Discovering Syndrome Regularities in Traditional Chinese Medicine Clinical by Topic Model

Jialin Ma[1,2*], Zhijian Wang[1]
[1] College of Computer and Information, Hohai University, Nanjing, China
majl@hyit.edu.cn, 51077061@qq.com
[2] Huaiyin Institute of Technology, Huaian, China

Abstract. Traditional Chinese Medicine (TCM) as one of most important approach for disease treatment in China for thousands of years. Lots of experience of famous experts in TCM is recorded in medical bibliography. The first vital work for TCM doctor is to diagnose the disease by the patients' symptoms, and then predict the syndromes which the patient has. Generally, this process reflects the medical skill of the TCM doctors. Therefore, TCM diagnose is easy to misdiagnose and difficult to master for TCM doctors. In this paper, we proposed a probabilistic model—the symptom-syndrome topic model (SSTM) to explore connected knowledge between symptoms and syndromes. In the SSTM, symptom-syndrome are modeled by generative process. Finally, we conduct the experiment on the SSTM. The results show that the SSTM is effective for mining the syndrome regularities in TCM data.

Keywords: TCM, syndrome, topic model, SSTM

1 Introduction

Traditional Chinese Medicine (TCM) as one of a important and independent medical theoretical and practical system has been existing thousands of years[1]. Furthermore, TCM is considered as one of an important complementary medical system to modern biomedicine. Especially in recent years, TCM is going abroad, more and more foreigners become accept and enjoy the treatment or health care of TCM[2]. Different from modern biomedicine, TCM doctors rarely resort to medical equipments to diagnose diseases. They usually utilize four diagnostic methods(observation, listening, interrogation, and pulse-taking) to understand the pathological conditions. Human body is regarded as a synthetic system in TCM. Therefore, TCM focuses on analysing the macro-level functional information of patients and utilize the much Chinese traditional naive philosophical theories and thoughts to adjust patients' body health and ecological balance. TCM emphasize individualized diagnosis and treatment which is most different from modern biomedicine[3].

In the past thousands years of the Chinese history, a large number of TCM doctor's experience, such as clinical case, ancient textbooks, classical ancient prescriptions etc., have been recorded. These records imply rich TCM knowledge. Mining regularities from TCM clinical records is one of the main approaches for TCM physicians to improve their clinical skills and empirical knowledge. Data mining is a useful computing approach for discovering hidden knowledge from large-scale data[4], and could be a potential solution for this issue. However, the huge scale of ancient TCM record is existing as the style of texts. Those experience and knowledge is described

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_15

by nature language. It is well known understanding semantics doesn't completely break through in the field of artificial intelligence. Fortunately, many researchers have paid attentions on the TCM record mining, such as[2, 4, 5].

The first and fundamental problem for TCM doctors is to diagnose patients' disease by four diagnostic methods(observation, listening, interrogation, and pulse-taking). Misdiagnosing would be lead to fatal results. Mining and learning diagnosing knowledge or experience from TCM record by computer is significative for doctors. It can guide or aid TCM doctors, especially young doctors to master diagnostic knowledge and experience[6]. In this paper, we proposed a probabilistic model—the symptom-syndrome topic model (SSTM) to explore latent knowledge between symptoms and syndromes. In the SSTM, symptom-syndrome is modelled by generative process which can help to acquire diagnosing knowledge or experience from TCM record. Finally, We conduct experiment on the SSTM. The results show that SSTM can extract effective latent semantics information and relation.

The paper is organized as follows: Section 2 reviews the related work about SMS spam filtering technologies. Section 3 presents our method in detail. Section 4 shows the experiments and discussion. Finally, we conclude and discuss further research in Section 5.

