A Tag Mining framework for Disease Inference from Health related data.

Ms. Vidhi L. Chawda
M.E. Computer Engineering
SSGM College of Engineering,
Shegaon,(M.S.) 444203, INDIA
vlchawda1993@gmail.com

Vishwanath S. Mahalle
Asst. Professor, Department of
Computer Science & Engineering
SSGM College of Engineering
Shegaon,(M.S.) 444203, INDIA
vsmahalle@gmail.com

Abstract— Normally people use Google to search their queries and that search engine respond them with the answer but that answer is in scattered format. User not gets exact answer for his / her queries. So we are going to implement this paper, we first report a user study on the information needs of health seekers in terms of questions and then select those that ask for possible diseases of their manifested symptoms for further analytic. We next propose a learning scheme to finding the possible diseases given the questions of health seekers. The proposed scheme comprises of two key components. The first globally mines the discriminate medical signatures from raw features. The second deems the raw features and their signatures as input nodes in one layer and hidden nodes in the subsequent layer, respectively. Meanwhile, it learns the inter-relations between these two layers via pre-training with pseudo labeled data. Following that, the hidden nodes serve as raw features for the more abstract signature mining. With incremental and alternative repeating of these two components, our scheme builds a sparsely connected deep architecture with three hidden layers. Overall, it well fits specific tasks with fine-tuning. Extensive experiments on a real-world dataset labeled by online doctors show the significant performance gains of our scheme.

Keywords- Hidden layers, Community-based Health Services, Question Answering, Disease Inference and Deep Learning.

I. INTRODUCTION

In today's increasing development of every country, its expenditure on healthcare and emergence in computer technologies, these all are the major reasons for creation and innovation of automatic health seeking system. During survey it is easy to understand that most of the people make usage of emerging technologies such as computers, journals, magazines and internet technologies. The mining of Knowledge in health record is one of the important aspects for clinical decision making, patient management and population management.

The large number of information they capture over time pose challenges not only for medical practitioners, but also for the information analysis by machines and local users they are intended to get satisfied with their need. Day to day the lifestyle of people get changes fast. So health related issues are plays important role in today's life. Health is demand of busy lifestyle is not an easy thing to get, but is best managed by regularly reviewing and assessing your priorities. Lot of people searches the answers for their health related questions on the internet but they not get satisfied every time. So I motivate to from this concept to choose such topic which will provide the better answers to the users health related questions. Diseases tracing plays important role in daily life. Every one cares about himself or herself health. According to some social study, lot of people spends their time on online searching of health related issues. By browsing they get lot of information about the medical concepts and health related issues.

Many of us are surfing internet to get any disease related information but still they did not get the appropriate information they require so for them our system will give accurate information. Disease Inference system which will give the disease information which he/she is facing on the basis of health related questions. In a less amount of time he/she will get to know what he/she is facing and that to by sitting at the home. The disease inferring architecture is as shown in the below figure 1.

Community-based health service is very time consuming process for health seekers to get their posted questions resolved in a particular period of time. Sometimes this resolving time could vary from hours to days some may be more than these. This leads to reduction in efficiency. Using our system health seeker will get immediate response as compared to the existing system. Our approach is distinctly different in that we are trying to build a general predictive system which can utilize a less constrained feature space. There are various methods for this system like sparse deep learning, SVM (Support Vector Machine), etc.

Because of the growing aged population coupled with lack of medical services and healthcare services in most of the developing countries, the traditional health- care system meets challenging problems caused by its high operating cost and unscalability. Compared to the conventional healthcare system, there is a need of more accurate and easy to access system to improve quality of health care services. The following are the some of the parameters we needed to be improved for the health care services. They are improving the quality of medical service, Improving the utilization of medical helps and care by enabling remote medical services, and supporting the development of the health industry. The existing system mainly focuses on healthcare service in a physiological and psychological aspect with the following two undesirable features etc. The existing literatures [9],[5] and [8],[6] are explaining some discriminant features. The literatures [3],[4],[2],[7] following some earlier methods of health services. There are various methods or we can say that various algorithms are used in each paper represents different methods for disease inference so that user can get the information for their query. There are various limitations inside existing system which may overcome in this proposed system.

