SPANDIDOS PUBLICATIONS

EXPERIMENTAL AND THERAPEUTIC MEDICINE 12: 288-296, 2016

Study of TCM clinical records based on LSA and LDA SHTDT model

FAN LIN, ZHIHONG ZHANG, SHU-FU LIN, JIA-SONG ZENG and YAN-FANG GAN

Software School of Xiamen University, Xiamen, Fujian 361009, P.R. China

Received October 20, 2015; Accepted April 20, 2016

DOI: 10.3892/etm.2016.3285

Abstract. Description of syndromes and symptoms in traditional Chinese medicine are extremely complicated. The method utilized to diagnose a patient's syndrome more efficiently is the primary aim of clinical health care workers. In the present study, two models were presented concerning this issue. The first is the latent semantic analysis (LSA)-based semantic classification model, which is employed when the classification and words used to depict these classfications have been confirmed. The second is the symptom-herb-therapies-diagnosis topic (SHTDT), which is employed when the classification has not been confirmed or described. The experimental results showed that this method was successful, and symptoms can be diagnosed to a certain extent. The experimental results indicated that the topic feature reflected patient characteristics and the topic structure was obtained, which was clinically significant. The experimental results showed that when provided with a patient's symptoms, the model can be used to predict the theme and diagnose the disease, and administer appropriate drugs and treatments. Additionally, the SHTDT model prediction results did not yield completely accurate results because this prediction is equivalent to multi-label prediction, whereby the drugs, treatment and diagnosis are considered as labels. In conclusion, diagnosis, and the drug and treatment administered are based on human factors.

Introduction

Treatment based on syndrome differentiation (TBOSD) is the feature and essence of traditional Chinese medicine (TCM) (1,2) and is the principle that should be adhered to when making a diagnosis and administering treatment. It has also been proven by long-term medical practice that TBOSD has its specificity, superiority and necessity (1). Irrespective of whether the type of disease, TBOSD constitutes a flexible method that can be employed according to the individual patient's specific condition, which largely enriches the capability of handling diseases of TCM (2). Syndrome differentiation of TCM is the production of long-term clinical practice. There are many types of syndrome differentiation, including that of viscera, etiological analysis and syndrome differentiation of triple energizer (3).

Data mining is a method of extracting potentially useful information from a database. This process uses computer programs, automatically searches the database and identifies modes or rules (3). Networks can be used to describe the associations of individuals, kinships, and network connections via use of computer. Increasingly, investigators use networks in the medical field, and conduct searches on the connection of the brain function (4), propagations of the diseases (5), study of drug efficacy and drug targets (6), gene regulatory networks (7) and protein interactions (8).

The application of quantitative modes and data mining is developing rapidly. Decision tree, KNN, and bayes are classifying methods (9-12) with their own approaches, and can be successfully employed in certain situations. However, TCM is a traditional medicine that captures the variations of the disease based on the concept of wholism. The traditional approaches did not reveal the meaning of the four diagnostic methods because the associations pertaining to the information are complex. TCM has strong correlation content, and is therefore processed more adequately in semantic space (13-15). In the present study, two models are posited, depending on whether classifications have been confirmed according to different situations.

Materials and methods

Latent semantic analysis (LSA) based semantic classification model. The LSA model is used when a classification and the words thereof have been confirmed. The LSA-based semantic classification model of syndrome differentiation, is dependent on the feature of TCM whereby each syndrome and organ has their own major clinical manifestation collection (Table I). This model includes three major steps: i) Decomposition of the matrix of syndromes/organs and clinical manifestation using singular value decomposition (SVD), ii) construction of the semantic space of syndromes/organs and clinical manifestation, and iii) conducting semantic matching of syndromes and organs as per correlative degrees, which are in descending order.

Correspondence to: Zhihong Zhang, Software School of Xiamen University, General office 308B, Xiamen, Fujian 361009, P.R. China
E-mail: zhihong@xmu.edu.cn

Key words: latent semantic analysis, tradition Chinese medicine diagnosis, potential Lejeune Dirichlet allocation model

Table I. Semantic description form of syndromes and organs.

Syndromes/organs	Major clinical manifestation
Xin qi (kui) xu zheng (syndromes)	Palpitation, shortness of breath, mental weariness, spontaneous sweating, pale face, pale tongue, weak pulse
Fei qi (kui) xu zheng (syndromes)	Cough, shortness of breath, asthma, clear thin phlegm, low voice, spontaneous sweating, anemophobia, pale tongue, weak pulse
Pi qi (kui) xu zheng (syndromes)	Consumption of less food, abdominal distension, thin loose stools, mental weariness, Physical weariness, pale tongue, weak pulse
Xin qi xu xue hen zheng (syndromes)	Palpitation, shortness of breath, chest tightness, cardiodynia, mental weariness, dark purple face, lilac tongue, weak pulse
Shen qi (kui) xu zheng (syndromes)	Tinnitus, soreness of waist, attenuated libido, dizziness, unconsciousness, weak pulse
Xin xi lei zheng (organs)	Palpitation, hang-ups, chest tightness, dreaminess, insomnia, dizziness, red tongue, thirst, cardiodynia, intermittent pulse, fever, red face, mental weariness, cold chills, thready weak pulse, disorderly speech, unconsciousness, limb cooling, weak pulse, shortness of breath

If the syndrome has the highest correlative degree with a particular organ, the syndrome was classified into that organ.

