Research Article

SDTM: A Novel Topic Model Framework for Syndrome Differentiation in Traditional Chinese Medicine

Jialin Ma,1 Xiaoqiang Gong,2 Zhaojun Wang,3 and Qian Xie4

1Jiangsu Internet of Things and Mobile Internet Technology Engineering Laboratory, Huaiyin Institute of Technology, Huaian 223003, China
2AVIC Xi’an Aircraft Industry Group Company Ltd., Xi’an 710089, China
3Huaiyin Wu Jutong Institute of Traditional Chinese Medicine, Huaian 223000, China
4Jiangsu Eazytec Co. Ltd., Wuxi, China

Correspondence should be addressed to Jialin Ma; majl@hyit.edu.cn and Xiaoqiang Gong; 1769728791@qq.com

Received 9 September 2021; Accepted 20 December 2021; Published 4 January 2022

Academic Editor: Chinmay Chakraborty

Copyright © 2022 Jialin Ma et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Syndrome differentiation is the most basic diagnostic method in traditional Chinese medicine (TCM). The process of syndrome differentiation is difficult and challenging due to its complexity, diversity, and vagueness. Recently, artificial intelligence methods have been introduced to discover the regularities of syndrome differentiation from TCM medical records, but the existing DM algorithms failed to consider how a syndrome is generated according to TCM theories. In this paper, we propose a novel topic model framework named syndrome differentiation topic model (SDTM) to dynamically characterize the process of syndrome differentiation. The SDTM framework utilizes latent Dirichlet allocation (LDA) to discover the latent semantic relationship between symptoms and syndromes in mass of Chinese medical records. We also use similarity measurement method to make the uninterpretable topics correspond with the labeled syndromes. Finally, Bayesian method is used in the final differentiated syndromes. Experimental results show the superiority of SDTM over existing topic models for the task of syndrome differentiation.

1. Introduction

As an important complementary medical system to modern biomedicine, traditional Chinese medicine (TCM) has played an indispensable role in healthcare of Chinese people for several thousand years [1, 2]. In recent years, the TCM has become more and more popular all over the world [3]. Doctors usually adopt four diagnostic ways to obtain symptoms, that is, observation, listening, interrogation, and pulse-taking in TCM [4]. A syndrome can be summarized via a set of symptoms, which are intrinsically related to each other. This process is the key to differentiating syndromes. An example of syndrome is given in Figure 1, which is selected from [4]. It includes syndrome name, symptoms, pathogenesis, treatment, representative prescription, and common medicines [5–7].

One of the significant characteristics of TCM is to treat diseases based on syndrome differentiation. This is a process of comprehensive judgment based on analysis, induction, and reasoning via four-way information diagnosis [8]. This is also the key link for doctors to select proper prescriptions or therapies. Syndrome differentiation is a process through which doctors make a diagnosis based on subjective knowledge and experience in accord with the objective reality of a patient. Because of the differences in individuals and the limited knowledge or experience of doctors, one patient may be diagnosed with different syndromes by different doctors [9].

In order to accurately master the complex structure of syndromes and establish a diagnostic standard for TCM, in time, it is of great significance to analyze the principles of syndrome differentiation. This is beneficial for the
Syndrome Name: Syndrome of sinking of qi due to spleen deficiency  
Symptom: urinary turbidity has recurrent attacks, no cure for a long time, shaped like white pulp, sagging distention in the smaller abdomen, deity and weakness, lusterless complexion, fatigue or exacerbation after exertion, pale tongue with whitish coating, pulse asthenia soft.

Pathogenesis of Syndrome: spleen deficiency depression, spermatozoa leakage.

Treatment: strengthening the spleen and replenishing qi, ascending and clearing and fixing.

Representative Prescription: buzhong yiqi decoct add and subtract. It is used to buzhong yiqi decocti, ascending clear and descending turbid, used for sinking of qi of middle-jiao, spermatozoa leakage.

Common Medicine: codonopsis pilosula, astragalus membranaceus, bighead atractyloides rhizome, rhizoma dioscoreae, semen amomi Amari, fructus rosae laevigatae, semen nelumbinis, semen euryales, rhizome, radix bupleuri.

In this paper, we propose a novel topic model framework to dynamically characterize the process of syndrome differentiation of TCM. The overall framework of the SDTM is shown in Figure 3. First, we propose a novel LDA-based model approach to discover the latent semantic relationship between symptoms and syndromes in Chinese medical records. Then, the corresponding syndromes are labeled for these topics based on similarity measurement in order to improve interpretability of topics. Finally, we utilize Bayesian method to implement syndrome differentiation. Our method contributes to a better understanding of TCM diagnostic principles and provides an effective model for computer automatic diagnosis.

