Wikipedia-Based Automatic Diagnosis Prediction in Clinical Decision Support Systems

Danchen Zhang[1], Daqing He[1], Sanqiang Zhao[1], Lei Li[2]
[1]School of Information Sciences, University of Pittsburgh, USA
[2]Department of Information Management, Nanjing University of Science & Technology, China

Abstract
When making clinical decisions, physicians often consult biomedical literatures for reference. In this case, an effective clinical decision support system, provided with a patient's health information, should be able to generate accurate queries and return to the physicians with useful articles. Related works in the Clinical Decision Support (CDS) track of TREC 2015 demonstrated the usefulness of knowing patients' diagnosis information for supporting more effective retrieval, but the diagnosis information is often missing in most cases. Furthermore, it is still a great challenge to perform large-scale automatic diagnosis prediction. This motivates us to propose an automatic diagnosis prediction method to enhance the retrieval in a clinical decision support system, where the evidence for the prediction is extracted from Wikipedia. Through the evaluation conducted on 2014 CDS tasks, our method reaches the best performance among all submitted runs. In the next step, graph structured evidence will be integrated to make the prediction more accurate.
Keywords: Automatic diagnosis prediction; clinical decision support system; text mining
Citation: Editor will add citation

Contact: {daz45, dah44, saz31}@pitt.edu, lileiwelldone@gmail.com

1 Introduction

To accurately make clinical decisions, physicians often consult published biomedical articles for reference. Many biomedical literature search engines, such as PubMed[1], have been developed to facilitate this information access process. However, most platforms only support keywords retrieval, and thus they impose high requirements for the physicians to construct accurate queries. Nevertheless, physicians usually have very complicated information needs and limited query refinement time. Therefore, an effective decision support system, provided with the patient health information, should be able to automatically expand physicians' original queries so that useful results can be returned.

Under the goal of exploring the information retrieval technologies to support this kind of clinical decision support (CDS) tasks, Text Retrieval Conference (TREC) hosted CDS track[2] in 2014 and 2015. Sixty electronic health records (EHRs), after removing patient personal and sensitive information, were released as sample patient information, and the participants of the track needed to provide relevant biomedical articles for three types of clinical questions (Simpson, et al., 2014):

- What is the patient's diagnosis?
- What tests should the patient receive?
- How should the patient be treated?

The CDS track results in 2015 showed that the correctly identified diagnosis information, after being integrated with the original query, can significantly improve the retrieval effectiveness -- the mean infNDCG increasing from 20.99% to 28.70%, and median infNDCG increasing from 22.88% to 32.12% (Roberts, et al., 2015). However, such diagnosis information is not often available, which significantly affect the retrieval performance. This motivates us to explore an automatic diagnosis prediction method to enhance the retrieval in CDS tasks.

When examining the content, EHRs can provide patients' information such as: (1) demographic information including age and gender, and (2) disease related information, including disease history, current and past symptoms, and testing results (see Figure 1 for an example). However, the diagnosis information about current disease is not in the EHRs, and it is difficult to obtain reliable diagnosis of the disease without expert knowledge.

Therefore, in this paper, we study the usefulness of utilizing Wikipedia as an external recourse to be combined with disease related medical concepts extracted from EHRs for predicting the disease

[1] http://www.ncbi.nlm.nih.gov/pubmed
[2] http://trec-cds.appspot.com/

diagnosis. The two research questions we explored are: (1) How accurate is Wikipedia-based automatic diagnosis prediction? and (2) How well can query expansion using predicated diagnosis improve retrieval effectiveness?

```
▼<topic number="1" type="diagnosis">
  ▼<description>
     A 58-year-old African-American woman presents to the ER with episodic pressing/burning anterior
     chest pain that began two days earlier for the first time in her life. The pain started while she
     was walking, radiates to the back, and is accompanied by nausea, diaphoresis and mild dyspnea, but
     is not increased on inspiration. The latest episode of pain ended half an hour prior to her
     arrival. She is known to have hypertension and obesity. She denies smoking, diabetes,
     hypercholesterolemia, or a family history of heart disease. She currently takes no medications.
     Physical examination is normal. The EKG shows nonspecific changes.
   </description>
  ▼<summary>
     58-year-old woman with hypertension and obesity presents with exercise-related episodic chest pain
     radiating to the back.
   </summary>
 </topic>
```

Figure 1. An example topic from CDS task in TREC 2014.