2 Related works

Topic model which is based on probability statistics theories, can detect latent semantic structure and information in large-scale documents[7]. Latent Semantic analysis (LSA) is one of the famous representative method in the early time[8]. It depends on capture word co-occurrence in documents. Therefore, LSA can bring semantic dimensionality between the text and words. Moreover, probabilistic latent semantic analysis (PLSA) is the further improvement of the LSA[9]. In PLSA, a document is regarded as a mixture of topics, while a topic is a probability distribution over words. In order to improve the defects of the PLSA, LDA proposed for the first time by Blei in 2003[10], which added Dirichlet priors in the distributions. LDA is a more completely generative model and achieves great successes in text mining and other artificial intelligence domains. With the rapid development of internet and social media, mass of short texts have been produced. Short texts data analysis (such as microblog) become advanced research hotspot. Many researchers are eager to mine social media date by topic model[11]. But the thorny problem is the lack of statistics information about terms in the short texts. Except directly applying the standard LDA, many improved topic model are researched to apt at short texts[12]. The famous researches in this aspect are Author-Topic Model(ATM)[13] and Twitter-LDA[12]. Moreover, Yan etc al. proposed a Biterm Topic Model(BTM), which can learn topics over short texts by directly modeling the generation of biterms in the whole corpus[14,22].

Many researchers have been eager to Knowledge discovering and data mining (KDD) in biomedicine field for a long times[2]. By contrast, TCM is to become a research hotspot in the recent years. The reviewing references of TCM mining are[4, 15-17]. Zhang et al. [1] proposed a data mining method, called the Symptom-Herb-Diagnosis topic (SHDT) model, to automatically extract the common relationships among symptoms, herb combinations and diagnoses from large-scale CM clinical data. Jiang et al.[2] apply the Link Latent Dirichlet Allocation (LinkLDA), to automatically extract the latent topic structures which contain the information of both symptoms and their corresponding herbs. Yao et al.[18] proposed a framework which mines the treatment pattern in TCM clinical cases by using probabilistic topic model and TCM domain knowledge. The framework can reflect principle rules in TCM and improve function prediction of a new prescription. They evaluate our model on real world TCM clinical cases.

These mentioned studies have devoted to mining knowledge from TCM case data. They focus on knowledge the compatibility of TCM, diagnose law, or the rules of ‘‘Li-Fa-Fang-Yao”[19].

3 Our Work

Different from modern biomedicine, TCM doctors rarely resort to medical equipments to diagnose diseases. They usually utilize four diagnostic methods(observation, listening, interrogation, and pulse-taking) to understand the pathological conditions. Misdiagnosing would be lead to fatal results. Mining and learning diagnosing knowledge or experience from TCM records, literatures, or clinic cases by computer is significative for doctors. Nevertheless, understanding semantics from TCM records isn’t a easy thing in the state of the art.

Conventional topic models, like PLSA[9] and LDA[10], reveal the latent topics within corpus by implicitly capturing the document-level word co-occurrence patterns. We propose a new topic model SSTM to capture the relationship between symptoms and the syndromes from TCM clinical data. In the SSTM, symptoms is distribution on syndrome. Different from LDA, explicit variable symptoms is divided into cardinal symptoms in diagnosis and secondary symptoms. A cardinal symptom is main feature for a specific disease, but the secondary symptoms is subordinate for the disease. This is more accord with the process of TCM diagnose.

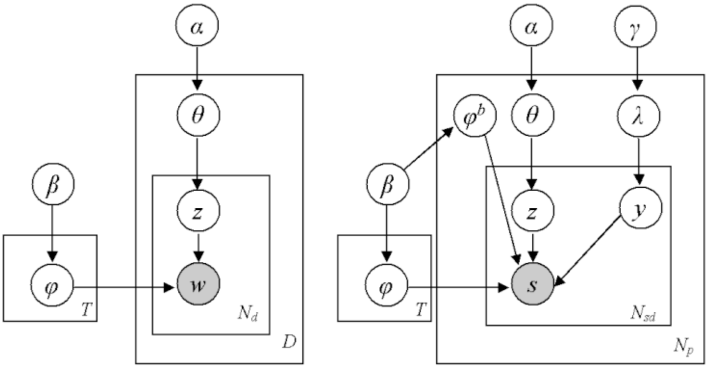

Fig. 1. Graphical models for (a)LDA,, (b) SSTM

Fig. 1(a) shows the graphical model for the “standard topic model”(LDA). *D* is the number of documents in the corpus and document *d* has N_d words. The process includes two steps: first, assign a topic number from document-topic distribution *θ*;

then, draw a word from topic-word distribution φ. All documents share T topics. Document-topic and topic-word distributions all obey the multinomial distributions, and each of them is governed by symmetric Dirichlet distribution. α and β are hyper-parameters of symmetric Dirichlet priors for θ and φ. Parameters θ and φ can be obtained through a Gibbs sampling.