The rest of this paper is organized as follows: Section 2 reviews some related works. Section 3 shows the specific differentiation process of syndromes. The experimental results are analyzed in Section 4. Finally, conclusion and future work are given in Section 5.

2. Related Works

2.1. TCM Knowledge Discovery. Knowledge discovery and data mining have become popular topics in healthcare and biomedicine [27]. The research of TCM knowledge discovery is summarized by Feng et al. [21], Lukman et al. [22], Wu et al. [23], and Liu et al. [27]. Many methods have been proposed to discover some regularities in TCM diagnosis and treatments. Zhang et al. [13] proposed a novel method based on author-topic model, called the symptom-herb-diagnosis topic model (SHDTM), to automatically extract the relationships between symptoms, herb groups, and diagnoses from TCM clinical data. Erosheva et al. [14] used link latent Dirichlet allocation (LinkLDA) to extract the latent topics with both symptoms and their corresponding herbs in clinical cases. Yao et al. [1] applied LDA and TCM domain knowledge to mine treatment patterns in TCM clinical cases.

2.2. Topic Model. Recently, topic model, as a popular text analysis method, can detect latent topics in large-scale documents [24]. It is known that two classical topic models
have been extensively applied to document analysis. They are probabilistic latent semantic analysis (PLSA) and latent Dirichlet allocation (LDA) [25]. In PLSA, a document is regarded as a mixture of topics, where a topic is determined by the probability distribution over words. In order to solve the limitation of PLSA, LDA adds Dirichlet priors in the distributions; it is a complete generative model and achieves great successes in text mining. Moreover, LDA can also be utilized in the tasks of health and biomedicine mining [13, 27–30]. For instance, Yao et al. [15] discovered some important treatment patterns in TCM clinical cases by exploiting the supervised topic model and domain knowledge. Chen et al. [20] demonstrated that the configuration of functional groups in metagenome samples can be inferred by probabilistic topic model. Huang et al. [29] mined the latent treatment patterns for clinical pathways through topic model. In addition, some improved topic models are also proposed for short text analysis, such as author-topic model (ATM) [26] and block-LDA [30].

However, a standard LDA still cannot be directly used for TCM mining, because it is an unsupervised topic model, which is unable to express the relationships between syndromes and symptoms [31–33]. Furthermore, the above-mentioned research failed to consider the syndrome differentiation principles [34–38]. Therefore, we propose a novel topic model framework called syndrome differentiation topic model to dynamically characterize the process of TCM syndrome differentiation.

3. Method
In this section, we present the framework named SDTM to characterize how a syndrome is generated according to TCM theory. It consists of three steps: topic modeling of Chinese medical records, syndrome labeling, and syndrome differentiation.

3.1. Topic Modeling of Chinese Medical Records
In the process of diagnosis and treatment, the TCM doctors usually obtain symptoms through four diagnostic ways, i.e., observation, listening, interrogation, and pulse-taking, and
3.1.1. Model Generative Process. The graphical representation of topic modeling of Chinese medical records is given in Figure 4. The meaning of notations is illustrated in Table 1.

When modeling the Chinese medical records in the frame SDTM, let $M$ be the number of medical records, where each medical record $m$ owns $N_{sm}$ symptoms, $s_{mn}$ is the $n$th symptom in medical record $m$, and $z_{mn}$ ($n = 1, 2, \ldots, N$) is the latent syndrome distribution for $s_{mn}$. For instance, the medical record in Figure 2 has $N_{s} = 18$ symptoms, and the latent syndrome distribution for the symptom “diuresis” should be “two deficiency syndrome of liver and kidney” or “syndrome of dampness-heat blocking collaterals.” Let $K$ be the number of topics, a topic $k \in \{1, 2, \ldots, K\}$ represent a syndrome, and $\varphi_k$ be the $N$-dimensional syndrome-symptom multinomial for syndrome $k$, where $N$ is the number of all unique symptoms in $M$ medical records. $\theta_m$ is the $K$-dimensional medical records-syndrome multinomial for medical record $m$. $\alpha$ and $\beta$ are the hyperparameters of the Dirichlet priors on $\theta_m$ and $\varphi_k$, respectively.

