Discovering Symptom-herb Relationship by Exploiting SHT Topic Model

Lidong Wang¹,a) Keyong Hu¹ Xiaodong Xu²

Received: July 7, 2017, Accepted: August 29, 2017

Abstract: TCM has been widely researched through various methods in computer science in past decades, but none digs into huge amount of clinical cases to discover the meaningful treatment patterns between symptoms and herbs. To meet the challenge, we explore the unstructured and intricate experiential data in clinical case, and propose a method to discover the treatment patterns by introducing a novel topic model named SHT (Symptom-Herb Topic model). Combinational rules are incorporated into the learning process. We evaluate our method on 3,765 TCM clinical cases. The experiment validates the effectiveness of our method compared with LDA model and LinkLDA model.

Keywords: Traditional Chinese Medicine, topic model, SHT model, combinational rules

1. Introduction

Traditional Chinese medicine (TCM) has been attracting more and more attention because of its complementary therapeutic effects to western medicines. TCM involves multiple types of entities, such as “herb,” “prescription” (a composition that consists of certain herbs), “symptom,” and “syndrome” (“Zheng” in Mandarin Chinese, a complex pattern of symptoms, which is used as a holistic summary of a patient’s status). Multiple types of relations can exist between these heterogeneous entities, such as composition relations between herbs and prescription, treatment relations between symptoms and herbs. TCM clinical cases describe how doctors diagnose and cure the disease. The unstructured TCM clinical cases involve the symptoms of a patient, the corresponding herbs, the initial visit information and the return visit information. How to dig into enormous clinical cases to mine the valuable relations between symptoms and herbs remains a challenging task.

Data mining approaches play critical roles in TCM related topics, such as new drug discovery [1], syndrome differentiation [2], [13], herbal combinational rule mining [4], [5], intelligent diagnosis [7], and patient classification [14]. Related works on relation extraction from the TCM literature are scarce. Wu et al. [6] was one of the pioneering works on this subject. The authors used a bootstrapping method to extract syndrome-disease associations from a corpus of data. In a recent work by Wang et al. [4], the authors created a herbal network based on attribute similarity calculation, and employed random walk based community detection to discover the latent combinational relations between two herbs. Chen et al. [7] designed a data mining approach to examine the relationship among symptoms, syndromes and herbs. This tripartite information network derived more accurate information than linking symptom and herb alone. Wan et al. [10] used a heterogeneous factor graph model (HFGM) to infer the multiple types of relations (e.g., herb-syndrome, herb-disease) from the entire corpus of TCM literature. Zhao et al. [13] found that a novel machine learning algorithm, minimum reference set-based multiple instance learning, was superior to other machine learning algorithms for TCM syndrome differentiation.

Recently, more and more researchers have adopted topic models to discover the relations between TCM objects. Lin et al. [3] proposed a symptom-herb-therapies-diagnosis topic model to diagnose the disease and administer appropriate drugs and treatments given a patient’s symptoms. Zhang et al. [8] proposed a hierarchical topic model (HSHT) to automatically extract the hierarchical latent topic structures with both symptoms and their corresponding herbs in the TCM clinical data. Yao et al. [11] employed Labeled-LDA (Labeled Latent Dirichlet Allocation) to mine treatment patterns in TCM clinical cases, but it only discovered the treatment patterns between herbs and disease by supervised model, which required labeled training data. The main goal of our paper is close to Zhang et al. [8]. However, we are different from theirs because: 1) We propose separate modeling for symptoms and herbs; 2) combinational rules between herbs are incorporated into the process of topic modeling, which is more consistent with TCM theory; that is, when two herbs are used together, their interaction should display their superiority over a single herb in the treatment of diseases.

In TCM, a syndrome can be inferred from symptoms. The process of the treatment is to determine syndromes by observing a patient’s symptoms and then determine appropriate herbs. Thus, we consider that the symptoms of a patient and the corresponding Chinese herbs have the same latent topic, which is known as “syndrome.” Based on this, we propose a topic model named SHT to automatically discover treatment patterns between symp-
toms and herbs from TCM clinical cases. After topic modeling, we can obtain the probability distribution of symptoms and its corresponding list of herbs in one topic (syndrome). The mining results provide valuable auxiliary information for TCM clinical diagnosis. Specifically, TCM doctors can use these associations to assist clinical treatment, since the mining results show the treatment patterns between symptoms and herbs. For example, to cure a patient with the disease “dyspnea with cough,” the doctor can navigate the results and find out the corresponding herbs for reference (see Table 2). In addition, the extracted relations may promote the understanding of TCM in Western countries.

