ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

AUTOMATIC DISEASE INFERENCE MODEL WITH MEDICAL ONTOLOGY SUPPORT

J.U.Paruvatha Kumari[1] , Dr.R.Velumani[2]
[1,2]Department of Computer Science & Engineering
[1,2]K.S.R College of Engineering, Tamilnadu, India

ABSTRACT

Internet is served as a diagnosis tool to facilitate patient-doctor communication. Online health resources are categorized into two categories. They are reputable portals and community based health services.. A reputable portal provides up to date health information by releasing the accurate and well structured health knowledge. WebMD and MedlinePlus are popular reputable portals. Community based health services offer interactive platforms to provide Question and Answer (QA) based medical support. HealthTap and HaoDF are the popular community based health services.

Community based health services supports automatic disease inference identification for online health seekers. Question and Answer(QA) sessions are suffered with the vocabulary gap and incomplete information. Correlated medical concepts and limited high quality training samples makes an impact on inferring results. Diseases and symptoms are collected and used in the QA based health analysis tasks. Deep learning scheme is applied to infer the possible diseases using QA data values. Global leaning component is used to mine the discriminant medical signatures from raw features. In local learning raw features and their signatures are updated into the input layer and hidden layer. Sparsely connected deep learning scheme is applied to infer various kinds of diseases.

The sparse deep learning scheme is enhanced to fetch discriminant features from health data values. Medical terminology based Ontology is used for inference estimation process. Feature analysis is carried out with the conceptual relationship based weight values. Question and Answer (QA) data values are evaluated with symptom priority levels.

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

1. INTRODUCTION

Internet users support each other online in two major ways: through online communities and through personal emails. In previous studies, the Pew Internet Project has found that 84% of Internet users have contacted online interest groups of varying sorts, from hobbies to politics to religion.14 Participation in health related online groups and communities has been steadily rising. In May-June 2001, we found that 36% of Internet users had visited a Web site that provides information or support for people interested in a specific medical condition or personal situation. In September 2002, that number grew to 47% of Internet users, and by December 2002, to 54% of Internet users, or about 63 million Americans.

In addition, about 32 million Americans seek support in a more private form; 30% of email users have sent or received health-related email. About a quarter of email users exchange email with family members about health or medical issues; another quarter do the same with friends. Only 7% exchange emails with doctors or health professionals. Women, better-educated, and more experienced Internet users are more likely to exchange health-related email. Of all those who email about health issues, about 90% find the email useful.

This usefulness and popularity of online support translates into enthusiasm and even passion from e-patients and caregivers for electronic communications. In comments, they describe the value from email and support groups in both emotional and practical terms. A number of themes emerge. On the emotional side, empathy is highly valued; giving support is as important as getting it. On the practical side, support leads to tangible results.

2. RELATED WORKS

Given a corpus of reviews, words highly correlated with the class label can be identified by many approaches such as class conditional probability of words, information gain, association rules, point wise mutual information (PMI), etc. These approaches, unfortunately, suffer from a severe problem: it is difficult to understand the

underlying aspects or concepts from just a set of words correlated with a class label. There is no intuitive algorithm to group the words so that each group conveys one or a few easily understandable concepts.

Aspect-based opinion mining is becoming popular in recent years. Frequency based approach extracts high frequency noun phrases which meet the specified criteria or constraints from the reviews as aspects. On the other hand, relation based approach identifies aspects based on the aspect-sentiment relation in the reviews. These two kinds of approaches, may not be applicable to drug reviews as aspects are often not indicated explicitly by authors and descriptions of side effects and people's experiences is diverse. Moreover, grouping of the extracted noun phrases is another challenge as they cannot be grouped just based on semantic meanings. In contrast, topic modeling identifies aspects based on the co-occurrence of words in reviews. It has an advantage that aspect identification and grouping are performed simultaneously.

Topic modeling [10] is a popular probabilistic approach in understanding a corpus. With this approach, a set of topics, which are represented by multinomial distributions over vocabulary words, are inferred. When the words of a topic are sorted according to the probabilities, high probability words of a topic are usually semantically correlated and the concept or aspect of the topic can be captured manually. For example, Topic Sentiment Mixture (TSM), Joint Sentiment/Topic (JST) model and Aspect and Sentiment Unification Model (ASUM) [1] were proposed to extract both the aspects and predict their associated sentiments. These aspect based opinion mining methods may not be appropriate to address the problem defined in the extracted aspects may not be related to the specified class labels and the performance depends on the manual selection of seed words.

Topic modeling with supervised label information has become an interest of research. Blei and McAuliffe proposed the supervised LDA (sLDA) that can take care of different forms of supervised information during topic inference. Mimno and McCallum introduced Dirichlet-multinomial regression to handle different kinds of meta information. Ramage et al. proposed DiscLDA to process discriminative information and find topics specific to individual classes as well as topics shared across different classes.