The graphical representation of Chinese medical records is given as follows:

(1) For syndrome $k$ in 1, 2, , , $K$, draw $\varphi_k \sim \text{Dir}(\beta)$.
(2) For medical record $m$, draw $\theta_m \sim \text{Dir}(\alpha)$.
(3) For each of the $N_{sm}$ symptoms in medical record $m$:
   (a) Draw a symptom $z_{mn} \sim \text{Mult}(\theta_m)$.
   (b) Draw a symptom $s_{mn} \sim \text{Mult}(\varphi_k)$.

Here, Dir is a convenient distribution on the simplex. It is in the exponential family and has finite dimensional sufficient statistics. It is conjugate to the multinomial distribution [9]. Mult represents the multinomial distribution.

3.1.2. Model Inference and Learning. Gibbs sampling is an effectively and widely used Markov chain Monte Carlo algorithm for latent variable inference [24, 25]. We use Gibbs sampling to extract latent syndrome distributions $z_{mn}$; it is defined as follows:

\[
p(z_{mn} = k | s_{mn}, z_{-mn}, z, \alpha, \beta) \propto \frac{n_{km}^m + \alpha}{\sum_{k=1}^{K} n_{km}^m} \times \frac{n_{km} + \beta}{\sum_{i=1}^{N} n_{ik}^m} = N \beta
\]

where $k$ represents a syndrome, $s_{mn}$ represents all symptoms except $s_{mn}$, $z_{mn}$ represent the syndrome distributions for all symptoms except $s_{mn}$, $z$ represent the syndrome distributions for all symptoms, $n_{km}^m$ is the number of times syndrome $k$ occurs in medical record $m$, and $n_{km}^m$ is the number of times $s_{mn}$ is assigned to syndrome $k$.

According to Gibbs sampling, $\theta_m$ and $\varphi_k$ can be calculated as follows:

\[
\theta_m(k) = \frac{n_{km}^m + \alpha}{\sum_{k=1}^{K} n_{km}^m},
\]

\[
\varphi_k(s_{mn}) = \frac{n_{km}^m + \beta}{\sum_{i=1}^{N} n_{ik}^m + \beta}
\]

3.2. Syndrome Labeling. Although topic modeling of Chinese medical records is successful in discovering hidden topics from medical records, each of these topics lacks an identifiable label, which results in low interpretability. Therefore, to improve the interpretability of topics, we label a syndrome on each topic by mapping symptoms in a topic to syndromes in TCM domain. First, we select data from [4] to build a standard syndrome database with $d$ syndromes. Then syndrome $y_j$ ($j \in \{1, 2, \ldots, d\}$) in the syndrome database is assigned to topic $k$ ($k \in \{1, 2, \ldots, K\}$) based on the similarity between $k$ and $y_j$, which is calculated using Jaccard similarity coefficient as follows [25]:

\[
y = \arg \max \text{sim}(k, y_j) = \arg \max \frac{|K \cap y_j|}{|K \cup y_j|}
\]

where $d$ is the number of syndromes in standard syndrome database and $y_j$ represents the $j$th syndrome in the standard syndrome database.

3.3. Syndrome Differentiation. After these syndromes are assigned, probability $p(k|H(s))$ of syndrome (topic) $k$ for medical record $H(s)$ can be computed using the Bayesian formula as follows:

\[
p(k|H(s)) \propto \frac{\sum_{s \in H(s)} p(s|k)p(k)}{|H(s)|}
\]

where a new medical record is represented by a set of symptoms $H(s)$, $p(k|H(s))$ is the probability of syndrome $k$ given medical record $H(s)$, $p(s|k)$ is the probability of symptom $s$ given syndrome $k$ which is equal to $\varphi_k(s)$, $p(k)$ is the prior of syndrome $k$ which can be regarded as a constant, and $|H(s)|$ is the number of symptoms in the new medical record $H(s)$.

To differentiate the syndromes for a given medical record, we exploit the symptom vector to represent the medical record:

\[
H(s) = (s_1, \ldots, s_l, \ldots, s_n),
\]
where symptom $s_i$ is a binary indicator; if a medical record contains $s_i$, it is equal to 1; otherwise, it equals 0.

We take the posterior vector as the feature vector of medical record $H(s)$:

$$\theta_{H(s)} = \{p(1|H(s)), \ldots, p(i|H(s)), \ldots, p(K|H(s))\},$$

where $p(i|H(s))$ represents the probability of syndrome $i$ which is calculated via (4).