2. SHT Topic Model

2.1 Topic Modeling

With respect to our topic model based method, a clinical case is considered as a “document.” A clinical case involves the symptoms of a patient and the corresponding Chinese herbs, so herbs and symptoms are treated as “words” in the document. TCM doctors have to select a set of herbs to cure a syndrome, which is reflected by a pattern of symptoms. In this way, a clinical case is a mixture of “topics,” syndromes are “topics” of the clinical case (“document”). And a “corpus” is a collection of clinical cases.

Let $C = \{c_1, c_2, \ldots, c_N\}$ be the set of clinical cases, $Z = \{z_1, z_2, \ldots, z_M\}$ be the set of syndromes, $H = \{h_1, h_2, \ldots, h_K\}$ be the set of herbs, $S = \{s_1, s_2, \ldots, s_L\}$ be the set of symptoms. The generative process of clinical cases is shown in Fig. 1.

This process is analogous to the generative process of probabilistic topic model [15]. Topic models, like Latent Dirichlet Allocation (LDA) [15], model each document as a mixture of underlying topics. Traditional LDA model generates a single word from one topic. Here, we generate a single symptom and a single herb from one “syndrome.” The generative processes for symptoms and herbs are very similar. Topics for symptoms are the probability distribution on symptom set, topics for herbs are the probability distribution on herb set. Note that Syndromes can be considered as the semantic bridge between symptoms and herbs. However, the efficiency of a single herb is usually limited in TCM. When two herbs are used together, their interaction should display their superiority over a single herb in the treatment of diseases, we say that these two herbs have compatibility rule. Thus, it is more meaningful to analyze paired herbs than a single herb.

Based on the above, we propose a novel topic model named SHT model to discover the treatment patterns between symptoms and herbs, and incorporate compatibility rules into the model. We introduce a variable $x_i$ to indicate whether herb $h_j$ has compatibility rule with herb $h_j$. If $x_i = 1$, then $h_i$ and $h_j$ are paired herbs; otherwise, they are generated from the distributions associated with their corresponding syndromes. The graphical model of SHT model is shown in Fig. 2.

In Fig. 2, plates represent replications, shaded circles represent observed variables, and unshaded circles represent hidden variables. The outer plate represents clinical cases, while the inner plates represents the repeated choice of topics (syndromes) and words (symptoms and herbs) within a clinical case. $h_i$ and $h_j$ are herbs, $s$ denotes symptoms. $z_i$ denotes the topic assigned for symptoms, $z_h$ denotes the topic assigned for herbs. $D$ is the number of clinical cases, $K$ and $L$ represent the topic number of symptoms and herbs, $M$ and $N$ represent the number of unique herbs and unique symptoms. $\gamma$ is the prior parameter for variable $x_i$. Dirichlet priors $\alpha$ and $\beta$ are set over the clinical case and topic distributions, respectively. SHT generates a collection of clinical cases by the process below:

(i) For each clinical case $c_i$, $i \in [1 \ldots D]$ in the collection, draw $\theta_i$ from a Dirichlet distribution with parameter $\alpha$. Each $\theta_i$ represents the probability of certain topic (syndrome) in clinical case $c_i$.

(ii) For symptoms in each clinical case, draw $\delta_k$ from a Dirichlet distribution with parameter $\beta$. Each $\delta_k$ represents the probability of seeing all symptoms given topic $k, k \in \{1 \ldots K\}$.

(iii) For herbs in each clinical case, draw $\phi_l$ from a Dirichlet distribution with parameter $\beta$. Each $\phi_l$ represents the probability of seeing all herbs given topic $l, l \in \{1 \ldots L\}$.

(iv) For each symptom index $s \in \{1 \ldots N\}$ in clinical case $c_i$: (a) draw a topic $z_s$ from $\theta_i, z_s \in \{1 \ldots K\}$; (b) draw a topic $s$ from $\delta_{z_s}$.