2 Related works

Medical text retrieval is a very popular topic. TREC hold Medical Records tracks in 2011 and 2012 targets on medical records retrieval, and then hold clinical decision support tracks in 2014 and 2015 targeting on the medical literatures retrieval. Both tracks have found that query expansion is a common and effective procedure to boost the retrieval system performance (Voorhees, et al., 2012; Roberts et al., 2015). For example, Soldaini et al. (2015) proposed to expand the query with selected terms extracted from pseudo relevant documents. Song, et al. (2015) proposed to expand the query with terms extracted from top 10 Google returned webpage titles. In addition, NLP and machine learning techniques are also frequently integrated into this area. For example, Wang et al., (2015) assign each term in the query with a Part-Of-Speech related weight, significantly improved the retrieval performance. Choi et al., trained classifiers to rank retrieved documents based on documents types (therapy versus non-therapy, diagnosis versus non-diagnosis), which gives best run in TREC 2014 CDS track.

Different from previous works, we propose to predict diagnosis automatically, and use it to expand the query. Esfandiari et al. (2015) gave a comprehensive review of the studies in area of automatic diagnosis prediction. Through their summarization, most studies regard diagnosis prediction as a classification task. For example, Yeh et al. (2011) utilized the patients' history diseases, blood test results and physical exam results as the features, and trained classifiers to predict the probability of getting a cerebrovascular disease. Tsipouras et al. (2008) used the rules extracted from a decision tree to construct a fuzzy model to do automatic diagnosis prediction of coronary artery disease.

Past works in automatic diagnosis prediction area usually target on sensitively and accurately predicting a small range of diseases. However, in this paper, we need not to predict the diagnosis so sensitively, but the system is expected to cover the diseases as many as possible. Meanwhile, the past studies may concentrate more on diagnosis prediction with well-formatted features, while the diagnosis prediction procedure in this paper post a high requirement for free-text processing ability. This is the difference between this work and past studies. To our best knowledge, it is the first automatic method on large-scale diagnosis prediction.

3 Methodology

Our method consists of three steps. Firstly, all the medical concepts appearing in EHRs are extracted, then they are used in Wikipedia to predict possible diagnosis. Finally, the identified diagnosis is used to expand the query for retrieval. We will talk about these three steps in details below.

3.1 Disease-related medical concepts extraction

Our medical concept extraction from EHRs utilizes MetaMap, a popular tool in medical text mining area that was released by NIH. In this paper, only medical concepts belonging to the semantic group Disorders[3] are kept, which consists of 12 semantic types, all describing the functional abnormality or disturbance.

[3] See more in https://metamap.nlm.nih.gov/Docs/SemGroups_2013.txt

3.2 Wikipedia-based automatic diagnosis prediction (WADP)

To predict a diagnosis based on the extracted medical concepts, we use Wikipedia as the external knowledge. Typically, a wiki page talking about a disease contains the information of symptoms, causes, and pathophysiology. Because the medical concepts extracted from EHRs are usually about the same information of the disease, it is more likely to these concepts appearing in the wiki page. Therefore, by examining the appearance of these medical concepts, we can make predictions about the diagnosis of the correct disease. Based on this hypothesis, our method uses those medical concepts to construct a query to search in the Wikipedia collection. Based on the returned ranked list of wiki pages, we check whether the title of the page can be recognized as a disease based on MetaMap. If the title of the wiki page is a disease, then this title is the diagnosis of the disease. Otherwise, the next wiki page in the ranked list is checked, until we find a wiki page whose title can be recognized as a disease.

Our Wikipedia collection consists of only English Wikipedia pages, which was downloaded on March 5th, 2016 as a dump of enwiki[4]. We only kept the title and the page content, and removed the tags, Reference, External Link and See Also sections. The Wikipedia data were indexed by Indri, and the search was performed using unigram-based KL divergence language model with Dirichlet smoothing. Dirichlet smoothing has a parameter μ, indicating the smoothing degree.