Symptom-herb-therapies-diagnosis topic (SHTDT). SHTDT is used in a situation where classification and the relevant words have yet to be confirmed. The core ideas of the SHTDT model posited in the present study involve the assumption that a patient has multiple combinations of symptoms and the corresponding TCM, diagnosis and treatment. The first step involves combination of the symptoms and TCM, extracting the symptoms theme and considering the treatment and diagnosis as the description of symptoms, and extract the multinomial distribution on the theme. The SHTDT model allows selection of drug therapy based on the specific symptom, the treatment selection for combating the combination of symptoms, the possible disease patients suffer from, and can predict the possible drugs, treatment and diagnosis for the patients.

Table II. Semantic description form of syndromes and organs.

Val	Label
188	Palpitation
48	Shortness of breath
24	Mental weariness
24	Spontaneous sweating
28	Pale face
64	Pale tongue
66	Weak pulse
148	Chest tightness
66	Cardiodynia
26	Dark purple face
54	Lilac tongue
26	Astringent weak pulse

LSA model

Constructing model. As an algebraic model of information retrieval, LSA was suggested by Susan and other investigators working at Bell Telephone laboratories in 1988 (16-18). It is a calculation theory and method that has been used for knowledge acquisition and representation. With 20 years of development, LSA, which has advantages including strong computability and a decreased requirement of patient involvement, surpasses the disadvantage of the vector space model (VSM) analytical method. In the present study, an LSA-based semantic classification model of syndrome differentiation was used (Fig. 1).

The latent semantic space constructed by SVD was the core of the model. As the basic semantic meaning of syndromes and organs was described in the clinical manifestation collection, the semantic meaning of syndromes and organs is labeled with clinical manifestation collection. Subsequently, classification occurs in the latent semantic space and the correlative degrees with space vectors of syndromes and organs are computed and sorted, and the corresponding organ whose correlative degree is highest as the belonging class is selected.

This model is relatively easy to extend and can be used to classify any aspect on the condition that the classes and objects to be classified have the same description collection, such as the 5-classes classification (the 5 elements) in the present study.

SVD. The definition of SVD is as follows:

i)

$$A = U \sum V^T$$

where U is a mxm dimensional orthogonal matrix whose column vectors are left singular vectors of matrix A. V is an

Table III. Frequency matrix of syndromes and organs.

Clinical manifestation	Xin qi (kui) xu zheng	Fei qi (kui) xu zheng	Pi qi (kui) xu zheng	Xin qi xu xue hen zheng	Shen qi (kui) xu zheng	Xin xi lei zheng
Cough	0	1	0	0	0	0
Palpitation	1	0	0	1	0	1
Shortness of breath	1	1	0	1	0	0
Pale tongue	1	1	1	1	0	0
Spontaneous sweating	1	1	0	0	0	0
Chest tightness	0	0	0	1	0	1

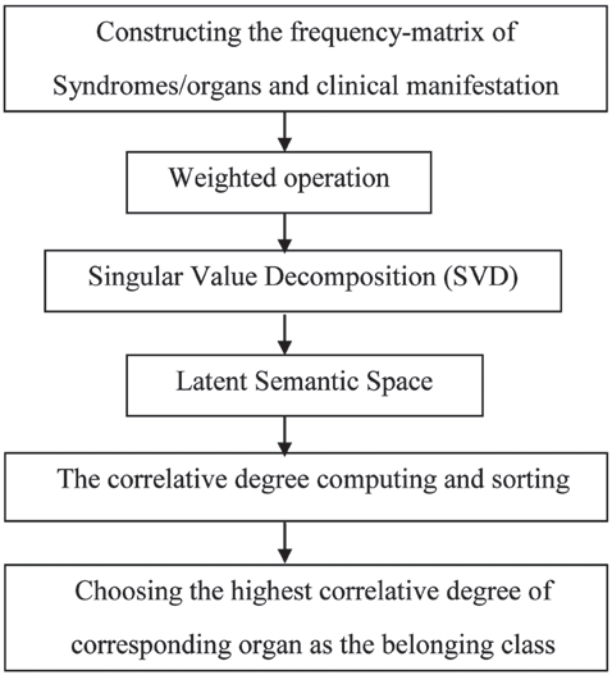

Figure 1. The latent semantic analysis-based semantic classification model of syndrome differentiation.