We use (6) to determine syndromes of medical record $H(s)$:

$$\text{Syndrome}_{H(s)} = \{k|p(k|H(s)) > T, \quad k \in 1, 2, \ldots, K, H(s) \in R^n\},$$

where $T$ is the syndrome differentiation threshold and $n$ is the number of symptoms in $H(s)$.

4. Experimental Results

In the section, we evaluate our framework, SDTM, on three experimental tasks for Chinese medical records. In particular, we want to determine the following:

(i) Can our SDTM achieve the best generalization performance compared to other topic models?

(ii) Can our SDTM differentiate syndromes for a set of symptoms?

(iii) Can our model reflect the patterns of TCM syndrome differentiation?

All experiments are tested in MATLAB 2015a and implemented on a computer with Intel Core i3-7100, 3.90 GHz CPU, 8 GB RAM, and Windows 10 64-bit operating system. Each experiment is run 10 times.
4.1. Dataset. Chronic kidney disease (CKD) is a common condition in clinical practice. The basic clinical manifestations of the disease include proteinuria, hematuria, hypertension, and edema. The disease has insidious cause, long course, and slow change of state, so its clinical treatment is difficult. Although modern medicine has adopted such means as controlling hypertension, reducing proteinuria and lipid, the prognosis is not good. Traditional Chinese medicine has significant advantages in the treatment of the disease, such as reducing adverse drug reactions and inhibiting relapse of the disease. We collected 1959 medical records on CKD from Beijing Dongzhimen Hospital, which include 948 (48.4%) females and 1011 (51.6%) males. The dataset mainly contains 4 syndromes, i.e., “deficiency of Qi and blood,” “retention of dampness and blood stasis,” “chronic nephritis,” “hypertension,” “cerebral embolism,” “hyperuricemia,” “hyperlipidemia,” “membranous nephropathy,” and “IgA nephropathy.” For example, a medical record is shown in Figure 2, where the texts in red are the symptoms in test medical records, $s_{\text{test}}$ are symptoms in test medical records, $s_{\text{test}}$ are symptoms in test medical records, $\overline{s}_{p}$ are symptoms in medical record $p$ of the test set, $\overline{u}_p$ are syndromes in medical record $p$ of the test set, $T_{\text{test}}$ is the number of medical records in the test set, $N_{u_p}$ is the number of syndromes in test medical record $p$, $s_{\text{test}}$ is the symptoms in test medical records, and $s_{\text{test}}$ is the symptoms in test medical records.

4.2. Baselines. We compare our method with the following baselines:

(1) Author-topic model (ATM) [26]: ATM is an extended LDA model, which extracts the topic distribution by utilizing the author information contained in documents. Here, we regard syndromes as authors and symptoms as words.

(2) LinkLDA [28]: LinkLDA is also a probabilistic generative model, which considers both the words in documents and the reference document information of these words. Here, we regard symptoms as words and references.

(3) Block-LDA [30]: Block-LDA is an extended LinkLDA model which links between certain types of entities. Here, we regard symptoms as words and regard symptom-pair set extracted from all training medical records as the external links.

(4) Symptom-syndrome topic model (SSTM): SSTM proposed in previous work [11] is an LDA-based topic model, which regards syndromes as topics and symptoms as words.

4.3. Evaluation Metrics. Here, we use the differentiated perplexity to evaluate the generalization performance of topic models. A lower perplexity means generalization performance of the topic model is better. The differentiated perplexity of a set of test symptoms is defined as follows [24]:

$$\text{perplexity} = \exp \left( \frac{-\sum_{p=1}^{P_{\text{test}}} \log p(u_p | s_{\text{test}})}{\sum_{p=1}^{P_{\text{test}}} N_{u_p}} \right),$$

$$p(u_p | s_{\text{test}}) = \prod_{u_{pl} \in u_p} p(u_{pl} | s_{\text{test}}),$$

$$= \prod_{u_{pl} \in u_p} \frac{1}{N_{u_p}} \sum_{s_{pl} \in s_{\text{test}}} p(u_{pl} | s_{pl}).$$

4.4. Parameter Settings. For all the models in comparison, we set hyperparameters $\alpha = 50/K$, $\beta = 0.01$, and the number of standard syndromes $d = 137$. We use 1000 Gibbs sampling iterations to train all topic models.

