Framework for Disease Prediction from Symptoms and Health related data

Miss Swati Y. Dugane (ME, CSIT)
HV PM'S Amravati
Maharashtra, India.

Prof. Karuna G. Bagde
HVPM'S Amravati
Maharashtra, India.

ABSTRACT

Diseases tracing plays important role in life. Every one cares about health. According to some social study, lot of people spend time on online searching of health related issues. Normally, people use Google to search their queries and that search engine respond them with the answer but that answer is in scattered format. So user usually ends up with a bulk of information from which it becomes difficult to make a sense out of that whole chunk. A lot of work has already been done related to the information needs of health seekers in terms of question-answers or in a flowchart manner so as to lead the user towards diagnosis of his/her ailments based on symptoms manifested. In this dissertation, further restructuring of the question and answer has been done in order to get most appropriate answer of a query. A tag mining framework for health information seekers will be used, aimed to analyze the symptoms in detail; some specific questions may be asked by system to the user to further elaborate the symptoms so as to arrive at a particular disease/s or to rule out a particular disease/s. Once the system suggests the possibility of a disease/s, user will be then directed to some basic investigations to arrive at a more precise symptomatic provisional diagnosis. System will also give detail information of the disease/s in question like epidemiology, causes, symptoms and signs, diagnostic aids, probable treatment, precautions etc. to help user to understand the disease condition. After providing this basic information, system will suggest the concerned specialty/superspeciality (eg. Endocrinologist for diabetes, gastroenterologist for intestinal disorder) best suited for the disease condition in question. Once the user decides the speciality, he/she will be provided with the choices of all available doctors/hospitals/clinics, including government, charitable, private located in and around his/her locality suited to his/her requirement, timing, affordability etc. for further course of action.

Keywords - Tag Mining, health seeker, Datasets, Symptoms.

I) INTRODUCTION

Recent years have seen a flourishing of community-driven question answering (cQA) Portals, which have emerged as an effective paradigm for disseminating diverse knowledge, seeking precise information, and locating outstanding expert. Around 40% of the questions in the emerging social-oriented question answering forums have at most one manually labeled tag, which is caused by incomprehensive question understanding or informal tagging behaviors. Information extraction from medical text is the basis for other higher-order analytics, such as representation, classification, and clustering. However, accurately and efficiently inferring disease information is non-trivial, especially for community-based health services due to the

incomplete information, correlated medical concepts, and limited high quality training samples.

To solve such problems of incomplete information and correlated medical concepts, the proposed dissertation will develop the scheme which studies the user information and health related data. It will infer a deep learning of the possible diseases suggested on the basis of questions/symptoms of health seeker [1].The prime intention of deep learning comprises of two key components, the first globally mines the discriminant medical signatures from raw features. The raw features and their signatures serve as input nodes in one layer and hidden nodes in the subsequent layer, respectively. The second learns the inter-relations between these two layers via pre-training. With incremental and alternative repeating of these two components, our scheme builds a sparsely connected deep learning architecture with three hidden layers .In this, the user request will be compared with the different datasets. The datasets will be created using the information taken from authenticated medical literature (like well known textbooks, guidelines issued by authorized medical bodies and consultation with experts in the field). The system will save the medical information in the form of datasets and infer the possible diseases on the basis of questions/symptoms of health seeker.

II) BACKGROUND

Information extraction from medical text is the basis for other higher-order analytics, such as representation, classification, and clustering. In 2007, Y. Zhang and B. Liu present a paper on "Semantic text classification of disease reporting [4] trained an infectious disease model with the sentence-level semantic features, and obtained promising performance.

In 2008, the research on online health data is relatively rare Luo et Al. [5] built a medical Web search engine called iMEd, which employed medical knowledge and an interactive questionnaire to help searchers form queries.

In 2010 the work by S. Doan and H. Xu in , "Recognizing medication related entities in hospital discharge summaries using support vector machine," [6] utilized SVM to recognize the medication related entities in hospital discharge summaries, and classified these atomic elements into pre-defined categories, such as treatments and conditions.