nxn dimensional orthogonal matrix whose row vectors are right singular vectors of matrix A. ∑ is an mxn dimensional diagonal matrix whose elements, σ1≥σ2≥…≥σr [r≤ min (m, n)], are singular values of matrix A. Decomposition such as this can be applicable to any matrix. In addition, the rank of matrix A can be the total numbers of non-zero singular values. The definition is as follows:

ii)

$$\|A_F\| = \|U\sum V^T\|_F = \|\sum\|_F = \sqrt{\sum_{j=1}^{\gamma_A}\delta_j^2}$$

where the top γ_A columns of matrix U is based on the column vectors of matrix A, and the top γ_A rows of matrix V are based on the row vectors of matrix A. To obtain the similar matrix of matrix A, A_k ($k \le \gamma_A$), singular values (in addition to the k highest ones) are altered to zeros. As is shown in the theory of SVD by Brain (19), the distance between matrix A and its similar matrix is determined by minimizing similar matrix A_k. This is indicated as follows:

iii)

$$\|A - A_k\|_F = \min_{rank(X)<k}\|A - X\|_F = \sqrt{\delta_{k+1}^2 + \delta_{k+2}^2 + \cdots + \delta_{tA}^2}$$

where $A_k = U_k \sum_k V_k^T$, U_k is a tx k dimensional matrix whose columns are the top k columns of matrix U, and V_k is a dxk dimensional matrix whose rows are the top k rows of matrix V. ∑ is a kxk dimensional diagonal matrix whose diagonal elements are the highest k singular values of matrix A. In the case that k is known, it is possible to identify the optimal similar matrix A_k by using SVD (19).

Matrix generation and weighting function

Matrix generation. Major features from samples of each organ are extracted, which represents a certain organ by degrees of vertices. The more degrees the vertex is, the higher the possibility the vertex is to be selected. As shown in Table II, palpitation has a high degree and is likely to be selected. The frequency matrix of syndromes and organs are then constructed based on Table I, as shown in Table III. Subsequently, the frequency matrix in the first step with weighting function was processed, in order to calculate the final mxn dimensional matrix of syndromes and organs $X = [x_{ij}]$:

iv)

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{m1} & x_{m2} & \cdots & xmn \end{bmatrix}$$

where x_{ij} is the weight of the clinical manifestation i in syndromes/organs j; row vector $x_i = [x_{i1}, x_{i2}, ..., x_{in}]$, $i = \{1, 2, ... m\}$ is the weight of clinical manifestation i in each syndrome/organ corresponding to one row of matrix x; column vector $x_j = [x_{1j}, x_{2j}, ..., x_{mj}]^T$, $J = \{1, 2, ..., N\}$ is the syndromes/organs vector corresponding to one column of matrix X (20,21).

Weighting function. The weight in traditional vector space is obtained using the method term frequency/inverse document frequency (TF/IDF) from statistical computing on the marked frequency of clinical manifestation in syndromes/organs. However, the simple structure TF/IDF cannot effectively provide the expression that indicates the importance and distribution of clinical manifestation. Therefore, it is inappropriate to continue using TF/IDF in the LSA-based semantic

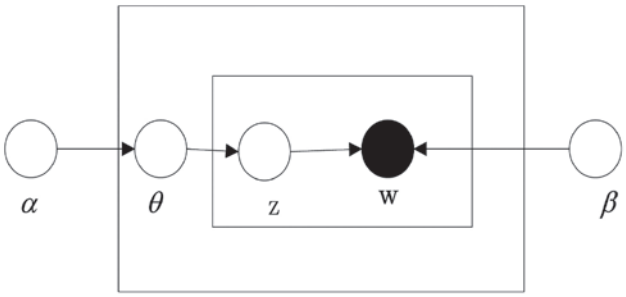

Figure 2. Lejeune Dirichlet allocation model graph.

annotation. Thus, the improved computing method shown earlier (22) was employed to compute the weight, which was divided into a) the global weight of clinical manifestation, and b) the global weight of syndromes/organs.

Global weight of clinical manifestation. Global weight of clinical manifestation, marked $T_w(i)$, is defined as:

v)

$$T_w(i) = 1 - \sum_j \left(\frac{p(i,j)}{\sum_{j=1}^{|R|} p(i,j)}\right) \times \log_2\left(\frac{p(i,j)}{\sum_{j=1}^{|R|} p(i,j)}\right)$$

where |R| is the quantity of syndromes/organs as the total frequencies of clinical manifestation Ti in the whole syndromes/organs collection Rj. H(R|Ti) can be zero. Adding '+1' created a positive number.

Global weight of syndromes/organs. The global weight of syndromes/organs, marked-$R_w(j)$, was:

vi)

$$R_w(j) = 1 + \sum_i \left(\frac{p(i,j)}{\sum_{j=1}^{|T|} p(i,j)}\right) \times \log_2\left(\frac{p(i,j)}{\sum_{j=1}^{|T|} p(i,j)}\right)$$

where |T| is the quantity of clinical manifestation; $\sum_{i=1}^{|\gamma|} P(i, j)$ is the total frequency of clinical manifestation Ti in syndromes/organs Rj. Adding '+1' in the formula $R_w(j)$ created a positive number.