For all tests, we use Jaccard similarity coefficient to measure the similarity between syndromes $X$ and $X'$, which is defined as follows:

$$\text{Sim}(X, X') = \frac{|X \cap X'|}{|X \cup X'|},$$

where $X$ represents a syndrome in a test medical record and $X'$ represents a predicted syndrome in Syndrome $H(s)$. For similarity threshold $C$, if $\text{Sim}(X, X') > C$, then $X'$ is a true syndrome. In the stage of syndrome differentiation, we need to determine threshold $T$ so that we can differentiate syndromes for each medical record. However, there is no theoretical guidance for automatically selecting an optimal threshold for syndrome differentiation. Therefore, when $K$
and C are both fixed, we use different thresholds \( T \) to compare the perplexity and accuracy.

As shown in Table 3, the value of \( T \) has a significant influence on the syndrome differentiation results. When \( T = 1e^{-7} \), all methods achieve the best syndrome differentiation results, and SDTM outperforms ATM, LinkLDA, Block-LDA, and SSTM in terms of perplexity and accuracy, so we select \( T = 1e^{-7} \) as an optimal threshold.

In the stage of syndrome evaluation stage, we need to determine similarity threshold \( C \) so that we can select true syndromes from the syndromes differentiated by SDTM. Therefore, when \( K \) is fixed and \( T = 1e^{-7} \), we use different thresholds \( C \) to compare the accuracy of all models. As shown in Figure 5, for different models, the accuracy of syndrome differentiation varies with the value of \( C \). It is clearly seen that when \( C = 0.6 \), all models obtain the highest number of true syndromes, and SDTM substantially outperforms the other four models in terms of accuracy, so we take \( C = 0.6 \) as an optimal similarity threshold for selecting true syndromes.

### 4.5. Experimental Results

#### 4.5.1. Generalization Performance

Figure 6 shows the variation of perplexity with the increase of topics. It is seen that the average perplexity of SDTM is less than those of the other four models. This demonstrates that our model is more efficient in the task of syndrome differentiation. When \( K \) is equal to 40, SDTM achieves the minimum perplexity, which means that the best generalization performance is achieved.

#### 4.5.2. Syndrome Differentiation

Figure 7 shows the variation of accuracy with increasing of topics. The average accuracy of SDTM is higher than that of the other four models in Figure 7. When \( K \) is equal to 40, the SDTM achieves the highest accuracy.

In summary, from Figures 6 and 7, we can see that when \( K \) is equal to 40, the SDTM has the best generalization performance.
performance and syndrome differentiated power, so we take $K = 40$ as the optimal number of topics.

4.5.3. Discovery of Syndrome Pattern. The top five topics generated by several baseline methods are shown in Tables 4–8, respectively. The top ten symptoms in each “syndrome” topic are also shown, where italicized symptoms are not related to the syndrome. Compared with the other four methods, our SDTM can discover the best differentiated results of syndromes, and most of symptoms in each “syndrome” topic can be validated effectively by the true syndromes in [4]. From Tables 4–8, we draw the following results for the discovered syndrome patterns.

The first “syndrome” topic is “two deficiency syndrome of liver and kidney.” The results are shown in Tables 1–8: (1) ATM cannot discover a good topic; only the symptoms “inhibited defecation,” “bowel 1 per day,” and “weak” are related. (2) LinkLDA discovers one topic with five related symptoms. (3) Block-LDA and SSTM discover seven related symptoms. (4) SDTM discovers a good topic with nine related symptoms.

The second “syndrome” topic is “syndrome of dampness-heat blocking collaterals.” We find the following results: (1) ATM cannot provide a good topic again; only
Table 4: Topics learned by ATM with \( K = 40 \).

| ATM                                                                 |
|---------------------------------------------------------------------|
| Two deficiency syndrome of liver and kidney                         |
| Syndrome of dampness-heat blocking collaterals                      |
| Syndrome of dampness-heat diffusing downward                       |
| Syndrome of yang deficiency of spleen and kidney                   |
| Syndrome of yin deficiency and dampness-heat                       |
| Inhibited defecation                                               |
| Leg swelling                                                       |
| Hypermenorrhea                                                     |
| Stomachache                                                        |
| Phlegm yellow                                                      |
| Bowel 1 per day                                                    |
| No hard stool                                                      |
| Weak                                                               |
| Dark red tongue                                                    |
| Bloody stool                                                       |
| Palpitation                                                       |
| Soreness of waist                                                 |
| Soreness of waist                                                 |
| Sallow complexion                                                 |
| Sunken pulse                                                       |
| Debility of the legs                                               |
| Irritability                                                       |
| Bowel 1 per day                                                    |
| Abnormal diet                                                      |
| Bowel 1 per day                                                    |
| Bowel 1 per day                                                    |
| Bowel 1 per day                                                    |
| Normal diet                                                        |
| No hard stool                                                      |
| Hard stool                                                         |
| Red tongue                                                         |
| Yellow fur                                                        |
| Depression                                                        |
| Less urine volume                                                 |
| Hand edema                                                         |
| No hard stool                                                      |
| Leg swelling                                                       |
| Loose stool after bowel hard                                      |
| Dark red tongue                                                    |
| Bloody stool                                                       |
| Normal sleep                                                       |
| Dark red tongue                                                    |
| Lip color: purple                                                 |
| Normal sleep                                                       |
| Yellow fur                                                        |
| Italics represent the values correctly predicted by the model.     |

