle

Dept. of Computer Science and Engineering, SSGMCE, Shegaon, Maharashtra, India

ABSTRACT: In the system 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. In this 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 paper, 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, Truth Discovery.

I. INTRODUCTION

The young generations would prefer to search the information readily available on the Web, or ask the doctors through the Internet. 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 service, so the crowd-generated information grows tremendously. In this system any patients ask question and many doctors give an answer on that question. Then the system checks the truth discovery of that answer using entity extraction. In the previous, valuable information of crowd-sourced Q&A websites how to use in real time is the big question. Therefore we extract knowledge from such information and use to development of many online health services. From that knowledge we can construct the robot doctor for answering the new health related questions. The main challenge of extracting knowledge from the medical crowd-sourced Q&A websites is that the quality of question-answer pairs is not guaranteed. Therefore I develop unsupervised learning methods that extract knowledge from noisy crowd-sourced data on medical Q&A websites. The key component in the medical knowledge extraction task is to find reliable answers without any supervision. Some existing truth discovery methods assume that the answers from high-expertise users are reliable, and the users who provide reliable answers should have high expertise. Thus these methods can iteratively estimate the source reliability (i.e., user expertise) and infer reliable answers from user-contributed data. Therefore I develop unsupervised learning methods that extract knowledge from noisy crowd-sourced data on medical Q&A websites. The key component in the medical knowledge extraction task is to find reliable answers without any supervision. Truth discovery methods are developed for

ISSN (Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 1, January 2018

structured data (e.g., relational database), but for crowd-sourced Q&A websites, all the inputs are unstructured data (i.e., texts). We address this challenge by transforming unstructured texts to entity-based representation.

Following are the objectives of the MKE system:

i. Find reliable answer using MKE(Medical Knowledge Extraction) system.
ii. Transforming unstructured texts to entity-based representation.
iii. Evaluate the quality of question-answer pairs without any supervision.
iv. 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.

II. LITERATURE SURVEY

Sr No.	Paper Name/ year	Author Name	Proposed System/ Algorithms	Advantage and Disadvantage
1.	**"Bridging the vocabulary gap between health seekers and healthcare knowledge," Year-2015.**	**L. Nie, Y.-L. Zhao, M. Akbari, J. Shen, and T.-S. Chua**	A novel scheme that is able to code the medical records with corpus-aware terminologies	Advantage : holds potential to large-scale data Disadvantage : the problem of information loss and low precision due to the possible lack of some key medical concepts in the medical records
2.	**"Disease inference from health-related questions via sparse deep learning," year-2015.**	**L. Nie, M. Wang, L. Zhang, S. Yan, B. Zhang, and T.-S. Chua**	The discriminate medical signatures from raw features	Advantage: build a disease inference scheme that is able to automatically infer the possible diseases of the given questions in community-based health services. Disadvantage: unable to directly compare with the target values.
3.	**"Truth discovery with multiple conflicting information providers on the web," year-2007**	**X. Yin, J. Han, and P. S. Yu**	Find true facts from a large amount of conflicting information on many subjects that is provided by various websites.	Advantage: identifies trustworthy websites better than the popular search engines. Disadvantage: we cannot compute the trustworthiness of a website by adding up the weights of its facts
4.	**"Integrating conflicting data: The role of source dependence," Year-2009.**	**X. L. Dong, L. Berti-Equille, and D. Srivastava**	To find true values from conflicting information whenthere are a large number of sources, among which some may copy from others.	Advantage: detects dependence and discovers truth from conflicting information. Disadvantage: did not study how to discover the dependence
5.	**"Knowing what to believe (when you already know something)," Year-2010.**	**J. Pasternack and D. Roth**	Incorporating prior knowledge into any factfinding algorithm, expressing both general "common-sense"	Advantage: if multiple claims are mutually exclusive of each other, select the one asserted by the most sources. Disadvantage: Fact-finding algorithm typically does not have meaningful