The key points of this paper are a) To provide more accurate information to the user for its query. b) To infer the possible diseases, given the questions of health seekers. c) The system can remove the obstacles such as the vocabulary gap, incomplete information, correlated medical concepts. d) System can able to find out the proper solution for users query.

The existing system is not working with insufficient data, if user input the insufficient data then system does not work for that query. The medical concepts stored in the form of the dataset. But the relation between dataset not associated properly. This all limitations may get overcome in this system. So limitations are a) Users not get answers on his query in case of insufficient data. b) Because of lot of dataset, the system performance is very slow. c) Problem occurs in the association of datasets i.e. maintaining relationship between the dataset is difficult task.

II. Disease Inference

Disease Inference is nothing but one prediction of disease given by system to the user for their query. This will help the user to find one disease prediction from which they may suffer from. There are various problems like vocabulary gap, incomplete information, inter-dependent medical attributes and limited ground truth have greatly hindered the performance of classic shallow machine learning methods. To tackle these problems, we propose a novel deep learning scheme to infer the possible diseases given the questions of health seekers. Compared to shallow learning, deep learning has several advantages.

First, it is able to learn representative and scalable features from other disease types. Consider one example of Lung Cancer Inference Learning, while building the classifier data can be liver cancer or other disease samples rather than strictly constrained to lung cancer. This will shows the limited ground truth and necessity of disease-aware feature extraction. Second, inherited from its deep architectures, here hidden layer architecture is get used. So layer by layer more compact patterns are get obtained. This advantage helps the user to get accurate answer for their query. This enables the system to mine the underlying connections among medical attributes.

Third, deep learning can seamlessly integrate signatures as hidden nodes. As we analyzed, signatures infer the incomplete information. With deep learning, each data instance will be ultimately represented by a mixture of very high-level abstract patterns, which are semantic Descriptors and thus are more robust of data inconsistency caused by vocabulary gap.

III. Overview OF Disease Inferring Scheme

The general disease inferring architecture is as shown in below figure 1:

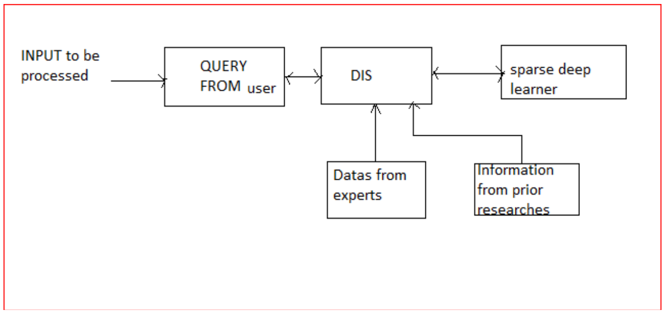

Figure 1: Disease Inferring Architecture.

The above figure shows the architecture for disease inference for heath seekers or users as shown in figure. The main aim of this paper is to give prediction of diseases when user or health seeker fires a query. The query here is in the form of symptoms that the user is going through it. After the Disease inference system will uses Deep learning algorithm to provide accurate result to the health seekers. In the above figure shown we will not give you information from the experts, as we have added some new concept so that every time there is no need to upgrade information in database. Here users or health seekers can ask health oriented questions to the inferring system and system will provide them the accurate answer or trustworthy answers. In above figure DIS is nothing but Disease Inferring System.

This inferring architecture is used as basic in every disease inference scheme. In this architecture, DIS (Disease Inferring System) contains the concept various hidden layers. The complete working of this system depends on these hidden layers. Here how hidden layers get completely work and gives more accuracy in question answering between system and users is done.