Table 5: Topics learned by LinkLDA with \( K = 40 \).

| LinkLDA                                                            |
|--------------------------------------------------------------------|
| Two deficiency syndrome of liver and kidney                        |
| Syndrome of dampness-heat blocking collaterals                     |
| Syndrome of dampness-heat diffusing downward                       |
| Syndrome of yang deficiency of spleen and kidney                  |
| Syndrome of yin deficiency and dampness-heat                       |
| Less urine volume                                                 |
| Hand edema                                                         |
| No hard stool                                                      |
| Leg swelling                                                       |
| Loose stool after bowel hard                                      |
| Bloody stool                                                       |
| Dizziness                                                          |
| No hard stool                                                      |
| Leg swelling                                                       |
| Loose stool after bowel hard                                      |
| Bloody stool                                                       |
| Normal sleep                                                       |
| Dark red tongue                                                    |
| Heartburn                                                          |
| Chest, epigastric fullness, and distress                          |
| Profuse spittle                                                    |
| Loose stool                                                        |
| Normal sleep                                                       |
| Dark red tongue                                                    |
| Heartburn                                                          |
| Borborygmus                                                       |
| Dark red tongue                                                    |
| Normal urination                                                  |
| Profuse spittle                                                    |
| Loose stool                                                        |
| Normal sleep                                                       |
| Dark red tongue                                                    |
| Lip color: purple                                                 |
| Stringy pulse                                                     |
| Italics represent the values correctly predicted by the model.     |

Table 6: Topics learned by block-LDA with \( K = 40 \).

| Block-LDA                                                        |
|------------------------------------------------------------------|
| Two deficiency syndrome of liver and kidney                      |
| Syndrome of dampness-heat blocking collaterals                   |
| Syndrome of dampness-heat diffusing downward                     |
| Syndrome of yang deficiency of spleen and kidney                 |
| Syndrome of yin deficiency and dampness-heat                     |
| Soreness of waist                                                |
| Dark red tongue                                                  |
| Weak                                                             |
| Slippery pulse                                                   |
| Bowel 1 per day                                                  |
| No hard stool                                                    |
| Lip color: purple                                                |
| Normal sleep                                                     |
| Yellow fur                                                       |
| Hard stool                                                       |
| Red tongue                                                       |
| Rapid pulse                                                      |
| Bowel 3 per day                                                  |
| Normal urination                                                 |
| No abdominal distention                                          |
| Hard stool                                                       |
| Yellow fur                                                       |
| Spiritlessness                                                   |
| Normal diet                                                      |
| Normal urination                                                 |
| No abdominal distention                                          |
| No abdominal distention                                          |
| No abdominal distention                                          |
| Italics represent the values correctly predicted by the model.    |
“palpitation,” “abnormal diet,” and “dark red tongue” are related symptoms. (2) LinkLDA discovers a little better topic with four related symptoms. (3) Block-LDA and SSTM discover six related symptoms. (4) SDTM discovers eight related symptoms.

The third “syndrome” topic is “syndrome of dampness-heat diffusing downward.” We find the following results: (1) ATM discovers a little better topic with five related symptoms. (2) LinkLDA cannot discover a meaningful topic including only three related symptoms, namely, “thin fur,” “soreness of waist,” and “hard stool.” (3) Block-LDA and SSTM discover six related symptoms. (4) SDTM discovers nine related symptoms.

From the abovementioned five topics, we find that SDTM can discover “syndrome” the most related topics.