Labeled LDA is another generalization of LDA. It allows multi-label supervision and associates each label with one topic in direct correspondence.

Apart from probabilistic algorithms, deterministic methods for topic modeling such as non-negative matrix factorization (NMF) were also proposed. By decomposing the data matrix into two low rank matrices, topics can be identified. Semi-supervised NMF (SSNMF) [12] is an extension proposed recently to incorporate the supervised information into NMF. The topics identified are more closely related to the supervised information.

3. ONLINE HEALTH RESOURCES

The community-based health services have several intrinsic limitations. First of all, it is very time consuming for health seekers to get their posted questions resolved. The time could vary from hours to days [3]. Second, doctors have to cope with an ever-expanding workload, which leads to decreased enthusiasm and efficiency. Taking HealthTap as an example, as of January 2014, it had gathered 50 thousand doctors and accumulated more than 1:1 billion answers, i.e., on average each doctor has online replied approximately 23 thousand times since its foundation in 2010. Third, qualitative replies are conditioned on doctors' expertise, experiences and time, which may result in diagnosis conflicts among multiple doctors and low disease coverage of individual doctor [4]. It is thus highly desirable to develop automatic and comprehensive wellness systems that can instantly answer all-round questions of health seekers and alleviate the doctor's workload.

The biggest stumbling block of automatic health system is disease inference. According to health seekers frequently ask for: (1) supplemental cues of their diagnosed diseases; (2) preventive information of their concerned diseases; and (3) possible diseases of their manifested signals. The former two genres usually involve the exact disease names and expected sub-topics or sub-problems of the given diseases, such as the side effects of specific medications, and treatments. They can be automatically and precisely answered by either directly matching the questions in the archived repositories or

syntactic information extraction from the structured health portals. The existing automatic question answering techniques are applicable here [5], [6]. The third genre conveys parts of the health seekers' demographic information, physical and mental symptoms, as well as medical histories, in which they do not know what conditions they might have and expect the doctors to offer them some forts of online diagnosis. If the diseases are correctly inferred, these questions are naturally transferred to the first genre. Hence a robust disease inference approach is the key to break the barrier of automatic wellness systems.

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. As a byproduct, we differentiate questions of this kind that require disease inference from other kinds. It is worth emphasizing that large-scale data often leads to explosion of feature space in the lights of n-gram representations [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. The first globally mines the latent medical signatures. They are compact patterns of inter-dependent medical terminologies or raw features, which can infer the incomplete information. The raw features and signatures respectively serve as input nodes in one layer and hidden nodes in the subsequent layer. The second learns the inter-relations between these two layers via pre-training. Following that, the hidden nodes are viewed as raw features for more abstract signature mining. With incremental and alternative repeating of these two components, our scheme builds a sparsely connected deep learning architecture with three hidden layers. This model is generalizable and scalable. Fine-tuning with a small set of labeled disease samples fits our model to specific disease inference. Different from conventional deep learning algorithms, the number of hidden nodes in each layer of our model is automatically determined and the connection between two adjacent layers is sparse, which make it faster. Extensive experiments on real world dataset labeled by online doctors were conducted to validate our scheme.

Community based health services supports automatic disease inference identification for online health seekers. Question and Answer (QA) sessions are suffered with the vocabulary gap and incomplete information. Correlated medical concepts and limited high quality training samples makes an impact on inferring results. Diseases and symptoms are collected and used in the QA based health analysis tasks. Deep learning scheme is applied to infer the possible diseases using QA data values. Global learning component is used to mine the discriminant medical signatures from raw features. In local learning raw features and their signatures are updated into the input layer and hidden layer. Sparsely connected deep learning scheme is applied to infer various kinds of diseases. The following problems are identified from the current online health resource system. Discriminate feature identification is not supported, Medical terminology relationship analysis is not provided, Feature priority factors are not considered and Limited inference accuracy levels.

4. QUESTION AND ANSWER BASED HEALTH SERVICES

To make more informed decisions towards better health, health seekers are getting increasingly savvy with their information needs. Specifically, each health seeker has very specific needs and knows what they expect when they look into the Internet. This leads to diverse, sophisticated and complex motivations and needs of online health seeking. To gain insights into health seeker needs, we randomly collected 5000 QA pairs from HealthTap, which cover a wide range of topics, including cancer, endocrine and pregnancy. We carefully went over all these QA pairs and observed that the health seeker needs can be abstracted into three main categories. Specific motivations and question examples are also provided to enhance the understanding of this categorization. It can be seen that the three categories do not mutually overlap and cover all the possible cases. This is because the health seeker with respect to a concerned health problem can only be in one state out of the three at one time: healthy status, suffering from diagnosed disease or undiagnosed disease.