Methodology Used	Author	Advantages	Disadvantages
Support Vector Machine	David Barbella	valuable and useful tool for making classifications.	they lack the natural explanatory value.
Signature Mining	F. Wang, N. Lee, 2013	This architecture enables the representation, extraction, and mining of high-order latent event structure.	No clinical assessment for visual interactive knowledge discovery in databases for users need.
Decision Tree	M. Shouman, T. Turner, and R. Stocker, 2011	The diagnosis of a disease has been investigated showing good levels of correctness.	It mostly use for identifying heart disease only.
Representation Learning	Y. Bengio, A. Courville, and P. Vincent, 2013	It has efficient machine learning algorithm.	Its disability to extract and organize the discriminative information from the data.
CQA services	Liqiang Nie and Tat-Seng	Its reliable and handy to use.	The result optimization is not proper.
Sparse Deep Learning	Liqiang Nie, Meng Wang, Luming Zhang, Shuicheng Yan, 2014	user will get accurate information for its query.	User not get answer in case of insufficient data.

Figure 2: Comparison between various methodologies for Disease Inference.

Here, sparse deep learning is machine learning algorithm which is used for complete processing of the system. This learning technique has some drawbacks as it get implemented in previous paper. So, to overcome the drawbacks novel deep learning is used in this paper, which has same working as of sparse deep learning, but this may remove or overcome the disadvantages in existing paper. For this we collected more than 900 popular disease concepts from EveryoneHealthy5, WebMD and Medline Plus.

There are various methodologies used in different papers which are as shown in figure 2. These methodologies have their advantages and disadvantages as they somewhat get overcome in this paper. We all know how disease tracing plays important role in everyone's life. So every user wants answer of their query from the system that they fired to the system. Each and every methodology used in previous papers want some modifications to overcome their drawback. So the proposed system introduces new method to answer users query. Due to this user can get more accurate and satisfied answers.

So, from this table of comparison between various methodologies we can say that every technique has some advantages and disadvantages, which may overcome in new papers that get published. To resolve the problems in existing system we proposed a new system or technique which gives better results as compared to previous existing system. The new system gives one new contribution to existing system of automatic prediction of diseases.

As we know in community-based health services this work is done manually by maintaining records of each and every patient in the hospital. But here this work is done automatically by the system.

IV. PROPOSED SYSTEM

The main aim of this paper is to build a disease inference scheme that is able to automatically infer the possible diseases of the given questions in community based health services.

We are going to implement a scheme to finding the possible diseases according to questions of users. The medical data will store in the form of dataset. Every time the system responded to the user query according to the raw dataset. When user fires any query then the system accept that request and compare with the collected dataset. The dataset are nothing but the raw data. When two users discuss on any medical concept the system automatically convert the discussion into something like array of text. From that text the medical concept will be sort out and converted into the datasets. The dataset forms according to different categories such as diseases name, symptoms, precautions, descriptions etc.

So to provide such facility of giving accurate to the users we will us the algorithm which is named as "Novel Deep Learning" algorithm. This algorithm is same as deep learning algorithm, but after adding some new features into it named as novel deep learning. The main thing that is new in this algorithm is Communication feature between doctors. This will help us not to update the system information manually by admin, about new diseases, their symptoms, precautions, etc. But this will automatically done by communication between doctors.

We proposed and develop the scheme who studies the user information and health related data. In our application, the user request is compared with the different dataset. The datasets are automatically created using the discussion of doctors. The doctors will discuss on various medical concepts. The system will save the medical information in the form of datasets. So one thing here we can say that, according to the doctors conversation the datasets inside the database get automatically updated. In the existing system, the admin has to manually update the information or we can say that he/she has to update the information about the new diseases, their symptoms, causes, precautions, etc. The main Objectives of this paper are:

A) To provide more accurate information to the user for their query:

In proposed system, every user will get more accurate information for their query. When compared with existing system, the proposed system gets more perfect result for users query. This will help the user for their disease prediction and the user can move to further procedure immediately as they get accurate disease inference.

B) To infer the possible diseases, given the questions of health seekers:

Proposed system identifies discriminant feature for specific disease. This means that there is no confusion in disease prediction according to users query. So that user can get accurate answer for their query. Distinguished answers will obtain to the user for the query. This means that for disease prediction proper query must be fired by health seekers. This will help the system to give proper prediction for user queries.