5. Conclusion and Future Work

We present a novel framework, SDTM, in this paper which can effectively analyze complex and changeable syndrome differentiation patterns from TCM historical clinic records. The framework SDTM conforms to the relevant theories of TCM. The experimental results on 1959 medical records show that SDTM can discover meaningful syndrome

| Table 7: Topics learned by SSTM with $K = 40$. |
|---------------------------------------------|
| **STM**                                    |
| Two deficiency syndrome of liver and kidney | Symptom of dampness-heat blocking collaterals | Symptom of dampness-heat diffusing downward | Symptom of yang deficiency of spleen and kidney | Symptom of yin deficiency and dampness-heat |
| Inhibited defecation                        | Knee pain                                    | Dumb                                        | Fissured tongue                           | Thirst without desire to drink |
| Hand edema                                  | Depression                                   | Dark red tongue                             | Soreness of waist                         | Brown macules on the skin   |
| Bulgy tongue                                | Chest, epigastric fullness, and distress     | Soreness of waist                           | Loose stool                               | Epistaxis                   |
| Difficulty in micturition                   | Dark red tongue                              | Emaciation                                  | No abdominal distention                   | Stringy pulse               |
| Stomachache                                 | Spontaneous perspiration                     | Borborygmus                                | Dizziness                                 | Dark red tongue             |
| Profuse spittle                             | Aversion to cold                             | Bloody stool                                | Lassitude                                 | Hematocchia                 |
| Aversion to cold                            | Arthralgia                                   | Nausea                                     |                                          | Hematuria                   |
| Palpitation                                 | Palpitation                                  | Greenish complexion                        | Sallow complexion                         | Dumb                        |
| Chest tightness                             | Indigestion                                  | Lochiotasis                                 | Lip color: purple                         | Normal sleep                |
| No abdominal distention                     | Hand edema                                   | Diuresis                                   | Turbid urine                              | Bowel 1 per day             |

| Table 8: Topics learned by SDTM with $K = 40$. |
|---------------------------------------------|
| **SDTM**                                    |
| Two deficiency syndrome of liver and kidney | Symptom of dampness-heat blocking collaterals | Symptom of dampness-heat diffusing downward | Symptom of yang deficiency of spleen and kidney | Symptom of yin deficiency and dampness-heat |
| Inhibited defecation                        | Lumbar flaccidity                            | Soreness of waist                           | Rapid pulse                               | Blurred vision              |
| Bulgy tongue                                | Knee pain                                    | Thin fur                                   | Sallow complexion                         | Stringy pulse               |
| Less urine volume                           | Weak knee                                    | Hard stool                                 | Effulgent gallbladder fire                 | Dark red tongue             |
| Hand edema                                  | Bowel 1 per day                              | Teeth-printed tongue                       | Emaciation                                | Irritability                |
| Loose stool after bowel hard                | Desire for drinking                          | Weak                                      | Soreness of waist                         | Dumb                        |
| Leg swelling                                | No swelling of the lower extremities         | Rib-side distention                        | Loose stool                               | Thirst without desire to drink |
| Difficulty in micturition                   | No pedal edema                               | Dumb                                      | Lassitude                                 | Brown macules on the skin   |
| Bowel 1 per day                             | Normal sleep                                 | Normal urination                           | Chest, epigastric fullness, and distress  | Normal diet                 |
| Normal sleep                                | Depression                                   | Normal sleep                               | Lip color: purple                         | Epistaxis                   |
| Normal diet                                 | Loose stool                                  | Bowel 3 per day                            | Abnormal diet                             | Bowel 1 per day             |

Symptoms indicate that the patterns of TCM syndrome differentiation have high quality.
patterns and outperforms several baseline methods. Furthermore, this study provides a framework for TCM intelligent diagnosis. However, this novel model requires annotated datasets which are often difficult to obtain.

In future work, we plan to incorporate more medical information into the model in our framework, such as disease location, pathogeny, and nature of disease in order to discover more accurate syndrome patterns. In addition, the same symptom could be described by different terms in the experimental data. This may degrade the performance of our method, so we will consider adopting metric learning for normalizing symptom in medical records in the future.

Data Availability

The authors collected 1959 medical records on CKD from Beijing Dongzhimen Hospital, which include 948 (48.4%) females and 1011 (51.6%) males only to support the research work. Because these records involve the patients’ privacy information, the authors have not obtained the authorization of the hospital. Therefore, the authors cannot publish them on the Internet for public sharing at present.

Conflicts of Interest

The authors declare that there are no potential conflicts of interest.