ISSN (Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 1, January 2018

			reasoning	semantics.
6	**"Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation," year-2014**	**Q. Li, Y. Li, J. Gao, B. Zhao,W. Fan, and J. Han**	To resolve conflicts among multiple sources of heterogeneous data types.	Advantage: find out both truths and source reliability Disadvantage: the problem in minimize the overall weighted deviation between the truthsand the multi-source observations..
7	**"A survey on truth discovery," Year-2015.**	**Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han**	Focus on providing a comprehensive overview of truth discovery methods, and summarizing them from dif- ferent aspects.	Advantage: eliminate conicts amongmulti-source data is to conduct majority voting or averaging. Disadvantage: selection depends on the speci_c application scenarios..
8	**"On truth discovery in social sensing: A maximum likelihood estimation approach," Year-2012.**	**D.Wang, L. Kaplan, H. Le, and T. Abdelzaher**	Truthfulness of Deep Web data in two domains where we believed data are fairly clean and data quality is important to people's lives: Stock and Flight	Advantage: data of quite high inconsistency and found a lot of sources of low quality. Disadvantage: did not find one method that is consistently better than others.
9	**"Truth finding on the deep web: Is the problem solved?" Year-2012.**	**X. Li, X. L. Dong, K. B. Lyons, W. Meng, and D. Srivastava**	Important interaction between the credibility of a statement, the trustworthiness of the user making that statement and the language used in the post containing that statement.	Advantage: retrieve the credibility label of unobserved statements given some expert labeled statements Disadvantage: does not specifically point to the occurrence of any side effect
10	**"People on drugs: credibility of user statements in health communities," Year- 2014.**	**S. Mukherjee, G. Weikum, and C. Danescu-Niculescu-Mizil**	Finds truth discovery From noisy social sensing data.	Advantage: returns the best guess regarding the correctness of each measurement. Disadvantage: the data are "incomplete" or the likelihood function involves latent variables.

III. PROPOSED SYSTEM

In the proposed system, a MKE (Medical Knowledge Extraction) system, conduct the medical knowledge extraction and doctor expertise estimation without any supervision. In medical crowd-sourced Q&A websites, patients ask questions with some intentions. For example, patients want to ask the disease related to their symptoms, or side effects of drug. Doctors provide answers to these questions those are play the essential role in the Q&A websites. Many

ISSN (Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 1, January 2018

doctors may give different answers due to their diverse expertise for the same question. The truth discovery method also proposed in proposed system and this method use 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 representation of truth discovery method, which outputs the medical knowledge triples <question, diagnosis, trustworthiness degree> and the estimated doctor expertise. Various real-world applications can be built on the based on these outputs. 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.

Advantages of Proposed System

1. The proposed system automatically generates doctors' expertise.
2. The proposed system is easily flexible in real time.