Fig. 1(b) shows the graphical model for the "Symptom-Syndrome Topic Model"(SSTM). N_p is the total number of clinical cases, T is the topic number, z is from the corpus-level topic distribution θ. N_{sd} is total number of syndrome the patient have. Let φ denote the symptom distribution for topics and φ_b denote the symptom distribution for secondary symptoms. Let λ denote the Bernoulli distribution which controls the indicator y for the choice between b cardinal symptom and secondary symptom. Φ, θ, *and* φ_b all obey multinomial distributions, each of term is drawn from symmetric Dirichlet (β), and Dirichlet (α) respectively. λ is drawn from Beta (γ).

The probability of a symptom s is described as the follow(14):

$$p(s) = p(y=0)\sum_{z} p(z)p(s \mid z) + p(y=1)p(s \mid y=1) \tag{1}$$

The generative probability of a *patient* is expressed as:

$$p(patient) = \sum_{n=1}^{N_{sd}}\sum (p(y=0)\sum_{z} p(z)p(s \mid z) + p(y=1)p(s \mid y=1)) \tag{2}$$

4 Experiment

In this following, we propose experiment on SSTM in order to verify the effect of mining syndrome regularities in TCM data.

The CTM data come from China National Scientific Data Sharing Platform for Population and Health. We selected one of the databases: TCM Clinical Database of Diabetes. It has been classified into 19 subjects and 4351 records. We just focus on the II Diabetes. The data has the number of 1162 records.

In the experiment, α = 50/ T and β= 0.01, which are common settings in the literature[20]. γ is a prior of Bernoulli distribution, we set γ=0.5, it refers to another similar study[21]. We conducted the experiment on the SSTM and set T=7 by observing the data. We select the first five symptoms for each topic. The results are showed in Table 1.

Table 1. The first five symptoms for each topic

Topics	Symptoms
Topic1	Weak, sweating, shortness of breath, spontaneous perspiration, thirst
Topic2	Thirst, tough, yellow, constipine, polyphagy, feel, upset
Topic3	emaciation, weak, thirst, polydipsia, diuresis
Topic4	dizziness, thirst, palpitation, Shortness of Breath, dry mouth
Topic5	hiccough, pale tongue, prospermia, chilly, backache
Topic6	weak,pectoralgia, dark tongue, umbness of limbs, sluggish pulse
Topic7	thirst, dizziness, weak, exhausted, spontaneous perspiration

We invited experienced TCM doctors to analysis these results about Table 1. They considered those symptoms topic 1-topic 7 are related with some syndromes, for example, topic 1 is similar to qi and yin deficiency, topic 1 is similar to deficiency of yin and excessive heat syndrome, topic 5 is similar to deficiency of both yin and yang, etc.

5 Conclusions and future Work

TCM treat patient according to syndrome differentiation. Therefore, predicting the syndromes is vital work in the diagnosing. TCM is one of important medical branch theory system. In thousands of years, huge number of medical records that is recorded by nature language is accumulated by famous TCM doctors. Our work devote to mining the syndrome rules in these records. The proposed probabilistic model—symptom-syndrome topic model (SSTM) is effective to capture the connected knowledge between symptoms and syndromes. The further work is to continue to prefect SSTM and analysis the detail and complex relationship between symptoms and syndromes.

Acknowledgments.

This work was supported by the Graduate Student Scientific Research and Innovation Project of Jiangsu Province, China (China Central University Basic Scientific Research Business Fund, Hohai university, Grant No.: 2015B38314), and Science and Technology Projects of Huaian (Grant No.: HAS2015033), the University Science Research Project of Jiangsu Province(Grant No.:15KJB520004), and the Science and Technology Projects of Huaian (Grant No.:HAG2015060).

References

1. Zhang, X., Zhou, X., Huang, K., Feng, Q., Chen, S., and Liu, B.: Topic model for Chinese medicine diagnosis and prescription regularities analysis: Case on diabetes, Chinese Journal of Integrative Medicine, 17, 307 (2011).