3.3 Query Expansion with Diagnosis

Once the diagnosis is obtained, the original query, which consists of disease-related medical concepts extracted from patient EHRs, is combined with the diagnosis information. In Indri query language, such expansion is conducted as follows:

#weight (α #combine(original query) (1- α) #combine(predicted diagnosis))

where α is the weighting parameter, ranging from 0 to 1. If a medical concept has multiple terms, we use #uw(term1 term2 …) to indicate that these terms would appear adjacent to each other.

4 Experiments

4.1 Dataset and the baseline

The target document collection for CDS task consists of 733,138 full-text medical articles from PubMed Central. Our index of the documents only contains the title, abstract and content of each article. Indri was again utilized to index the collection, and we performed stop word removal and stemming with Porter stemmer.

Both 2014 and 2015 CDS tasks provide 30 EHR topics. Each topic consists of a description and a summary. We extracted the medical concepts from both areas. Since 2015 CDS task provides the diagnosis for 20 topics, we use it as the training, and tested the system on 2014 CDS data.

Topic Id	Predicted Diagnosis	Topic Id	Predicted Diagnosis
11	Hypothyroidism**	21	Cyclospora cayetanensis
12	Meningitis*	22	Chronic pulmonary aspergillosis*
13	Epiglottitis**	23	Dengue fever**
14	Paroxysmal nocturnal hemoglobinuria**	24	Pneumonia*
15	Atrial fibrillation*	25	Meningitis
16	Asthma	26	Ectopic pregnancy**
17	Cervical cancer*	27	Iron-Deficiency Anemia**
18	Heart failure*	28	Lyme disease*
19	Chronic obstructive pulmonary disease**	29	Kawasaki disease**
20	Myoclonus	30	Shoulder problem

Table 1. Predicted diagnosis on CDS task in TREC 2015, ** means exactly same with the true diagnosis, * means closely related predictions.

First, we tuned the Dirichlet smoothing parameter μ_1 in Wikipedia based diagnosis prediction module. Since a μ value from 1,000 to 100,000 produced nearly the same prediction for all 20 topics, we set μ_1 with Indri's default value 2500. We then tuned the other two parameters based on 2015 CDS task and measured the retrieval performance using infNDCG: (1) the weighting parameter α in query

[4] https://dumps.wikimedia.org/enwiki/20160701/

expansion; and (2) the other Dirichlet smoothing parameter μ_2 in the biomedical literature retrieval module. Through tuning, the value for μ_2 = 1000 and α = 0.4.

In the experiment, query set of baseline was identical to WADP except that it does not have the query expansion based on predicated diagnosis.

4.2 How accurate is WADP's prediction?

From Table 1 we can see that, among 20 predictions, WADP can predict eight diagnoses with the same diseases as the true diagnoses. At the same time, it can make closely related predictions for another seven. Although it is not perfect, it can generate reasonable predictions for 75% of the disease prediction cases we tested. So WADP is reasonably accurate. Therefore, the question turns to whether these predictions are useful or not, which will be examined in the next section.

4.3 How well can WADP-based query expansion improve retrieval effectiveness?

Figure 2 shows that, comparing to the baseline that does not have the query expansion, the performance of the topics with correctly predicted diagnoses have been greatly improved, even among those seven topics with partially correct predictions, their retrieval performance has been improved too.

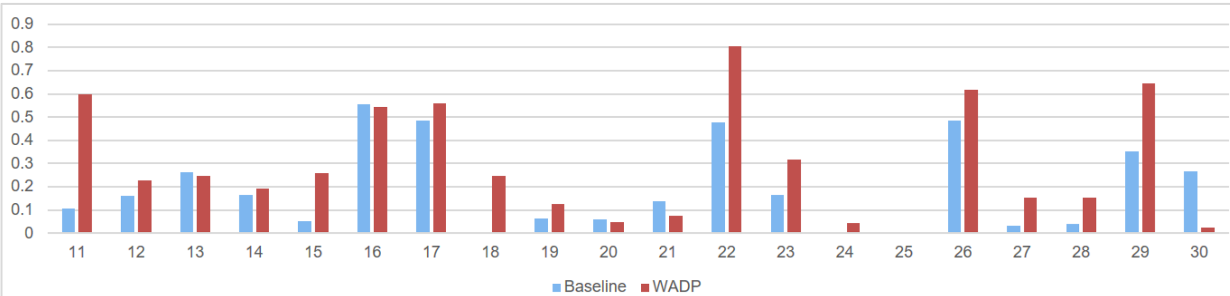

Figure 2. System performance on infNDCG, from topic 11 to 30 of CDS task in TREC 2015.

