RESEARCH ARTICLE OPEN ACCESS

Community-Based Health Service for Lexis Gap in Online Health Seekers

U. Vijaya [1], Y. Jahnavi [2], G. Subba Rao [3]
PG Scholar [1], Professor [2], Professor [3]
GIST, Nellore
AP – India

ABSTRACT

Automatic disease inference plays a major role in finding health issues of human beings based on their symptoms. However, illnesses are non-trivial, specifically for community-based health services. It is due to the vocabulary gap, incomplete information, correlated medical concepts and limited top quality training samples. For automatic disease inference, the developed model considers the following steps. In the first step, mining the medical signatures from raw features has been done. In the second step, raw features and medical signatures are used for representing a three layered neural network. By fine-tuning the components, the proposed model outperforms the existing automatic disease inference systems.

Keywords:- Community-based health services, question answering, disease inference, deep learning

I. INTRODUCTION

The present prevailing online health sources could be roughly categorized into two groups. The first is the trustworthy sites operated by official industries, famous organizations or any other professional health service providers. They are disseminating up-to-date health information by delivering probably the most accurate, well-structured and formally presented health understanding on various subjects. The graying of society, increasing costs of healthcare and burgeoning computer technology together are driving more customers to spend extended period online to understand more about health information [1]. Another category may be the community-based health services. They provide interactive platforms, where health seekers can anonymously ask health-oriented questions, while the doctors supply the knowledgeable and reliable solutions. To begin with, it is very time intensive for health seekers to have their published questions resolved. Second, doctors are getting to deal with a constantly growing workload, which results in decreased enthusiasm and efficiency. Third, qualitative replies are conditioned on doctors expertise, encounters and time, which may lead to diagnosis conflicts among multiple doctors and occasional disease coverage of person physician. It is thus highly desirable to build up automatic and comprehensive wellness systems that may instantly answer all-round questions of health seekers and alleviate the workload of doctors. The greatest obstacle of automatic health product is disease inference. The present automatic question responding to techniques are relevant here. The third generation conveys areas of the seekers' demographic information, mental and physical signs and symptoms. In addition to medical histories, that they do not know what conditions they may have and expect the doctors to provide them some forts of internet diagnosis. Hence a strong disease inference approach is paramount to interrupt the barrier of automatic wellness systems. Disease inference differs from subjects or tags assignment to short questions, where subjects or tags are direct summarizations of given data instances plus they may clearly come in the questions. While disease inference is really a reasoning consequence in line with the given question, this is nontrivial because of following reasons. First, vocabulary gap between diverse health seekers helps make the data more sporadic, as in comparison with other formats of health data. Second, health seekers describe their problems in a nutshell questions that contain 14:5 terms per question typically. Third, medical characteristics such as age, gender and signs and symptoms, are highly correlated and do not abnormally appear as compact designs to signal the problems [2]. These four elements limit the condition inference performance

that may be acquired by general shallow learning techniques.

II. LITERATURE SURVEY

Research on healthcare is the most part of science for humans. The existing literatures are diverse and roughly follow four lines of research. Information extraction[10][11][12][13][14]diseaseinference[15][16][17][1] preventive medicine [4][19][20][21] as well as medical search [22][23][24][25] [16][2] [8] [9].

Information extraction from medical data is the important for other higher-order analytics, such as representation, classification, and clustering. The work in [12] utilized SVM to recognize the medication related entities in hospital discharge summaries, and classified these elements into predefined categories, such as treatments and conditions. By the extraction, Sondhi et al[17] constructed entity graphs by exploring their co-occurrence relations and studied how to leverage such graphs to convert raw entity mentions into more useful knowledge, which is helpful for feature expansion. These efforts only consider the explicitly present medical entities, while they overlook the temporal aspect of data as well as the latent discriminative patterns across patient records[14].To deal with these two problems, Wang et al[10][11] proposed a nonnegative matrix factorization based framework to mine common and individual shift-invariant temporal patterns from heterogeneous events over different patient groups, which is able to handle sparseness and scalability problems. As a complementary work, a simple yet effective tool for visualizing the temporal associations among multiple records was designed in[13].