(v) For each herb $h_p, p \in \{1 \ldots M\}$ in clinical case $c_i$: (a) generate $x_p$ from Bernoulli distribution with parameter $\gamma$; (b) draw a topic $z_h$ from $\theta_i, z_h \in \{1 \ldots L\}$; (c) if $x_p = 0$, draw a herb $h_p$ from $\phi_{z_h}$; if $x_p = 1$, draw a herb
If \( x_i \) topic \( k \) estimates the probability of assigning the current symptom to each topic, we set \( n_{ik}^s \) in each clinical case, we use the assignment of symptom or herb to each topic, conditioned on the topic assignment to all other symptoms \( s \neq i \). During Gibbs sampling, we draw the topic assignment \( z_i \) and \( z_k \) according to Eq. (1) and Eq. (2).

\[
p(z_i = k | z_{-i}, s, -s) \propto n_{ik}^{s-} + \beta n_{ik}^{-s} + \alpha n_{ik}^{e-} + N\beta n_{ik}^{e-s} + K\alpha
\]

where \( z_i = k \) means assigning current symptom \( s \) to topic \( k \). \( z_{-i} \) denotes the topic assignments for all symptoms except symptom \( s \). The meanings of \( n_{ik}^{s-}, n_{ik}^{-s}, n_{ik}^{e-} \) and \( n_{ik}^{e-s} \) refer to the corresponding components in Table 1, but not including the current assignment instance \( s \) (represented by the token \(-s\)).

If \( x_i = 0 \):

\[
p(z_h = k | z_{-i}, h_i, -h_i) \propto n_{hk}^{h} + \beta n_{hk}^{h-} + \alpha n_{hk}^{e-h} + M\beta n_{hk}^{e-h-} + L\alpha
\]

If \( x_i = 1 \), \( h_i \) and \( h_j \) are regarded as a whole. We assign the topic for the unit \((h_i, h_j)\). After Gibbs Sampling iterations, we estimate the syndrome-herb distribution \( \varphi \), the syndrome-symptom distribution \( \delta \) and the document-syndrome distribution \( \theta \) as follows:

\[
\delta_i(x) = \frac{n_{ik}^s + \beta}{n_k^s + N\beta}
\]

\[
\theta(k) = \frac{n_i^k + \alpha}{n_i^k + M\beta}
\]

3. Results and Discussion

3.1 Setup

We collect 3,765 clinical cases from Professional Knowledge Service System for Chinese Herbal Medicine\(^1\). The symptoms and the herbs should be extracted by text matching according to Traditional Chinese Medical Subject Headings (TCM Mesh)\(^2\) and Chinese pharmacopoeia (2,000 edition).

We designed three experiments to validate our method: LDA-based method, LinkLDA-based method\(^9\) and SHT-based method. LinkLDA can simultaneously model the content of documents and citations in previous work\(^9\). We employed it to extract the latent topic structures which involve the symptoms and their corresponding herbs. To evaluate the performance of our topic model, we used two metrics: the perplexity and the accuracy of top 5 “words” discovered for latent topics. The former can be thought of as the effective number of equally likely words (symptoms or herbs) according to the model. It is a common way to evaluate the effectiveness of topic models on topic modeling. We computed the perplexity of the test sets with parameters learned from the corresponding training sets. Let \( C \) be the set of clinical cases, the definition of perplexity is defined as follows:

\[
\text{perplexity}(C_{test}) = \exp \left( \frac{\sum_{w\in C_{test}} \ln p(w|C_{train})}{\sum_{w\in C_{test}} n_i} \right)
\]

where \( C_{test} \) is the test data set, \( w_i \) is a vector of “words” in clinical case \( c_i \) of the test set, and \( C_{train} \) is the training set. \( N_i \) denotes the number of terms and herbs in clinical case \( c_i \). \( p(w|C_{train}) \) denotes the probability of the words \( w_i \) in a test clinical case \( c_i \) under the parameters trained by training set. Note that lower numbers denote better performance. The latter evaluation can be computed as follows:

\[
\text{Accuracy} = \frac{\text{The correct number of returned “words”}}{\text{The total number of returned “words”}}
\]

The correct number of returned “words” (symptoms and herbs) is determined by expert’s manual judgement. For each syndrome (topic), if a symptom can reflect the syndrome and a herb has therapeutic effects on the syndrome, then the “word” is correct. We randomly selected 20% clinical cases as the test set. In topic modeling process, we set the hyper-parameters for both two models as follows: \( \alpha = 50/K, \beta = 0.1 \), and the iteration number \( l = 100 \).