C) The system can remove the obstacles such as the vocabulary gap, incomplete information, correlated medical concepts:

There are various obstacles in the existing system like Vocabulary gap, incomplete information, correlated medical concepts. These obstacles are overcome in our proposed system. This will help the user to find the spellings of the diseases that they don't know. When we talk about the incomplete information, it means that only limited information is there in the system. Sometimes, this irritates the user and indirectly will affect on the reviews of the system.

D) System can able to find out the proper solution for users query:

All these objectives are key points in this proposed system, to get better results to the user so that every user gets satisfied results according to the query that user fired. So our system can be able to find out proper solution on the users query.

This paper aims to build a disease inference scheme that is able to automatically infer the possible diseases of the given questions in community-based health services. We first analyze and categorize the information needs of health seekers. It is worth emphasizing that large-scale data often leads to explosion of feature space in the lights of n-gram representations [8], [9], especially for the community generated inconsistent data. To avoid this problem, we utilize the medical terminologies to represent our data. Our scheme builds a novel deep learning model, comprising two components, as demonstrated in Figure 3.

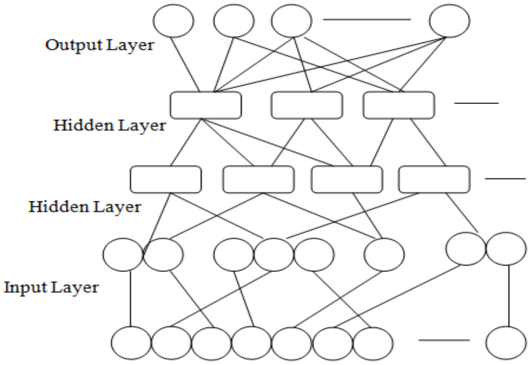

Figure 3. Process of sparsely connected Deep learning

Extensive experiments on real-world dataset labelled by online doctors were conducted to validate our scheme. The main contributions of this work consist of 3 things:

1) This is the first work done on automatic disease inference in the community-based health services. Distinguished from the conventional sporadic efforts that generally focus on only a single or a few diseases based on the hospital generated records with structured fields, our scheme benefits from the volume of unstructured community generated data and it is capable of handling various kinds of diseases effectively.

2) The information needs of health seekers in the community-based health services get categorizes and mines the signatures of their generated data.

3) It proposes a sparsely connected deep learning scheme to infer various kinds of diseases. This scheme is pre-trained with pseudo-labelled data and further strengthened by fine-tuning with online doctor labelled data.

There are two main modules of this system which includes User and Doctor. In the previous existing system, there is only one user module. Due to this second module the system updates automatically.

a. User Module:-

Every user has to registered first, after that he can login to the system using proper username and password. This user can search in the system or fired questionnaire to system. User can do two things inside this system, search for disease information and also can ask the system for disease inference. The system will respond it according to associated dataset.

When particular user wants information about the disease then that will be provided by the system in specified manner. This may include some introduction about that disease, after that symptoms, precautions, images, and provide audio or videos if available in the database. So these functions are done in the User Module. In the existing system, the user module cannot search for the disease information.

b. Doctor Module:-

In Doctor Module, doctors can make discussion on different medical concept and medical raw data. The system will follow them and gather the medical related information and place them in the form of dataset.

The dataset includes the information about disease, symptoms, precaution and description. And when user fires the query which is related to the dataset that get updated by the doctor's communication, then the user gets answer of that query. So manual updating of information about the diseases is not done in this system. Automatic data updating is done in this proposed system. This will help the admin not to update data manually every time. The general dataflow diagram of our system is as shown in below figure 4.

These two modules show the complete working flow of the proposed system in which user fires a query in terms of raw data, this query get checked inside the database with every single database layer by layer. And after question answering between system and user finally every user get their answer for that query.