Researchers have been increasingly attracted to use machine learning techniques to assist health professionals in the diagnosis of diseases. Shouman et al[17] and Ghumbre et al[21] have respectively explored decision tree and SVM in the inference of heart disease, which is the leading cause of death in the world over the past 10 years according to the reports from World Health Organization. A learning framework was presented in[16] that focused on Alzheimer disease inference from magnetic resonance images by integrating visual similarities and user feedback. Instead of building single disease related model, Zhang and Liu[18] trained an infectious disease model with the sentence-level semantic features, and obtained promising performance. Fakoor et al[15] in 2013 addressed the scalability and generality problems of these inference models, and utilized unsupervised feature learning method to enhance cancer types classification of cancer types.

Most of the current healthcare is reactive, triggered by the emerging symptoms of diseases. The irreversible consequences of reaction such as death have motivated the drive towards preventive medicine, where the primary concern is recognizing disease risk and taking action at the earliest signs. Khosla et al[20] presented an integrated machine learning scheme to predict stroke for early intervention and treatment, which achieved improved performance on the cardiovascular health study dataset. More practically, a novel system was designed in[4] which combined collaborative filtering methods with clustering to predict each patient 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. Zhou et al. recently modeled the progression of Alzheimer disease with multi-task learning [19] and fused sparse group lasso[21] respectively.

Medical retrieval is the dominant way for knowledge exchanging and sharing. Huang and Hu[17] proposed a re- ranking model for promoting diversity in medical search. Query-adaptive weighting methods that can dynamically aggregate and score various search results have been presented in [22][23]. These three approaches were all evaluated on the dataset[24].A more systematic evaluating framework for medical record search was developed in [25].The aforementioned approaches were designed for and validated on hospital or lab generated patient records, which are well-organized with structured fields.These approaches are not applicable to online health data due to two main reasons. From the perspective of data property, they have different data structure, quality and number of training samples. From the point of techniques, most of the previous efforts are unable to take advantages of other data types beyond the targeted ones, and hence are not scalable or generalizable. However, the

research on online health data is relatively rare. Luo and Tang[25] in 2008 built a medical web search engine called iMed, which employed medical knowledge and an interactive questionnaire to help searchers form queries. After that, needs of health seeker were explored from multiple persepctives, including intentions and attention[25] onset and persistence of medical concerns[24] as well as the comprehensive wellness search[23]. The existing work was monotonously conducted on retrieval.

III. METHODOLOGY

This paper aims to construct an illness inference plan that has the capacity to instantly infer the potential illnesses from the given questions in community-based health services. First evaluation and classification of the data needs of health seekers. Disease inference using their company kinds are differentiated It is worth emphasizing that giant-scale data frequently results in explosion of feature space within the lights of n-gram representations, specifically for the city produced sporadic data. To avert this problem, we make use of the medical terminologies to represent our data. A manuscript deep learning model composed of two components, as shown in table1 is developed. A sparsely connected deep learning architecture with three hidden layers. This model is generalizable and scalable. Fine-tuning having a small group of labeled disease samples fits our model to a particular disease inference. Not the same as conventional deep learning calculations, the amount of hidden nodes in every layer in our model is instantly determined and also the connections between two adjacent layers are sparse, which render it faster. Extensive experiments on real-world dataset labeled by online doctors were carried out to validate the proposed model.

Table 1: Disease identification based on their symptoms

Data	Medical Terminology	Result
What causes urinary frequency in young woman?	Urinary, urgency	Blood in Urine.
What could cause breast pain in a lady of 37 years of age?	Tight chest, wheezing, dynpnea.	Breast cancer.
I get feel breathing problem for more than two weeks.	Shortness of breath, breathless.	Asthma.