Definition of weighting function. The weighting function is composed by equations (v) and (vi) and was defined as:

vii)

$$W(i,j) = T_w(i,j) \times R_w(i,j)$$

The advantage of the weighting function is that it considers the TCM organ classification in its entirety: a) Each syndrome/organ is regarded as a point in the space that uses clinical manifestation as the dimension, and b) each clinical manifestation is regarded as a point in the space that uses syndrome/organ as the dimension.

Correlation calculation. Correlation calculation is considered based on the formula of similarity calculation. Commonly used similarity calculating formulas are the inner product formula, Pearson formula, Dice coefficient method formula, Jaccard coefficient method formula and cosine formula (22). As the information of the syndromes and organs was expressed as vectors, the cosine formula was used to calculate the degree of correlation. The syndromes and organ vectors D processed by LSA, were divided into part of a) organ, and b) syndrome vectors:

viii)

$$D = X \times T_k \times S_k^{-1}$$

Therefore, the degree of correlation was calculated in the k-dimension semantic space with D_d and D_q. The formula used to calculate C_q was:

ix)

$$C_q = sim(D_q, D_{d_j}) = \frac{\sum \left(D_{d_{ij}} \cdot D_{q_i}\right)}{\sqrt{\sum_{i=1}^{k} \left(D_{d_{ij}}\right)^2 \cdot \sum_{i=1}^{k} \left(D_{q_i}\right)^2}}$$

where sim (D_q, D_{dj}) was the angle cosine value of the syndrome q vector and organ vector dj. It was determined that the bigger the value, the greater the degree of correlation.

SHTDT model

Lejeune Dirichlet allocation (LDA). Given a document collection, LDA expresses each document as a theme set, each topic is a multinomial distribution and is used for capturing the relevant information between words (23). In the LDA, the themes are shared by all the documents and embodied by the specific vocabulary in the text (24). Therefore, the implicit theme may be considered as the probability distribution of the vocabulary, and a single document as the mixture of the implicit theme in specific proportion.

The LDA is a modeling method of the text theme information using probability (25). As shown in Fig. 2, it contains the words, topics and document of three institutions. The (α, β) is the parameters of the document collection layer, which determines the LDA model. In the document collection, α is used to describe the relative strength between the themes, β is used to describe the probability distribution of the implicit theme, and θ constitutes a document layer parameter, with the component of θ indicating the weight of each implicit theme of the target. The (z, w) constitutes a word layer parameter, z is the share of implicit theme each word accounts for, and w denotes the word vector of the target document.

SHTDT model. In the theory framework of the theme model and the background of the application of the TCM (24), an

Table IV. The Gibbs sampling process of SHTDT model based on weight.

Characteristics
1) For i=1 to n
2) Assign topic randomly $z_i \in 1 \ldots T$
3) According to the symptoms-drug frequency, select the drug of the greatest probability for the corresponding symptoms $x_i \in \|m_q\|$ ($m_q \in m_p$)
4) According to the symptoms-treatment methods word frequency, select the treatment of the greatest probability for the corresponding symptoms. //$\|m_p\|$ is the treatment set with patient p, with $\|m_p\|$ being the diagnosis set with patient p
5) According to the symptoms-diagnosis word frequency, select the diagnosis of the greatest probability for the corresponding symptoms. yi $\in \|r_q\|$ ($r_q \in r_p$)
6) Generate the initial distribution of the symptoms, Chinese medicine, treatments and diagnostic (ϕ, θ, η, ε) according to the formula (iv).
7) Repeat
8) For i=1 to n
9) For j=1 to $\|m_p\|$, where $\|m_p\|$ is the corresponding TCM for patient p
10) For k=1 to T
11) According to the formula (i), calculate the corresponding probability value, and obtain the theme k and TCM j that meet the condition of arg $\max_{j,k} p$ ($z_i=k$, $x_i=j$ | $w_i=n$, z_{-i}, x_{-i}, $weight_i$)
12) Update the symptoms and drug distribution. ϕ, θ according to formula (i)
13) For l=1 to $\|t_p\|$
14) Calculate the probability value of 1 to each theme j, obtain the treatment that meets the condition of arg $\max_i p$ ($u_i=l$ | $w_i=n$, $z_i=k$, u_{-i}, t_p, $weight_i$), and then update the treatment distribution η according to formula (ii)
15) For s=1 to $\|r_p\|$
16) Calculate the probability value of s to each theme j, obtain the diagnosis that meets the condition of arg $\max_s p$ ($y_i=s$ | $w_i=n$, $z_i=k$, u_{-i}, r_p, $weight_i$), and then update the diagnosis distribution ε according to formula (iii)

Repeat the process until the change is small enough to oversee or the the number of iterations reach the limit. SHTDT, symptom-herb-therapies-diagnosis topic; TCM, traditional Chinese medicine.