To precisely evaluate the effect coming from the predicted diagnosis, we conducted the Wilcoxon Signed Ranks Test. As shown in Table 2, a significant difference is found between the Baseline and WADP, with p-value<0.001 on both infNDCG and infAP, but there is no significant difference on iP10. This demonstrates the usefulness of WADP-based query expansion for the clinical decision support system.

Besides, we compare our method with SNUMedinfo, the best run submitted to CDS task in TREC 2014 (Choi, et al., 2014). Although WADP and SNUMedinfo have nearly same performance on infNDCG, but WADP outperforms SNUMedinfo on iP10 and infAP. This implies our method outperforms the state-of-the-art system.

Models	infNDCG	iP10	infAP
Baseline	22.95%	38.44%	5.85%
WADP	**26.98%**	**38.11%**	**9.3%**
SNUMedinfo (Choi, et al., 2014)	26.74%	36.33%	6.59%

Table 2. Performance on CDS task of TREC 2014.

We think that the original queries extracted from the EHRs mainly contain terms about the symptoms or disease history, terms describing patient's situation. WADP, on the other hand, clearly states the predicted diagnosis, which is the real information need of the physicians related to diagnosis, test, and therapy. Thus, WADP makes the queries more accurate.

5 Conclusion

In this paper, we propose a novel mechanism, Wikipedia-based automatic diagnosis prediction, is for enhancing the clinical decision related retrieval tasks. Given patients' disease-related information, we search through the Wikipedia collection to obtain a prediction of disease with the highest probability, and use it to expand the original query. our experiment trained on CDS task of TREC 2015 and tested on CDS task of TREC 2014 demonstrates the usefulness of our method, which makes it outperform the state of the art methods. For the future work, we plan to incorporate better prediction models such as Markov Random Field to make the prediction more robust and accurate.

6 References

Choi, S., & Choi, J. (2014). SNUMedinfo at TREC CDS track 2014: Medical case-based retrieval task.

Esfandiari, N., Babavalian, M. R., Moghadam, A. M. E., & Tabar, V. K. (2014). Knowledge discovery in medicine: Current issue and future trend. Expert Systems with Applications, 41(9), 4434-4463.

Roberts, K., Simpson, M. S., Voorhees E., & Hesh R. W. (2015). Overview of the TREC 2015 Clinical Decision Support Track.

Simpson, M. S., Voorhees E., & Hersh W. (2014). Overview of the TREC 2014 clinical decision support track.

Soldaini, L., Cohan, A., Yates, A., Goharian, N., & Frieder, O. (2015). Retrieving medical literature for clinical decision support. European Conference on Information Retrieval.

Song, Y., He, Y., Hu, Q., & He, L. (2015). ECNU at 2015 CDS Track: Two Re-ranking Methods in Medical Information Retrieval. Proceedings of the 2015 Text Retrieval Conference.

Tsipouras, M. G., Exarchos, T. P., Fotiadis, D. I., Kotsia, A. P., Vakalis, K. V., Naka, K. K., & Michalis, L. K. (2008). Automated diagnosis of coronary artery disease based on data mining and fuzzy modeling. IEEE Transactions on Information Technology in Biomedicine 12.4 (2008): 447-458.

Voorhees, E. M., & Hersh, W. R. (2012). Overview of the TREC 2012 Medical Records Track. TREC. 2012.

Wang, Y., Wu, S., Li, D., Mehrabi, S., & Liu, H. (2016). A Part-Of-Speech term weighting scheme for biomedical information retrieval. Journal of Biomedical Informatics 63 (2016): 379-389.

Yeh, D. Y., Cheng, C. H., & Chen, Y. W. (2011). A predictive model for cerebrovascular disease using data mining. Expert Systems with Applications, 38(7), 8970-8977.