In 2011, a more systematic evaluating framework for medical record search was developed in [7] using BLULab de-identified medical corpus.

In 2013, LiqiangNie and Tat-Seng presented a technique in "Beyond Text QA: Multimedia Answer Generation by Harvesting Web Information" [8]. It describes that Community question answering (cQA) services have gained popularity over the past years. It not only allows community members to post and answer questions but also enables general users to seek information from a comprehensive set of well-answered questions. However, existing cQA forums usually provide only textual answers, which are not informative enough for many questions.

III) EXISTING SYSTEMS

The D. A. Davis, N. V. Chawla, N. Blumm, N. Christakis, and A.-L. Barabasi presented the paper on "Predicting individual disease risk based on medical history," a novel system named CARE was designed in [9], which combined collaborative filtering methods with clustering to predict each patient's greatest disease risks based on their own medical history and those of similar patients. Understanding how the disease progress is of essential importance for proactive healthcare. The

paper "A Bayesian learning approach to promoting diversity in ranking for biomedical information retrieval," by X. Huang and Q. Hu, described Medical retrieval is the dominant way for knowledge exchanging and sharing. Huang et al. [10] proposed a re-ranking model for promoting diversity in medical search. Query-adaptive weighting methods that can dynamically aggregate and score various search results.

IV) ANALYSIS AND DISCUSSION

1. The previous used methodologies were unable to get exact answer because of incomplete information or insufficient dataset.

2. Because of bulky dataset, the system performance slows down.

3. Problem occurs in the association of datasets i.e. maintaining relationship between the dataset is a difficult task.

4. Internet connection is required.

5. The system is not able to identify the multiple symptoms in same sentences.

V) PROPOSED SYSTEM

Finding the possible diseases according to questions of users is very difficult. The medical data will be stored in the form of dataset. Every time the system will respond to the user query according to the raw dataset. The dataset forms according to different categories such as diseases name, symptoms, precautions, descriptions etc. A normal user can search in the system or can fire questionnaire to system. The system will respond it according to associated dataset, .this process is called deep learning.

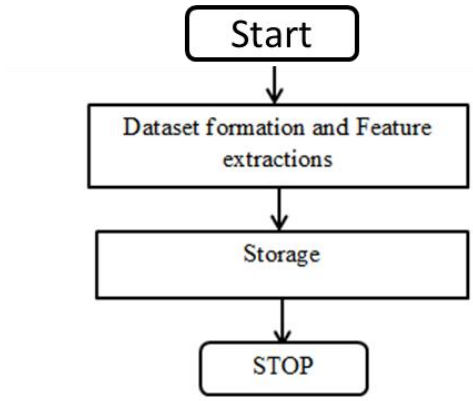

Fig1. Flowchart for Data Collection

1. Dataset formations and features extractions

The information within the database is transformed into the dataset. The dataset is the collection of related sets of information that is composed of separate elements but can be manipulated as a unit.

2. Deep Learning

Deep learning is also called as hierarchical learning or structured learning .It is the branch of machine learning based on a set of algorithms that attempt to model high-level abstractions in data by using multiple processing layers with complex structures. Deep learning presents this idea of hierarchical explanatory factors where higher level, more abstract concepts being learned from the lower level ones. These architectures are often constructed with a layer-by-layer method. Sparsely connected deep learning method contains number of layers depending on the health care application. These techniques mainly contain input and output layers. The nodes in the input layer contain number of raw features and nodes in the output layer contain result which denotes ultimate disease type. Remaining layers are constructed incrementally alternating between sub graph mining and pre-training. Each node in the hidden layer is

corresponding to a signature obtained by sub graph mining from a large graph, where the large no of raw features are assumed as nodes and edges, respectively. This deep learning method employs with the help of mining which help to find interdependent medical attributes from the large dataset.