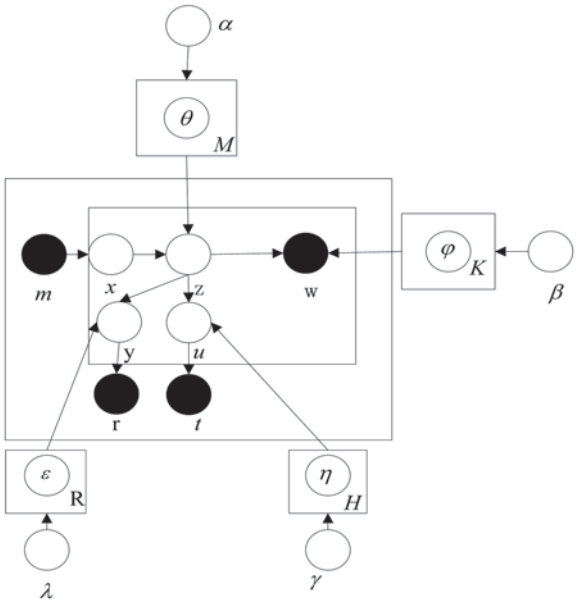

Figure 3. Symptom-herb-therapies-diagnosis topic model graph.

SHTDT model, inspired by the hidden structure method was established (Fig. 3). The four dark circles are the significant variables, w is the sample symptoms, m is the corresponding TCM for a collection of symptoms, t denotes the corresponding treatment under the theme, t denotes the corresponding diagnosis set under the theme, open circles are the hidden variables, the outermost rectangles is the number of patients, the internal rectangular box indicates the sample of N types of symptoms and their corresponding themes and drugs with regard to the patient, the sample of the treatment methods and the diagnosis in the corresponding theme concerning the patient.

Estimating the SHTDT model parameters using the Gibbs sampling method. Given the symptoms of the first n ($w_i=n$) and the drug vector m of the current patient, the Gibbs method was used to estimate the probability of current symptoms assigned to the topic $Z_i=k$ according to the distribution of the theme and the drugs of symptoms except w_i, based on the sampling form shown in formula i. At the same time, the Gibbs method was used to estimate the probability of each drug of the m ($x_i=j$) to the current symptoms with the corresponding theme. Subsequently, according to the treatment vector u and theme distribution $Z_i=k$, based on the form shown in formula ii, the Gibbs method was used to estimate the probability of each treatment ($u_i=1$) of u to the theme $Z_i=k$. Furthermore, according to the diagnosis vector r and theme distribution $Z_i=k$, based on the form shown in formula iii, the Gibbs method was used

Table V. Part of syndromes vector set.

Part of syndromes vector set
1.90945757e-002 -4.24046812e-002 3.65377378e-003 2.40735542e-002 -6.16202858e -003-7.23660460e-003 4.30610050e-002
4.00585125e-002 -3.72604253e-002 -2.31163110e-002 4.73920401e-002-3.88408266e -002-2.76001384e-002 4.73858843e-002
1.00939593e-002 -3.38087295e-002 1.44325399e-002 -5.04584184e-002 1.75236553e-002 3.83762814e-002 -9.25199848e-002
3.30016986e-002 -8.91442827e-002 1.60408602e-002 -4.24170056e-003 5.49615913e-003 1.34018652e-002 4.74668365e-002

Table VI. Part of organs vector set.

Part of organs vector set
1.52142266e-001 5.43477370e-002 2.68646576e-002 -1.55594463e-001 -1.89159278e-001 -3.98234074e-002 -3.36992832e-002
1.84964369e-001 -7.19531062e-003 -1.15767339e-001 -1.66075937e-001 8.40538933e-002 3.99222783e-002 -6.79457783e-002
1.85001434e-001 -1.14156252e-001 -6.39053502e-003 -6.37092024e-002
-1.12233101e-002 6.45070181e-002 -1.45155973e-001
1.71031085e-001 7.05841533e-002 1.79376694e-001 -9.93727433e-002 2.75777159e-002 3.29369106e-002 -4.84512266e-002
1.69728908e-001 -4.58665353e-002 2.18632149e-002 -1.04202173e-001 2.83948127e-002 8.66105955e-002 -3.05467470e-001

to estimate the probability of each diagnosis (r_i=s) of r to the theme Z_i=k.

x

$$P(z_i=k, x_i=j \mid w_i=n, z_{-i}, x_{-i}) \propto \frac{C_{nk}^{VK}+\beta}{\sum_{v'} C_{nk}^{VK}+V\beta} \frac{C_{jk}^{MK}+\alpha}{\sum_{k'} C_{jk'}^{MK}+K\alpha}$$

xi

$$p(u_i=l \mid z_i=k, w_i=n, u_{-i}) \propto \frac{C_{lk}^{HK}+\gamma}{\sum_{k'} C_{lk'}^{HK}+K\gamma}$$

xii

$$p(r_i=s \mid z_i=k, w_i=n, r_{-i}) \propto \frac{C_{sk}^{RK}+\lambda}{\sum_{k'} C_{sk'}^{RK}+K\lambda}$$