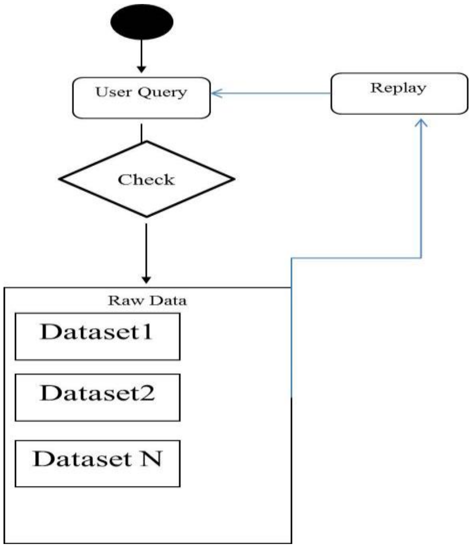

Figure 4. General Dataflow Diagram for Disease Inference

By using deep learning framework, disease analysis is getting done. One scheme is constructed i.e. disease inference estimation scheme to analyze the question and answers in online health services. Medical domain based Ontology is adapted to identify the disease inferences. The two operations, feature selection and categorization operations are integrated with the system. The system is distinguished in three major modules. They are Question and Answer Sessions, Tag Analysis and Deep Learning Process, The Question and Answers (QA) session module is designed to perform the data preprocess. Tags are identified and categorized under the tag analysis. Features and signatures are identified under deep learning process.

1) Question and Answer Sessions:-

The Question and Answer (QA) data sets are collected from online health services. These sessions are done between user and system. The QA data values are transferred into the database. Questions, answers and tags are extracted from the datasets inside the database. The datasets are filled with category information.

2) Tag Analysis:-

Here tags are nothing but the query fired by the user to the system. The tags and associated disease information are identified in the tag analysis. Overlapped tag details are also updated with category information. Feature identification is performed in the tag analysis. Features and associated tag labels are updated into the database.

3) Deep Learning Process:-

Pseudo labeled data and doctor labeled data are analyzed in the learning process. Signatures get identified from the raw data under the learning process. Input layers and hidden layers are updated with features and signatures. The layers are used in the inference identification process. Raw data is nothing but data which we get by different sites like Medline Plus, WebMD, etc. In this way the complete process is get done by the system with the help of these three modules.

The Algorithm i.e. Novel deep learning algorithm is one deep learning algorithm. The difference between Shallow learning and deep learning is "shallow" machine learning, which is often based on user having some prior knowledge which specific features of the input may help in disambiguating the correct answer. The emphasis in shallow learning is not always on feature engineering and selection while in deep learning the emphasis is on defining the most useful computational graph topology and optimizing parameters correctly. So novel deep learning algorithm that we used here is:

V. Conclusion and Future Work

This paper established a system that first performed user study to analyze the health seeker needs. This provides the insights of community-based health services. It then presented a deep learning scheme that is able to infer the possible diseases given the questions of health seekers. This scheme is constructed via alternative signature mining and pre-training in an incremental way. It permits unsupervised feature learning from other wide range of disease types. 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.

The biggest stumbling block of automatic health system in disease inference. So if the communication between doctors that is added in this system is not done then there is same need of manual data updating as in the existing system. So in our system also, the data updating is also depend on doctor communication. In future, we will pay more attention on this.

ACKNOWLEDGEMENT

This project is done as a part of Tag Mining Framework for Disease Inference using deep learning under the guidance of Prof. V. S. Mahalle, Dept. of Computer Science and Engineering, SSGMCE Shegaon.

REFERENCES

[1]Liqiang Nie, Meng Wang, Luming Zhang, Shuicheng Yan, Member, IEEE, Bo Zhang, Senior Member, IEEE, Tat-Seng Chua, Senior Member, IEEE ”Disease Inference from Health-Related Questions via Sparse Deep Learning” IEEE Transactions on Knowledge and Data Engineering ,May 2014.

[2] F. Wang, N. Lee, J. Hu, J. Sun, S. Ebadollah , and A. Laine, “A framework for mining signatures from event sequences and its applications in healthcare data,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013.

[3]F. Wang, N. Lee, J. Hu, J. Sun, and S. Ebadollahi, “Towards heterogeneous temporal clinical event pattern discovery: A convolutional approach,” in The ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2012.