To create more informed choices towards better health, health seekers are becoming more and savvier using their information needs. Particularly, each health seeker has very specific needs and knows the things they expect once they consider the web. This can lead to diverse, sophisticated and sophisticated motivations and requires of internet health seeking[4].To achieve information into health seeker needs, we at random collected a few pairs from Health Tap at random, that go over an array of subjects, including cancer, endocrine and pregnancy. These QA pairs and the health seeker needs could be abstracted into three primary groups. It is worth mentioning that exist some questions, in which the health seekers inquire about one undiagnosed disease, but who currently have been identified with another disease. Evidently from it, such questions do not fit in with any of the three groups. However, within the proposed work, our categorization targets in the health seeker needs instead of health seekers themselves, along with other information communicated within the questions is considered as contexts. A person study to research the seeker needs and voting approach to establish the ultimate classification of every QA pair were carried out. For cases when each class equally receiving one election, attorney at law was transported out one of the volunteers to get the ultimate decision. Some traditional approaches do not separate the good and bad contexts of medical concepts in medical records, which might avoid the learning or retrieval performance from being effective. Within the health towns, customers with diverse backgrounds do not always share exactly the same vocabulary. Sometimes, exactly the same medical subjects might be in modern language expressed with assorted medical concepts. Within this paper, we make use of these normalized medical characteristics to represent the city produced health data. The QA pairs using these terminologies and focus the seeker needs via QA pair classification were represented. The experiments to validate this research. The primary challenging condition in health domain may be the interdependent medical characteristics which are named as signature within this paper. As in comparison to individual raw feature, signatures are

crucial identifications for illnesses. It is also an indicator of several conditions that do not directly involve such as migraine, stroke and negative effects of medicines [5]. Therefore, the medical signatures tend to be more descriptive than raw features and can considerably lessen the dimension of feature space. However, it is not easy to extract such signatures from individual data instances. And also the structures are often unconditionally distributed on the large-scale dataset. Within the work, the latent signatures are seen as overlapping dense sub graphs. As aforementioned, vocabulary gap, incomplete information, inter-dependent medical characteristics and limited ground truth have greatly hindered the performance of classic shallow machine learning approaches. To tackle these complaints, we a manuscript deep learning plan to infer the potential illnesses was advised because of the questions of health seekers. In comparison to shallow learning, deep learning has lots of advantages. First, is the ability to learn representative and scalable features using their company disease types. Second, inherited from the deep architectures, it frequently discovers the greater abstract compact designs layer by layer. This allows the machine to mine the actual connections among medical characteristics. Third, deep learning can effortlessly integrate signatures as hidden nodes. Most significantly, with deep learning, each data instance is going to be ultimately symbolized by a combination of high-level abstract designs that are semantic descriptors and therefore tend to be more robust of information inconsistency brought on by vocabulary gap.

Our sparsely connected deep learning has L layers with d (1<l<L) nodes in each layer. The first layer contains the input n-dimension raw features and the L-th layers denote the output disease types. Deep learning architectures has nodes in the higher layer are signatures and connect to the nodes in its adjacent lower layer. These relations are indicated by affinity matrix W_{ij}^{1}

$$\begin{cases} Z^{1+1} = W^{1}O^{1} + b^{1}, \\ O^{1+1} = f(Z^{1+1}), \\ h_{W,b}(X) = O^{L} = f(Z^{L}), \end{cases}$$

Where,

$f(Z)$ and $h_{W,b}(X)$ are the activation and objective function.

Z^{1} and O^{1} denote the sum weighted inputs

b^{1} represents the bias term with node i in layer l+1.

W^{1} denotes the weighted output layer.

IV. EXPERIMENTAL ANALYSIS

This section evaluates the proposed technique. We discuss the experimental data, evaluation settings and the results.

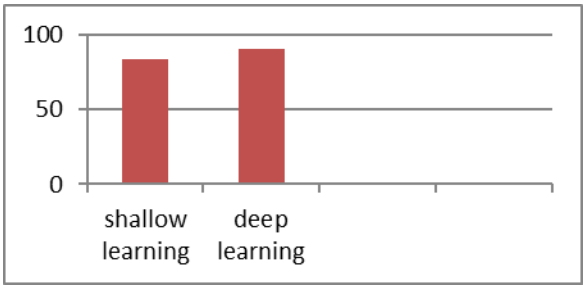

Fig 1: Performance evaluation

Here, we propose a deep learning scheme to infer the possible diseases given the questions of health seekers. Compared to shallow learning, deep learning has better performance.