According to the Gibbs sampling process, the process was iterated to obtain the distribution of ϕ, θ, η, ε, as indicated:

$$\varphi_{nk}=\frac{C_{nk}^{VK}+\beta}{\sum_{n'} C_{n'k}^{VK}+K\beta},\quad \theta_{jk}=\frac{C_{jk}^{MK}+\alpha}{\sum_{k'} C_{k'j}^{MK}+K\alpha},\quad \eta_{lk}=\frac{C_{lk}^{HK}+\gamma}{\sum_{k'} C_{k'l}^{HK}+K\gamma},\quad \varepsilon_{sk}=\frac{C_{sk}^{RK}+\lambda}{\sum_{k'} C_{sk'}^{RK}+K\lambda}$$

Determining the theme number. The theme vector was identified according to the distribution of the theme (from β matrix) in the V-dimensional word space. The similarity between the theme vector was measured by the standard vector cosine distance based on the formula shown in (iv).

xiii

$$corre(z_i, z_j)=corre(\beta_i, \beta_j)=\frac{\sum_{v=0}^{V}\beta_{iv}\beta_{jv}}{\sqrt{\sum_{v=0}^{V}(\beta_{iv})^2\sum_{v=0}^{V}(\beta_{jv})^2}}$$

xiv

$$avgcorre(struc)=\frac{\sum_{i=1}^{K-1}\sum_{j=i+1}^{K} corre(z_i, z_j)}{K*(K-1)/2}$$

where the smaller the *corre* (z_i, z_j), the smaller the correlation of the theme, and the stability of the structure of the theme according to the average similarity between all subjects based on the formula shown in (v).

SHTDT model based on the weight. Improvements to the SHTDT model were made based on the weight, initially when stacking for each set, and adding the *weight*$_i$, instead of not adding 1. The weight was derived from the distribution based on the Gaussian function. Since for the diagnosis and treatment data, each symptom in each patient sample generally appeared only once, TF=1. However, if weighted using TF-IDF, it would lead to an increase in the weight of the low-frequency words, and a decrease in the weights of the high-frequency words, which was not a viable result. In addition, during the SHTDT model's initialization, the random assignment of the variables was provided from a wider range of variables. However, the assignment accuracy was not high, which affected the subsequent circulation sampling link. Thus, use of the statistical principle leads to yielding statistics of the drug-symptom, treatment-symptom, and diagnosis-symptom correlations of the patients' record, resulting in the sorting of each combination. For example, to allocate drugs for certain symptoms, ininitally x_i=*rand.next* (0, the patient of the corresponding drugs). Based on the correlation of the statistics (drug-symptoms), in the corresponding drug episodes of the patients, the most frequent drug was employed. If several drugs were equally frequent, one drug was randomly selected. Of note is that the smaller the selected range, the higher the assignment accuracy. The algorithm of SHTDT model based on weight is shown in Table IV.

Results and Discussion

Results of LSA model. The dataset used in the present study is derived from the clinical data of the Zhongshan Hospital

of Xiamen University. There were 588 clinical manifestations used to semantically describe 251 syndromes and 5 organ cards (23). According to the dataset, a comparison was made of the performance of non-LSA and LSA prior to and following the experiment (Fig. 4). As there are many matrix computing in the model, so MATLAB is applied to do SVD. And cosine vector method is applied to compute correlative degree. Part of syndrome vectors and organ vectors that has been processed by LSA (k=7) are shown as Tables V and VI.

Fig. 4 shows that, the performance of LSA-based semantic classification model of the syndrome differentiation classifier was more effective than that of the non-LSA classifier (24). The main reason for this finding is that LSA maps the high dimension VSM to the low dimensional latent semantic space. At the same time, the 'noise' (irrelevant information) was also removed. Compared with the traditional vector space, the dimension of the latent semantic space is smaller and a semantic relationship is clearer.

As shown in Fig. 5, weighted-LSA performs more effectively than non-weighted-LSA. Thus, feature weighting improves the performance of classification. Feature weighting reduces the interference of high frequency words and stop words (reducing their representative) and improves the representative keywords' function (improving the differentiation) so as to increase the classification accuracy.

The experimental result showed that LSA was successfully applied in the TCM field, although additional studies should be conducted to confirm the results. Although the experiment was small scale, an advantage of LSA was identified. Thus, this method may be applied successfully in future.

Results of SHTDT model. The attributes of symptoms, Chinese medicine, treatment and diagnosis for each case were screened. After supplementing any lacking data, the deletion of redundant data, the uniform regulation of the symptoms of the term, the unitary drug name, and the specification of representation format, the best theme number was determined in accordance. In Fig. 6, k is identified as 12, following which the SHTDT model was run based on the weight. Two typical themes were identified, each of which lists the symptoms, Chinese medicine, treatment and diagnosis of the first 10.