3.2 Overall Performance and Discussions

In ancient TCM books, 917 paired herbs have been discovered by TCM experts\(^2\). The data set of paired herbs was incorporated into the SHT topic modeling process. We conducted treatment pattern mining through LDA, LinkLDA and SHT model, and calculated perplexity on different number of topics, which vary from

\(^1\) http://zcy.ckcest.cn/MedicalRecord/browse

\(^2\) http://pan.baidu.com/s/1JfPwU6
Table 2 shows the probability distributions of 5 discovered topics in SHT model.

| Topic | Diagnosis | Symptoms | Herbs |
|-------|-----------|----------|-------|
| 1     | Cough     | Cough, excessive phlegm, deep and adynamic pulse, thick yellow sputum, upward adverseness of gas to the chest | Alpiniae Oxyphyllae Fructus, Caulis Perillae, Platycodonis Radix, Cortex Morii, Morindae Officinalis Radix, Armeniaca Semen Amarum, Pinelliae Rhizoma |
| 2     | Thirsty   | Hot flush, diuresis, yello tongue, weak, polydipsia and polyphagia | Anemarrhenae Rhizoma, Ophiopogonis Radix, Puerariae Lobatae Radix, Scrophulariae Radix, Fructus Schisandrae Chinensis, Atractylodis Macrocephalae Rhizoma, Pinelliae Rhizoma |
| 3     | Asthenic | Abdominal Distension, yellow eyes, sallow complexion, heavy arms and legs, giddy and dazzled | Gentianae Macrophyllae Radix, Forsythiae Fructus, Atractylodis Rhizoma, Ephedra herba, Radix Angelicae Sinensis |
| 4     | Heatstroke| Headache, aversion to wind, pharyngalgia, puerperal fever, yellow tongue fur | Radix Puerariae Lobatae, Saposhnikoviae Radix, Forsythiae Fructus, Glycyrrhizae Radix, Bupleuri Radix |

Figure 3 shows the perplexity scores on different number of topics for LDA, LinkLDA and SHT. We can see that SHT and LinkLDA outperform LDA regardless of the topic number, which demonstrates the effectiveness of separate modeling for symptoms and herbs. Actually, the separate modelling for symptoms and herbs is a more realistic approach for TCM practitioners to conduct automatic diagnosis research and new prescription discovery. However, LDA cannot discover combinations of effective interacting herbs.
and herbs can improve the topic structure. If symptoms and herbs are regarded as a whole, the probability distribution for symptoms and herbs will be mixed together. SHT performs better than LinkLDA when $K \leq 60$, which means that concerning combinational rules can improve the modeling performance. However, the performance of SHT is close to LinkLDA when $K > 60$, this is possibly because the larger number of topics would decrease the number of paired herbs. Figure 4 shows the accuracy of discovered “words.” The accuracy has to be calculated manually by expert’s instruction, thus we set the topic number between 10 and 40 to alleviate the heavy workload for our task. Our results show that the SHT is more efficient at extracting symptom-herb relationship from the clinical case compared with the basic LDA model (increasing accuracy by 18%, 17%, 19% and 23% for different number of topics) and the LinkLDA (increasing precision by 5%, 12%, 1%, 2% for different number of topics).

However, some of our results can be improved upon, and our approach can be expanded upon in the future. Firstly, some other types of important TCM entities, such as prescriptions and diseases, are not incorporated into our model. If we can bring such entities into our unified model in the future, then more types of relations can be extracted. Secondly, most of the symptom names are manually extracted because there is not a standard or unified terminology glossary for TCM symptoms, so entity recognition techniques are needed to detect symptom entities in clinical cases.

4. Conclusions

This paper has presented a method of clinical records mining based on probabilistic topic model. We propose a novel topic model named SHT to discover the treatment patterns between symptoms and herbs. The combinational rules are incorporated into the SHT modeling process. Each discovered topic involves a list of symptoms and its corresponding list of herbs. The performance shows that our approach is superior to other topic models in extracting symptom-herb relations from TCM clinical cases. The results can provide valuable information for TCM automatic diagnosis or poly-pharmacology research.

The dosage of herbs in a prescription plays a key role in clinical treatment. The efficiency of a composition of herbs would change when we adjust the dosage of herbs. In the future, we plan to incorporate the dosage information into the topic modeling process. Besides, we intend to use the mining results to construct a calculation model for automatic diagnosis. Specifically, when the doctor provides the symptoms of a patient, our model may automatically return the corresponding combination of herbs to cure the disease.

Acknowledgments

This study was funded by Zhejiang Provincial Natural Science Foundation of China under Grant No. LQ14F020008 and No. LY17E070004, National Natural Science Foundation of China under Grant No. 61602402.