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

A user study to investigate the health seeker needs. Three volunteers were invited to manually classify each of the 5000 QA pairs into one of the three predefined categories. It is worth noting that each volunteer was pre trained with the definitions of category types as well as corresponding examples. We performed a voting method to establish the final classification of each QA pair. For cases where each class equally receiving one vote, a discussion was carried out among the volunteers to obtain the final decision. According to our statistics, the distributions of QA pairs over the three categories are 79, 6 and 15 percent, respectively. Even though the third category is not the majority, it greatly increases the bottlenecks of the automatic health system as we have analyzed before.

Automatically categorizing this community generated health data is somewhat difficult because of the negated language and vocabulary gap. Regarding the negated language, negated identifiers are frequently used by medical practitioners to indicate that patients do not have given conditions. Some traditional approaches do not distinguish between the positive and negative contexts of medical concepts in medical records, which may prevent the learning/retrieval performance from being effective. Take the following two short medical records as an example. Intuitively, their contexts are totally different, while a learning or search system may inaccurately consider such medical records to be equivalent.

5. MEDICAL ONTOLOGY BASED DISEASE INFERENCE SCHEME

The sparse deep learning scheme is enhanced to fetch discriminate features from health data values. Medical terminology based Ontology is used for inference estimation process. Feature analysis is carried out with the conceptual relationship based weight values. Question and Answer (QA) data values are evaluated with symptom priority levels.

The disease inference estimation scheme is constructed to analyze the question and answers in online health services. Medical domain based Ontology is adapted to identify the disease inferences. Feature selection and categorization operations are

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

integrated with the system. The system is partitioned into six major modules. They are Question and Answer Sessions, Tag Analysis, Investigation with Ontology, Deep Learning Process, Discriminatory Feature Selection and Inference Identification Process.

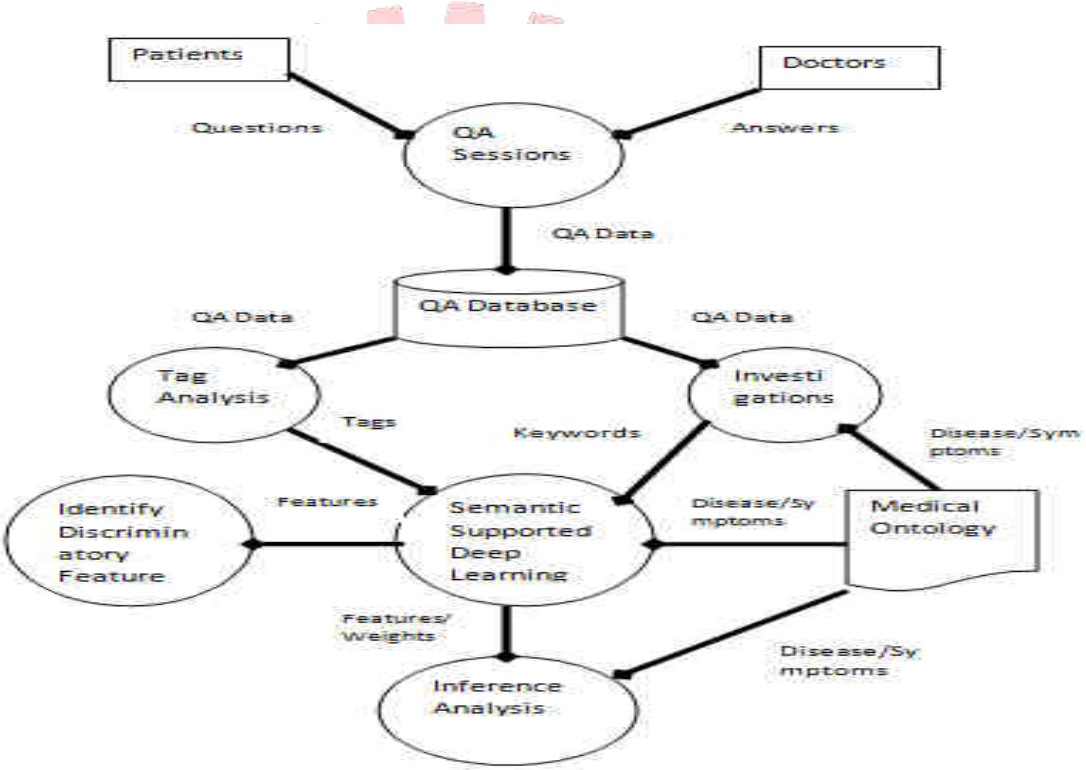

Fig. No: 5.1. Medical Ontology based Disease Inference Scheme

The Question and Answers (QA) session module is designed to perform the data preprocess. Tags are identified and categorized under the tag analysis. Question and Answer data values are analyzed with Ontology. Features and signatures are identified under deep learning process. Discriminatory feature identification is used to discover features for decision making process. Disease discovery is carried out under the inference identification process.