It is extremely difficult to comprehend semantic meaning at present. However, latent semantic comprehension is practically feasible. The application of LSA makes the meaning of vectors change as they reflect the distributed relationship of clinical manifestation, and reinforce the semantic meaning of vectors. Thus, vectors are based on lexemic and semantic strata. Performing a correlative analysis in such a new semantic space yields better results compared to the original feature vector. Because of SVD, the LSA-based semantic classification model of syndrome differentiation suppresses the 'noise' and reduces the dimensions of matrix. The semantic relationship between organs and syndromes is guaranteed. Additionally, it has high computability and strong operability and solves the issue of matrix sparsity. However, there are factors that remain to be investigated, such as obtaining k in SVD, and a more viable option of clinical manifestation. These factors are likely to affect the whole classification effect.

Table VII shows that, the theme pertains to spleen deficiency syndrome, and refers to lack of temper, weak symptoms

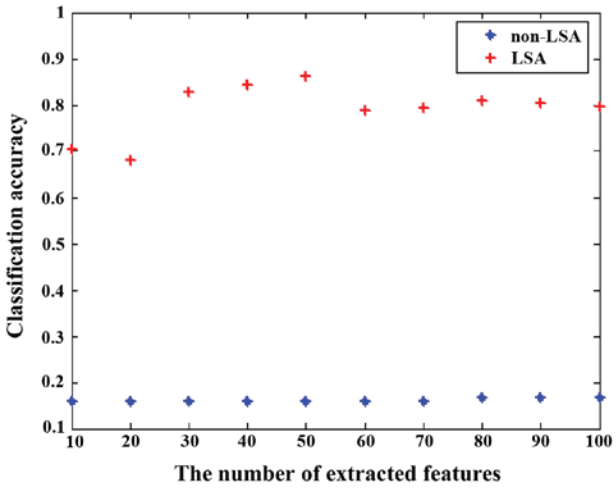

Figure 4. The performance of non-latent semantic analysis (LSA) and LSA.

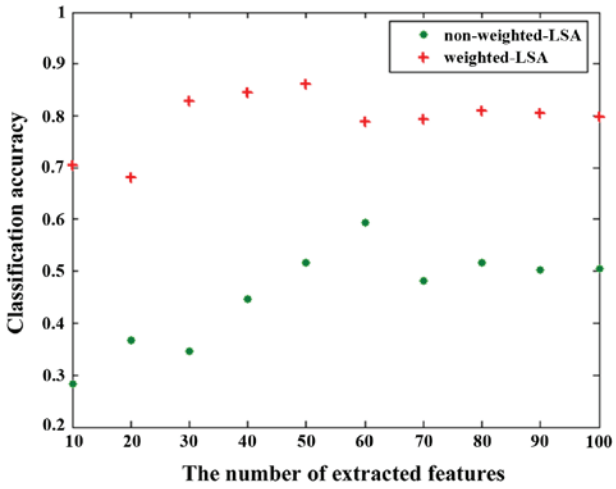

Figure 5. The performance of non-weighted-latent semantic analysis (LSA) and weighted-LSA.

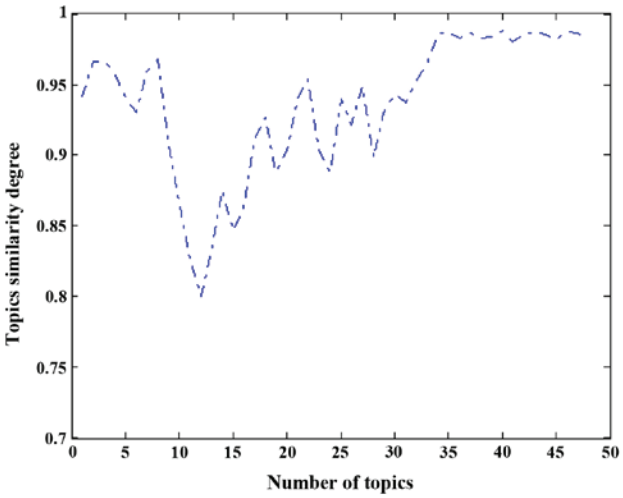

Figure 6. Determine the best theme number.

of transport and dereliction of digestion, with the general performance of consuming less, lack of blood, Shenpi fatigue,

Table VII. The probability distribution of symptoms, Chinese medicine, treatment and diagnosis concerning theme 3.