References

[1] Yang, H., Chen, J., Tang, S., Li, Z., Zhen, Y., Huang, L. and Yi, J.: New drug R&D of traditional Chinese medicine: Role of data mining approaches, Journal of Biological Systems, Vol.17, No.03, pp.329–347 (2009).
[2] Liu, X.L., Hong, W.X., et al.: Using Formal Concept Analysis to Visualize Relationships of Syndromes in Traditional Chinese Medicine, Medical Biometrics, Vol.6165, pp.315–324 (2010).
[3] Yang, T., Wu, C., Xu, Z. and Ding, Y.: The syndrome differentiation model and program of traditional Chinese medicine based on the fuzzy recognition, Proc. BIBM, pp.285–287, Shanghai, China, IEEE (2013).
[4] Wang, L., Zhang, Y. and Xu, X.: A Novel Group Detection Method for Finding Related Chinese Herbs, Journal of Information Science and Engineering, Vol.31, No.4, pp.1387–1411 (2015).
[5] Qiao, S.J. and Tang, C.J.: Mining the compatibility rule of multidimensional medicines based on dependence model sets, Journal of Sichuan University (Engineering and Science Edition), Vol.39, No.4, pp.134–138 (2007).
[6] Wu, Z., Zhou, X., Liu, B., et al.: Text mining for finding functional community of related genes using TCM knowledge, Proc. PKDD’04, Seattle, pp.454–470, WA, USA, ACM (2004).
[7] Chen, J., Poon, J., Poon, S.K., Xu, L. and Daniel M.Y.: Mining Symptom-Herb Patterns from Patient Records Using Tripartite Graph, Evidence-Based Complementary and Alternative Medicine, Vol.2015, 435085 (2015).
[8] Zhang, X., Zhou, X., Huang, H., Chen, S. and Liu, B.: A hierarchical symptom-herb topic model for analyzing traditional Chinese medicine clinical diabetic data, Proc. BMEI, pp.2246–2249, Yantai, China, IEEE (2010).
[9] Lin, Y. and Mizil, A.N.: Topic-Link LDA: Joint Models of Topic and Author Community, Proc. ICML, pp.665–672, Montreal, Canada, IEEE (2009).
[10] Wan, H., Moens, M., Luyten, W., Zhou, X., Mei, Q., Liu, L. and Tang, J.: Extracting relations from traditional Chinese medicine literature via heterogeneous entity networks, Journal of the American Medical Informatics Association, Vol.23, No.2, pp.356–365 (2016).
[11] Yao, L., Zhang, Y., Wei, B., Wang, 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, Vol.58, pp.260–267 (2015).
[12] Wu, L.: Chinese Traditional Medicine and Material Medical Subject Headings, Chinese Medical Ancient Books Publishing, Beijing (1996).
[13] Zhao, Y., He, L., Xie, Q., Li, G., Liu, B. and Wang, J.: A Novel Classification Method for Syndrome Differentiation of Patients with AIDS, Evidence-Based Complementary and Alternative Medicine, Vol.2015, 936290 (2015).
[14] Zhao, C., Li, G., Wang, C. and Niu, J.: Advances in Patient Classification for Traditional Chinese Medicine: A Machine Learning Perspective, Evidence-Based Complementary and Alternative Medicine, Vol.2015, 936290 (2015).
[15] Steyvers, M. and Griffiths, T.: Probabilistic topic models, Latent Semantic Analysis: A Road to Meaning, Landauer, T., et al. (Eds.), Lawrence Erlbaum (2006).
Lidong Wang was born on December 4, 1982. She received her M.S. degree in Computer Science from Ningbo University and her Ph.D. degree from the College of Computer Science and Technology, Zhejiang University. She is currently an Associate Professor at Hangzhou Normal University. Her current research interests include image processing, machine learning, and text mining.

Keyong Hu is currently a teacher of Electronic Information Engineering in Qianjiang College of Hangzhou Normal University. He has received Ph.D. degree in 2016 from Zhejiang University of Technology, Hangzhou China, in Mechatronic Engineering. His research interests include artificial intelligence, new energy technology.

Xiaodong Xu is currently a Professor at Zhejiang Chinese Medical University. He has (co)authored over 30 publications on the drug exploitation in Traditional Chinese Medicine. His research interests include prescriptions, drug exploitation, and combinational rule analysis.

(Communicated by Tatsuya Akutsu)