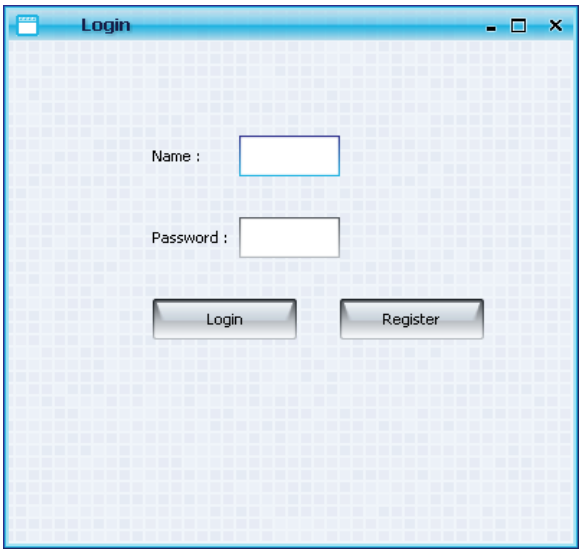

Fig 2: User login form

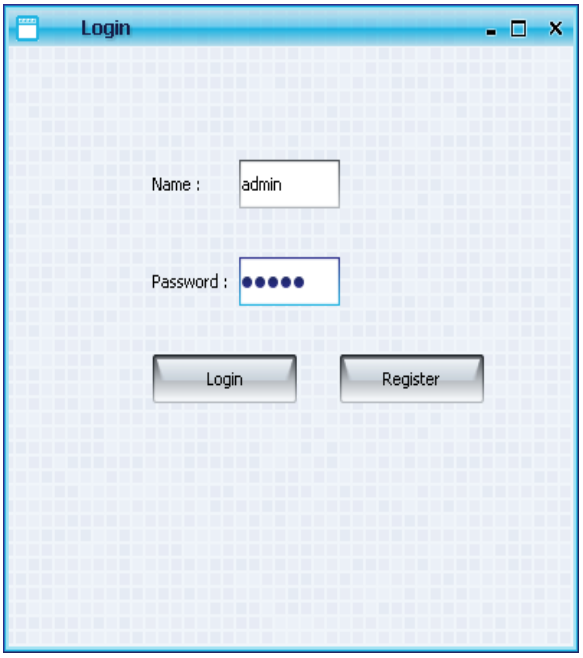

Fig 3: User registration form

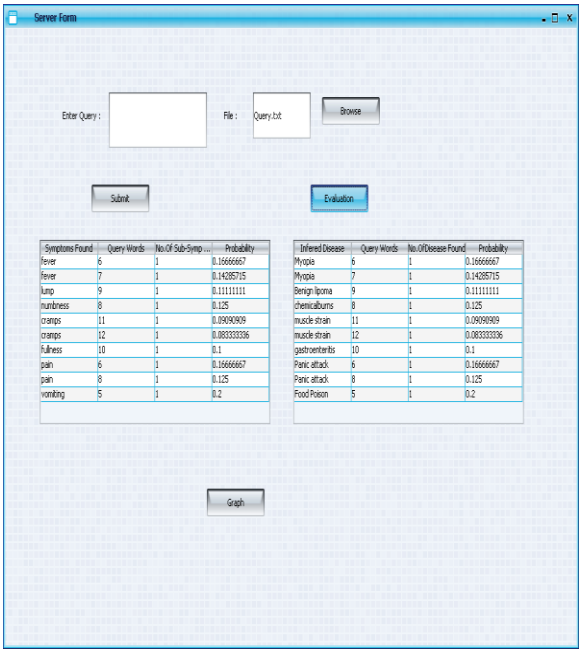

Fig 6: Disease inference evaluation form based on their symptoms

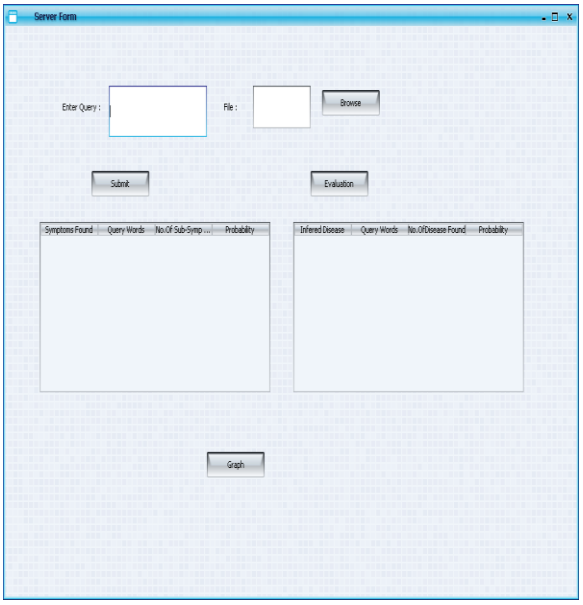

Fig 4: User query entry form

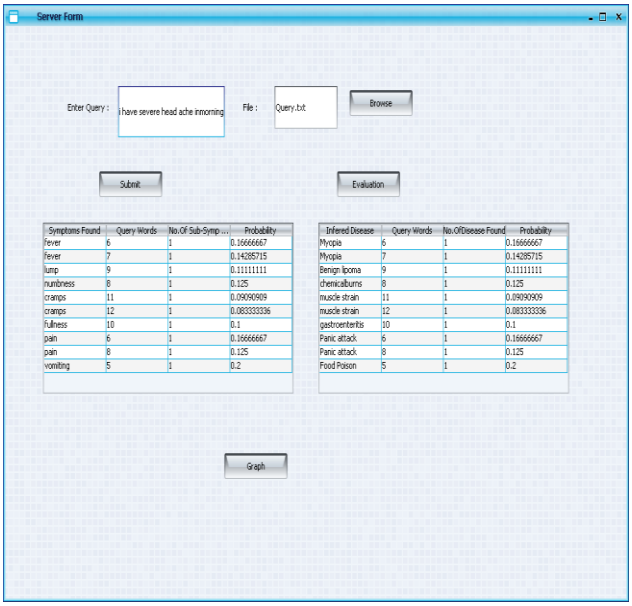

Fig 5: Disease inference form based on their symptoms

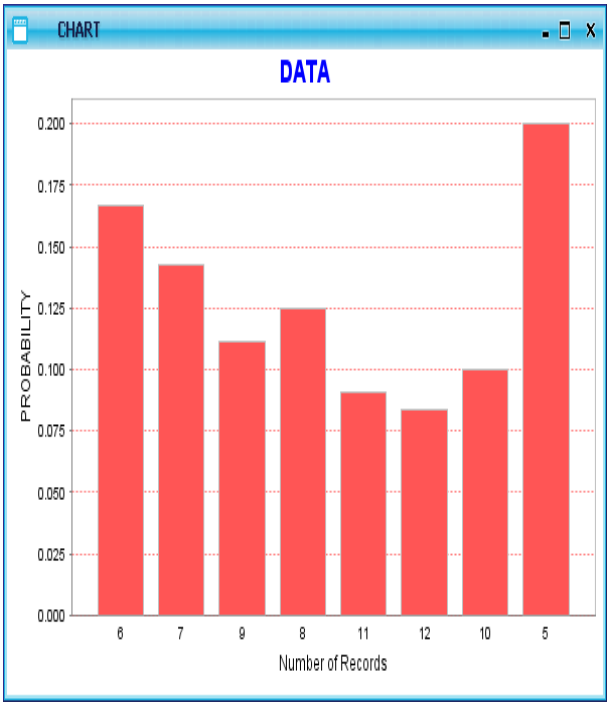

Fig 7: Performance evaluation graph based on their symptoms

IV. CONCLUSION

Classical deep learning architectures are densely connected and also the node number in every hidden layer is tediously modified. In contrast, the developed model is sparsely associated with enhanced learning efficiency and the amount of hidden nodes is instantly determined. This plan is built via alternative signature mining and pre-learning an incremental way. It permits without supervision feature gaining knowledge from other number of disease types. Therefore, it is generalizable and scalable as in comparison to previous disease inference using shallow learning approaches that are usually trained on hospital produced patient records with structured fields. This paper first carried out user study to evaluate the seeker needs. Our current model is not able to recognize discriminate features for every specific disease. Later on, we propose to pay more attention to that. This gives the information of community based health services. After that it presented a sparsely connected deep learning plan that has the capacity to infer the potential illnesses because of the questions of health seekers.