References

[1] L. Yao, Y. Zhang, B. Wei et al., “Discovering treatment pattern in Traditional Chinese Medicine clinical cases by exploiting supervised topic model and domain knowledge,” Journal of Biomedical Informatics, vol. 58, no. C, pp. 260–267, 2015.
[2] L. Yao, Y. Zhang, B. Wei, W. Zhang, and Z. Jin, “A topic modeling approach for traditional Chinese medicine prescriptions,” IEEE Transactions on Knowledge and Data Engineering, vol. 30, no. 6, pp. 1007–1021, 2018.
[3] X. Zhou, Y. Peng, and B. Liu, “Text mining for traditional Chinese medical knowledge discovery: a survey,” Journal of Biomedical Informatics, vol. 43, no. 4, pp. 650–660, 2010.
[4] M. H. Wu and X. Y. Wang, Internal Medicine Of Traditional Chinese Medicine, China Press Of Traditional Chinese Medicine, Beijing, China, 9nd edition, 2012.
[5] C. X. Liu and Y. Shi, “Application of data-mining technologies in analysis of clinical literature on traditional Chinese medicine,” Chinese Journal of Medical Library and Information Science, vol. 20, no. 9, pp. 6–8, 2011.
[6] X. Zhao, L. He, Q. Xie et al., “A novel classification method for syndrome differentiation of patients with AIDS,” Evidence-based Complementary and Alternative Medicine, vol. 2015, no. 6, 8 pages, Article ID 936290, 2015.
[7] J. Chen, D. Yang, Y. Cao et al., “Syndrome differentiation and treatment algorithm model in traditional Chinese medicine based on disease cause, location, characteristics and conditions,” IEEE Access, vol. 6, Article ID 71813, 2018.
[26] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, “The author-topic model for authors and documents,” in *Proceedings of the 20th Conference On Uncertainty In Artificial Intelligence*, pp. 487–494, Banff, Canada, July 2004.

[27] B. Liu, X. Zhou, Y. Wang et al., “Data processing and analysis in real-world traditional Chinese medicine clinical data: challenges and approaches,” *Statistics in Medicine*, vol. 31, no. 7, pp. 653–660, 2012.

[28] Z. Jiang, X. Zhou, X. Zhang, and S. Chen, “Using link topic model to analyze traditional Chinese Medicine Clinical symptom-herb regularities,” in *Proceedings of the 2012 IEEE 14th International Conference on e-Health Networking, Applications and Services (Healthcom)*, pp. 15–18, Beijing, China, October 2012.

[29] Z. Huang, X. Lu, and X. Duan, “Latent treatment pattern discovery for clinical processes,” *Journal of Medical Systems*, vol. 37, no. 2, 2013.

[30] R. Balasubramanyan and W. W. Cohen, “Block-LDA: jointly modeling entity-annotated text and entity-entity links,” in *Proceedings of the 11th SIAM International Conference on Data Mining*, pp. 450–461, Arizona, AZ, USA, April 2011.

[31] I. Yoo, P. Alafaireet, M. Marinov et al., “Data mining in healthcare and biomedicine: a survey of the literature,” *Journal of Medical Systems*, vol. 36, no. 4, pp. 2431–2448, 2012.

[32] Z. Huang, W. Dong, L. Ji, C. He, and H. Duan, “Incorporating comorbidities into latent treatment pattern mining for clinical pathways,” *Journal of Biomedical Informatics*, vol. 59, pp. 227–239, 2016.

[33] X. P. Zhao, X. Z. Zhou, H. K. Huang, Q. Feng, S. B. Chen, and B. Liu, Topic model for Chinese medicine diagnosis and prescription regularities analysis: case on diabetes,” *Chinese Journal of Integrative Medicine*, vol. 17, no. 4, pp. 307–313, 2011.

[34] A. Van Esbroeck, C.-C. Chia, and Z. Syed, “Heart rate topic models,” in *Proceedings of the 26th AAAI Conf. On Artificial Intelligence*, pp. 1635–1641, Ontario, Canada, July 2012.

[35] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, “Disease prediction by machine learning over big data from Healthcare communities,” *IEEE Access*, vol. 5, pp. 8869–8879, 2017.

[36] Y. Gu, Y. Wang, C. Ji et al., “Syndrome differentiation of IgA nephropathy based on clinicopathological parameters: a decision tree model,” *Evidence-based Complementary and Alternative Medicine*, vol. 2017, no. 2, 11 pages, Article ID 2697560, 2017.

[37] F. F. Li and P. Perona, “A Bayesian hierarchical model for learning natural scene categories,” in *Proceedings of the 2005*, vol. 2, pp. 524–531, San Diego, CA, USA, June 2005.

[38] E. Bart, M. Welling, and P. Perona, “Unsupervised organization of image collections: taxonomies and beyond,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 1, 2011.