[4]David Barbella1, Sami Benzaid2, Janara Christensen3, Bret Jackson4, X. Victor Qin “Understanding Support Vector Machine Classifications via a Recommender System-Like Approach” in Proceedings of the IJSR Conference, 2013.

[5]Lejun Gong∗, Ronggen Yang, Qin Yan, and Xiao Sun, “Prioritization of Disease Susceptibility Genes Using LSM/SVD” in Proceedings of the IJSR Conference, 2011

[6]M.Shouman, T. Turner, and R. Stocker, “Using decision tree for diagnosing heart disease patients,” in Proceedings of the Australasian Data Mining Conference, 2011.

[7]Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013.

[8]Refinement of the Facility-Level Medical Technology Score to Reflect Key Disease Response Capacity and Personnel Availability, Olumurejiwa A. Fatunde (Student Member, IEEE)1, And Timothy W. KOTIN (Student Member, IEEE) 2012.

[9] “Online health research eclipsing patient-doctor conversations,” Makovsky Health and Kelton, Survey, 2013.

[10] L. Nie, M. Wang, Z. Zha, G. Li, and T.-S. Chua, “Multi-media answering: Enriching text qa with media information,” in Proceedings of the International ACM SIGIR Conference, 2011.

[11] L. Nie, Y.-L. Zhao, X. Wang, J. Shen, and T.-S. Chua, “Learning to recommend descriptive tags for questions in social forums,”ACM Transactions on Information System, 2014.

[12] P. Sondhi, J. Sun, H. Tong, and C. Zhai, “Sympgraph: A framework for mining clinical notes through symptom relation graphs,” in The ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2012.

[13] L. Nie, M. Wang, Y. Gao, Z.-J. Zha and T.-S. Chua, “Beyond text qa: Multimedia answer generation by harvesting web information,” Multimedia, IEEE Transactions on, 2013.

[14] D. Zhu and B. Carterette, “An adaptive evidence weighting method for medical record search,” in Proceedings of the International ACM SIGIR Conference, 2013.

[15] R. Fakoor, F. Ladhak, A. Nazi, and M. Huber, “Using deep learning to enhance cancer diagnosis and classification,” in Proceedings of the International Conference on Machine Learning, 2013.

[16] 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, 2014.

[17] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, “Why does unsupervised pre-training help deep learning?” Journal of Machine Learning Research, 2010.

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

[19] T. C. Zhou, M. R. Lyu, and I. King, “A classification based approach to question routing in community question answering,” in The International World Wide Web Conference, 2012.

[20] R. W. White and E. Horvitz, “Studies of the onset and persistence of medical concerns in search logs,” in Proceedings of the International ACM SIGIR Conference, 2012.

[21] M.-A. Cartright, R. W. White, and E. Horvitz, “Intentions and attention in exploratory health search,” in Proceedings of the International ACM SIGIR Conference, 2011.

[22] B. Koopman, P. Bruza, L. Sitbon, and M. Lawley, “Evaluating medical information retrieval,” in Proceedings of the International ACM SIGIR Conference, 2011.

[23] C. B. Akgu¨ l, D. U¨ nay, and A. Ekin, “Automated diagnosis of alzheimer’s disease using image similarity and

user feedback," in Proceedings of the ACM International Conference on Image and Video Retrieval, 2009.

[24] S. Doan and H. Xu, "Recognizing medication related entities in hospital discharge summaries using support vector machine," in Proceedings of the International Conference on Computational Linguistics, 2010.

[25] J. Zhou, L. Yuan, J. Liu, and J. Ye, "A multi-task learning formulation for predicting disease progression," in The ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2011.

[26] K. I. Penny and I. Atkinson, "Approaches for dealing with missing data in health care studies," Journal of Clinical Nursing, 2012.

[27] H. Liu, L. Latecki, and S. Yan, "Fast detection of dense sub graphs with iterative shrinking and expansion," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013.

[28] S. Ghumbre, C. Patil, and A. Ghatol, "Heart disease diagnosis using support vector machine," in Proceedings of the International Conference on Computer Science and Information Technology, 2011.