This system first receives the input in the form of question/symptom from health seeker, it then tries to elaborate the symptoms in detail to discriminate its various aspects by throwing cross questions to the user. This allows in-depth analysis and restructuring of the received information to fit in the preformed structured dataset. It is then presented to sparsely connected deep learning scheme that is able to infer the possible diseases from the questions of health seekers. Therefore, it is generalizable and scalable as compared to previous disease inference using shallow learning approaches, which are usually trained on hospital generated patient records with structured fields. Classical deep learning architectures are densely connected and the node number in each hidden layers are tediously adjusted. But this previous scheme is not able to find the discriminate features of the each disease, and we are trying to remove this limitation. Our current model is able to identify discriminant features for each specific disease.

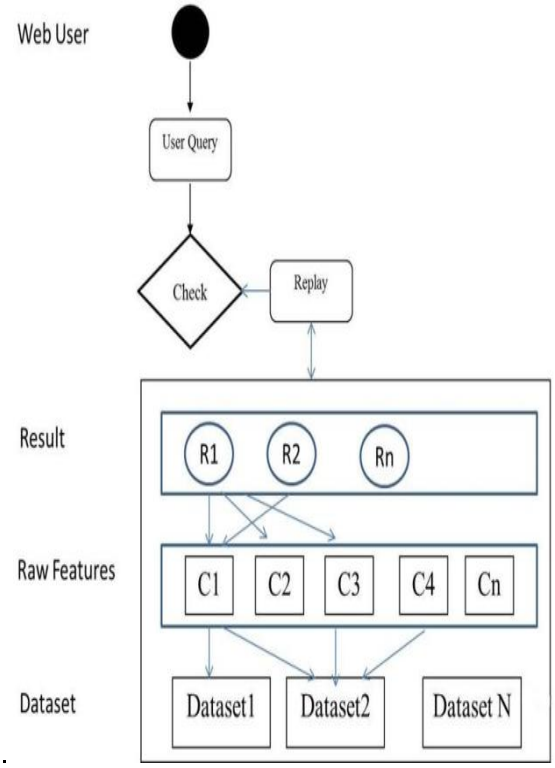

Fig2. System Architecture

VI) CONCLUSION

Numerous techniques are developed in the previous years like WebMD, MedlinePlus etc. The wide scope of this technique is presented in this paper. The accuracy level is achieved without huge time consuming unlike in previous works. The proposed system will help user to narrow down the list of differential diagnoses based on the symptoms, targeted medical history retrieved from the user and reports of some basic, noninvasive investigations. The ability of the system to arrive at or nearby a precise condition will certainly help the user to alleviate the curiosity as well as anxiety generated because of the symptoms. As the user will get basic first-hand knowledge of the medical condition, he/she will be able to make learned choices in the future. As the system will also direct the user to consult the best suited healthcare facility for further

course of action, it will add to the confidence, convenience and satisfaction of the user.

VII) FUTURE SCOPE

The growing vastness, superspecialities, complexities and advancements in the medical field requires that the user should also keep pace with these advancements. This certainly is not possible for everyone, so this system will help the user to make learned choices regarding health issues. A learned person cannot be misguided even by a healthcare professional and this gives added confidence to the user when consulting/dealing with healthcare professional. It also saves the user from inappropriate and unnecessary medical consultation as many symptoms are temporary, seasonal and nonsignificant. At the same time it may also help the user to become aware of some slowly developing disease, with casual symptoms. The timely awareness may lead to the diagnosis of disease like cancer in early curable stage. The proposed system will save the time as well as money, as it will be easy to approach the best medical facility in the locality as per user's affordability. To add to the convenience of user, system will also help arrange appointment with the hospital/consultant selected by the user.

References

[1] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," IEEE Transactionson Pattern Analysis and Machine Intelligence, 2013.

[2] S. Fox and M. Duggan, "Health online 2013," Pew Research Center, Survey, 2013.

[3] "Online health research eclipsing patient-doctor conversations," Makovsky Health and Kelton, Survey, 2013.

[4].Y. Zhang and B. Liu, "Semantic text classification of disease reporting," in Proceedings of the International ACM SIGIR Conference, 2007.

[5] G. Luo and C. Tang, "On iterative intelligent medical search," in Proceedings of the International ACM SIGIR Conference, 2008