The Question and Answer (QA) data sets are collected from online health services. The QA data values are transferred into the database. Questions, answers and tags are extracted from the data sets. The data sets are labeled with category information. The tags and associated disease information are identified in the tag analysis. Overlapped

tag details are also updated with category information. Keyword feature identification is performed in the tag analysis. Features and associated tags labels are updated into the database.

Medical Ontology is constructed with disease categories and term elements. Term feature relationship is also represented in the Ontology. QA data values are analyzed with Ontology elements. Keyword features are assigned with conceptual relationship weight values. Pseudo labeled data and doctor labeled data are analyzed in the learning process. Signatures are identified from the raw features under the global learning process. Input layers and hidden layers are updated with features and signatures. The layers are used in the inference identification process.

Unstructured community generated data values are analyzed to fetch the discriminatory features. Features are ranked with the conceptual relationship based weight values. Symptom priority levels are also used to identify the discriminatory features. Overlapped features are also verified in the feature selection process. Automated disease inference identification is carried out for the community based health services. Sparse deep learning based inference identification process is applied without concept relations. Concept hierarchy is used in the Ontology Supported Space Deep Learning method. Healthy status, diagnosed disease and undiagnosed disease labels are produced in the inference identification process

.

6. CONCLUSION

Online health services are deployed to provide remote medical assistance. Automatic disease inference estimation is carried out using Question and Answer (QA) based diagnosis details. Sparse deep learning scheme is improved with Ontology support. Discriminate feature identification mechanism is used to upgrade the inference estimation process. Efficient discriminatory feature identification model is adapted in the community based health services. The system provides Ontology support for Question and Answer (QA) based communication process. The Question and Answer (QA) based

system performs decision making with feature priority values. The system improves the accuracy in inference estimation process.

REFERENCES

[1] Y. Jo and A. Oh, "Aspect and sentiment unification model for online review analysis," in Proc. 4th ACM Int. Conf. WSDM, New York, NY, USA, 2011, pp. 815–824.

[2] L. Nie, M. Akbari, T. Li and T.-S. Chua, "A Joint Local-Global Approach For Medical Terminology Assignment," in Proc. Int. ACM SIGIR Workshop, 2014, pp. 24–27.

[3] T. C. Zhou, M. R. Lyu and I. King, "A Classification-Based Approach To Question Routing In Community Question Answering," in Proc. 21st Int. World Wide Web Conf., 2012, pp. 783–790.

[4] D. A. Davis, N. V. Chawla, N. Blumm, N. Christakis and A.-L. Barabasi, "Predicting Individual Disease Risk Based On Medical History," in Proc. 13th Int. Conf. Inf. Knowl. Manage., 2008, pp. 769–778.

[5] L. Nie, M. Wang, Z. Zha, G. Li and T.-S. Chua, "Multimedia Answering: Enriching Text Qa With Media Information," in Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2011, 695–704.

[6] L. Nie, M. Wang, Y. Gao, Z.-J. Zha and T.-S. Chua, "Beyond Text QA: Multimedia Answer Generation By Harvesting Web Information," IEEE Trans. Multimedia, vol. 15, no. 2, pp. 426–441, Feb. 2013.

[7] A. R. Aronson and F.-M. Lang, "An Overview Of Metamap: Historical Perspective And Recent Advances," J. Amer. Med. Informat. Assoc., vol. 17, no. 3, pp. 229–236, 2010.

[8] L. Nie, Y.-L. Zhao, M. Akbari, J. Shen and T.-S. Chua, "Bridging the Vocabulary Gap Between Health Seekers And Healthcare Knowledge," IEEE Trans. Knowl. Data Eng., vol. 27, no. 2, pp. 396–409, Jun. 2014.

[9] M. Galle, "The Bag-of-Repeats Representation of Documents," in Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2013, pp. 1053–1056.

[10] M. Rosen-Zvi, C. Chemudugunta, T. Griffiths, P. Smyth, and M. Steyvers, "Learning author-topic models from text corpora," ACM TOIS, vol. 28, no. 1, pp. 1–38, Jan. 2010.

[11] Victor C. Cheng, C.H.C. Leung, Jiming Liu, Fellow, IEEE, and Alfredo Milani, "Probabilistic Aspect Mining Model for Drug Reviews", IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 8, August 2014

[12] H. Lee, J. Yoo, and S. Choi, "Semi-supervised nonnegative matrix factorization," IEEE Signal Process. Lett., vol. 17, no. 1, pp. 4–7, Jan. 2010.