IV. PROPOSED SYSTEM IMPLEMENTATION

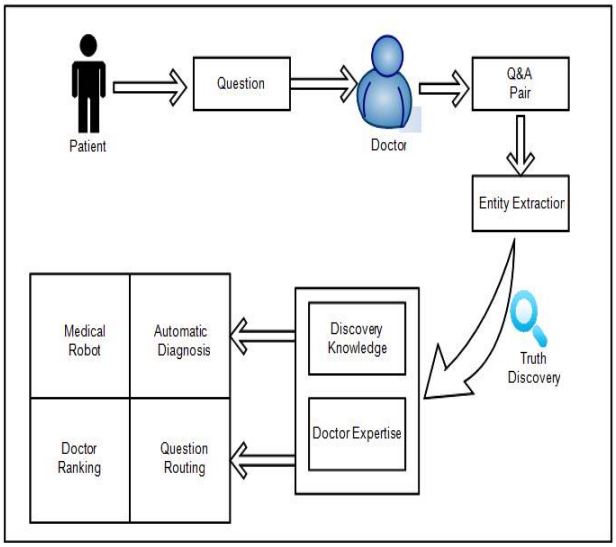

Fig 1: System Architecture

A MKE (Medical Knowledge Extraction) system, conduct the medical knowledge extraction and doctor expertise estimation without any supervision. In medical crowd-sourced Q&A websites, patients ask questions with some intentions. The MKE system extracts the information from noisy data that is Q/A pairs and summarizes into medical knowledge. After that find the truth discovery method for find the truth answer. For that design the 3 module that is Patient, Doctor, and Admin. The patients are asking question and see the answers. Doctors give the answers to the patient's questions. And the admin is work on truth discovery method, checks the truth answer and see to user. And admin also see the all patient's data and the doctor's data.
The MKE system begins with the preprocessing of the text. Then based on an available entity dictionary, MKE system conducts entity extraction. In that step, different types of entities are extracted for different tasks. For example, if the question contains statements of symptoms, then the patient is asking for the possible disease, so the disease type will be extracted from the answer text. MKE system also extracts the patients' age from the question text and incorporates it

ISSN (Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 1, January 2018

into the representation. Next, MKE system constructs the tuples <{age, entities from question text}, entity from answer text, doctor ID> as the input to the truth discovery approach. Finally, MKE system applies the proposed truth discovery method to build knowledge triples <question, diagnosis, trustworthiness degree>and compute the doctor expertise $\{w_d\}$

V. RESULT DISCUSSION

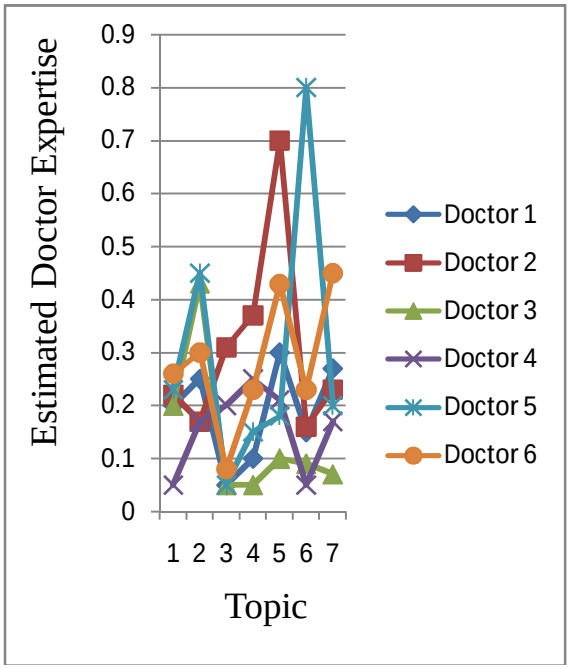

Figure clearly show that the estimated doctor expertise scores are quite different over topics. Means doctor cannot be experts in all topics, some doctors are experts in some topic and other also. For example, doctor 1 might be an expert on topic 5, but not on other topics. Doctor 5 has a high expertise score on topic 7, while his expertise scores on other topics are quite low. These observations confirm the intuition about the necessity of fine-grained doctor expertise estimation.

VI. CONCLUSION

The medical crowd-sourced Q&A websites provide valuable information but noisy health related information. Medical Knowledge Extraction (MKE) system is proposed, to extract high quality medical knowledge from the question-answer pairs. 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. Using the entity-based representation to remove noisy text input and merge similar questions; 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. In 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

ISSN (Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 1, January 2018

REFERENCES

[1] 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, vol. 27, no. 2, pp. 396–409, 2015.
[2] 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.
[3] X. Yin, J. Han, and P. S. Yu, "Truth discovery with multiple conflicting information providers on the web," in Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'07), 2007, pp. 1048–1052.
[4] 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.
[5] 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.
[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] 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.
[8] D.Wang, L. Kaplan, H. Le, and T. Abdelzaher, "On truth discovery in social sensing: A maximum likelihood estimation approach," in Proc. of the International Conference on Information Processing in Sensor Networks (IPSN'12), 2012, pp. 233–244.
[9] X. Li, X. L. Dong, K. B. Lyons, W. Meng, and D. Srivastava, "Truth finding on the deep web: Is the problem solved?" The Proceedings of the VLDB Endowment (PVLDB), vol. 6, no. 2, pp. 97–108, 2012.
[10] S. Mukherjee, G. Weikum, and C. Danescu-Niculescu-Mizil, "People on drugs: credibility of user statements in health communities," in Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'14), 2014, pp. 65–74.