2. Jiang, Z., Zhou, X., Zhang, X., and Chen, S.: Using link topic model to analyze traditional Chinese Medicine
Clinical symptom-herb regularities, in IEEE International Conference on E-Health Networking, Applications
and Services, pp. 15 (2012).

3. Zhou, X., Chen, S., Liu, B., Zhang, R., Wang, Y., Li, P., & Yan, X.: Development of traditional
Chinese medicine clinical data warehouse for medical knowledge discovery and decision support,
Artificial Intelligence in Medicine, 48, 139 (2010).

4. Zhou, X., Peng, Y., and Liu, B.:Text mining for traditional Chinese medical knowledge discovery: A survey, Journal of Biomedical Informatics, 43, 650 (2010).

5. Liu, C. X., and Shi, Y.: Application of data-mining technologies in analysis of clinical literature on traditional Chinese medicine, Chinese Journal of Medical Library & Information Science (2011).

6. Liu, B., Zhou, X., Wang, Y., Hu, J., He, L., Zhang, R., Chen, S., and Guo, Y.: Data processing and analysis in real-world traditional Chinese medicine clinical data:challenges and approaches, Statistics in Medicine, 31, 653 (2012).

7. Blei, D. M.: Probabilistic topic models, Communications of the ACM, 55, 77 (2012).

8. Thomas K. Landauer, P. W. F., Darrell Laham.:An Introduction to Latent Semantic Analysis, Discourse Processes, 25, 259 (1998).

9. Hofmann, T.: Probabilistic latent semantic indexing, in Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, ACM, pp. 50 (1999).

10. Blei, D. M., Ng, A. Y., and Jordan, M. I.: Latent dirichlet allocation, the Journal of machine Learning research, 3, 993 (2003).

11. Hong, L., and Davison, B. D.: Empirical study of topic modeling in Twitter, Proceedings of the Sigkdd Workshop on Social Media Analytics, 80 (2010).

12. Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E. P., Yan, H., and Li, X.:Comparing Twitter and Traditional Media Using Topic Models, in In ECIR, pp. 338 (2011).

13. Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P., The author-topic model for authors and documents: in Proceedings of the 20th conference on Uncertainty in artificial intelligence, AUAI Press, pp. 487 (2004).

14. Yan, X., Guo, J., Lan, Y., and Cheng, X.: A biterm topic model for short texts, in Proceedings of the 22nd international conference on World Wide Web, International World Wide Web Conferences Steering Committee, pp. 1445 (2013).

15. Yi, F., Wu, Z., Zhou, X., Zhou, Z., and Fan, W.: Knowledge discovery in traditional Chinese medicine: State of the art and perspectives, Artificial Intelligence in Medicine, 38, 219 (2006).

16. Lukman, S., He, Y., and Hui, S. C., Computational methods for Traditional Chinese Medicine: A survey, Computer Methods & Programs in Biomedicine, 88, 283 (2007).

17. Wu, Z., Chen, H., and Jiang, X.: 1 – Overview of Knowledge Discovery in Traditional Chinese Medicine 1, 1 (2012).

18. Yao, L., Zhang, Y., Wei, B., Wang, W., Zhang, Y., and Ren, X.: Discovering treatment pattern in traditional Chinese medicine clinical cases using topic model and domain knowledge, in IEEE International Conference on Bioinformatics and Biomedicine, pp. 191 (2014).

19. Liang, Y., Yin, Z., Wei, B., Wei, W., Zhang, Y., Ren, X., and Bian, Y.: Discovering treatment pattern in Traditional Chinese Medicine clinical cases by exploiting supervised topic model and domain knowledge, Journal of Biomedical Informatics, 58, 425 (2015).

20. Heinrich, G.: Parameter estimation for text analysis, Technical Report (2004).

21. Chemudugunta, C., Smyth, P., Steyvers, M.: Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model Vol. 19, MIT Press (2007).

22. Ma J, Zhang Y, Wang Z, et al.: A Message Topic Model for Multi-Grain SMS Spam Filtering. International Journal of Technology and Human Interaction (IJTHI), 2016, 12(2): 83-95.