REFERENCES

[1] A. Khosla, Y. Cao, C. C.-Y. Lin, H.-K. Chiu, J. Hu, and H. Lee, "An integrated machine learning approach to stroke prediction," in Proc. 16th ACM SIGKDD Conf. Knowl. Discovery Data Mining, 2010, pp. 183–192.

[2] B. Koopman, P. Bruza, L. Sitbon, and M. Lawley, "Evaluating medical information retrieval," in Proc. 34th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2011, pp. 1139–1140.

[3] C. B. Akgul, D. Unay, and A. Ekin, "Automated diagnosis of alzheimer's disease using image similarity and user feedback," in Proc. ACM Int. Conf. Image Video Retrieval, 2009.

[4] D. A. Davis, N. V. Chawla, N. Blumm, N. Christakis, and L.Barabasi, "Predicting individual disease risk based on medical history," in Proc. 13th Int. Conf. Inf. Knowl. Manage., 2008.

[5] D. Zhu and B. Carterette, "An adaptive evidence weighting method for medical record search," in Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2013.

[6] F. Wang, N. Lee, J. Hu, J. Sun, S. Ebadollahi, and A. Laine, "A framework for mining signatures from event sequences and its applications in healthcare data," IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, Feb. 2013.

[7] F. Wang, N. Lee, J. Hu, J. Sun, and S. Ebadollahi, "Towards hetero- geneous temporal clinical event pattern discovery: A convolution approach," in Proc. 18th ACM SIGKDD Conf. Knowl. Discovery Data Mining, 2012.

[8] G. Luo and C. Tang, "On iterative intelligent medical search," in Proc. 31st Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2008.

[9] I. Batal, L. Sacchi, R. Bellazzi, and M. Hauskrecht, "A temporal abstraction framework for classifying clinical temporal data," in Proc. Amer. Med. Informat. Assoc., 2008.

[10] J. Zhou, L. Yuan, J. Liu, and J. Ye, "A multi-task learning formulation for predicting disease progression," in Proc. 17th ACM SIGKDD Conf. Knowl. Discovery Data Mining, 2011.

[11] J. Zhou, J. Liu, V. A. Narayan, and J. Ye, "Modeling disease progression via fused sparse group lasso," in Proc. ACM SIGKDD Conf. Knowl. Discovery Data Mining, 2012.

[12] L. Nie, T. Li, M. Akbari, J. Shen, and T.-S. Chua, "Wenzher: Comprehensive vertical search for healthcare domain," in Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2014.

[13] M. Shouman, T. Turner, and R. Stocker, "Using decision tree for diagnosing heart disease patients," in Proc. 9th Australasian Data Mining Conf., 2011.

[14] M. Galle "The bag-of-repeats representation of documents," in Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2013, pp. 1053–1056.

[15] N. Limsopatham, C. Macdonald, and I. Ounis, "Learning to combine representations for medical records search," in Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2013.

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

[17] P. Sondhi, J. Sun, H. Tong, and C. Zhai, Sympgraph: "A frame- work for mining clinical notes through symptom relation graphs" in Proc. 18th ACM SIGKDD Conf. Knowl. Discovery Data Mining, 2012.

[18] R. Fakoor, F. Ladhak, A. Nazi, and M. Huber, "Using deep learning to enhance cancer diagnosis and classification," presented at the Int. Conf. Mach. Learn., Atlanta, GA, USA, 2013.

[19] S. Doan and H. Xu, "Recognizing medication related entities in hospital discharge summaries using support vector machine," in Proc. Int. Conf. Compute Linguistics, 2010, pp. 259–266.

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

[21] T. D. Wang, C. Plaisant, A. J. Quinn, R. Stanchak, S. Murphy, and B. Shneiderman, "Aligning temporal data by sentinel events: Discovering patterns in electronic health records," in Proc. SIGCHI Conf. Human Factors Comput. Syst., 2008.

[22] 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.

[23] X. Huang and Q. Hu, "A Bayesian learning approach to promoting diversity in ranking for biomedical information retrieval," in Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2009.

[24] Y. Zhang and B. Liu, "Semantic text classification of disease reporting," in Proc. 30th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2007.

[25] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," IEEE Trans. Pattern Anal. Mach.Intell, vol. 35, Aug. 2013.