Symptoms probability	TCM probability	Treatment probability	Diagnosis probability
Shortness of breath 0.0564	Lobelia 0.6766	Help breathing 0.2166	Deficiency of lung 0.4020
Pale complexion 0.0432	Hyacinth 0.6303	Detoxification 0.1715	Qi phlegmy heat 0.3987
Epigastric discomfort 0.043	Nourish 'Yin' 0.1687	Nourish 'Yin' 0.1687	Spleen-lost-all-blood 0.3906
Less bloating 0.0268	Yuan hu 0.5210	Anti-cancer 0.1592	Moisture to stay 0.3359
Moderate sleep effect 0.0251	Scutellaria barbata 0.5840	Invigorating spleen 0.1561	Qi and Yin injury 0.3245
Fatigue 0.01988	North Adenophora 0.4391	Reinforcing stomach 0.1522	Spleen Qi deficiency 0.3133
Poor appetite 0.0181	Bai ji 0.4079	Eliminate bloating 0.1381	Blood stagnation 0.2968
Sweating 0.0166	Amomum 0.3831	Consumer product 0.1368	Gas-and-Yin-deficiency 0.2756
Anorexia 0.0165	Lily 0.37715	Moist lung 0.1332	Gas and blood deficiency 0.2683
Emaciation 0.01589	Pseudostellaria-heterophylla 0.3668	Antiperspirant 0.1271	Physically weak and poison accumulation 0.2614

TCM, traditional Chinese medicine.

Table VIII. The probability distribution of symptoms, Chinese medicine, treatment and diagnosis concerning theme 7.

Symptoms probability	TCM probability	Treatment probability	Diagnosis probability
Poor sleep 0.0198	Sanqi powder 0.2998	Digestion 0.1061	Diarrhea 0.1744
Moss thin white 0.0105	Rhubarb 0.2579	Nourishing blood 0.1053	Food retention abdominal pain 0.1732
Consuming less 0.0059	Hawthorn 0.2420	Solid off 0.1028	Stagnation stomach 0.1706
Poor appetite 0.0058	Coke hawthorn 0.2417	Warming the kidney 0.1017	Kidney deficiency blood stagnation 0.1498
Poor appetite 0.0056	Psoralea corylifolia 0.2153	Removing stagnation 0.0957	Cold blood 0.1459
Constipation 0.0055	Pear skin 0.2080	Synthesis 0.0954	Heart deficiency and timidity 0.1410
Anorexia 0.0051	Corydalis 0.2059	Consumer product 0.0933	Qi stagnation 0.1377
Dizziness 0.0045	Curcuma 0.2052	Transfer Qi 0.0924	Chill condensation 0.1354
Nausea, vomit 0.0043	Japonica rice 0.2034	Sweet moisturizing 0.0920	Colorectal hot and humid 0.1348
Irritability 0.0042	Notopterygium 0.2028	Resuscitation 0.0912	Alpine dysentery 0.1344

TCM, traditional Chinese medicine.

heart palpitations, shortness of breath, pale complexion, less bloating, pale tongue, white coating, and weak pulse. The nourishing Qi Diet is recommended for symptoms including food overconsumption, overexertion, lassitude, inadequate endowment, the elderly and infirm, chronic diseases, and critical diseases.

Table VIII shows that the theme primarily pertains to constipation symptoms. Patients with constipation often experience headache, fatigue, poor appetite, bloating, indigestion and other symptoms, which may be associated with poor diet, sedentary lifestyle and personal spiritual factors (25). Therapies include becoming involved in the Qi, laxatives, and regulation of blood lipids for relief (26). In addition, attention should be focused on symptoms with the occult that may be confused with other clinical signs of disease. For example, the cause of hyperlipidemia is that the content of cholesterol, triglycerides, β-lipoprotein and other lipid components in the blood are higher than normal, reflecting a series of pathological changes in the body, including clinical dizziness, chest tightness, palpitations, Shenpi fatigue, insomnia, forgetfulness, numbness and other symptoms as the main performance (26). Similar symptoms are shown in Table VII. Hyperlipidemia, a type of 'rich disease', because of its slow onset, which may trigger coronary heart disease, stroke, diabetes, obesity, and fatty liver disease. Therefore, manifestation of the above symptoms requires seeking medical assistance to prevent disease progression.

In conclusion, it is difficult to comprehend semantic meaning at present, although latent semantic comprehension is practically feasible. The application of LSA makes the meaning of vectors change. They reflect the distributed relationship of clinical manifestation, and reinforce the semantic meaning of vectors. Thus, vectors are based on lexemic and semantic strata. Performing a correlative analysis in such a new semantic space yields a better result compared to the original feature vector. Because of SVD, the LSA-based semantic classification model of syndrome differentiation suppresses the 'noise' and reduces the dimensions of matrix. The semantic relationship between organs and syndromes is guaranteed. In addition, it has high computability and strong operability and

solves the issue of matrix sparsity. However, there are factors that remain to be investigated, such as obtaining k in SVD, and the optimal choice of clinical manifestation. These factors may affect the whole classification effect.

For the TCM diagnosis, a variety of subjective factors exist, but the symptoms and drugs may be considered objective factors. To identify clinic rules from these two objective factors, a model of TCM rules based on the statistics was created and the SHTDT model was suggested. The experimental results have been identified by the Chinese clinical doctors, and the model generated and the results obtained are of great clinical significance. Given patient symptoms, we can predict the theme, drug application, the treatments and diagnosis of the patient using this model. The results of the experiments show that the SHTDT model prediction results were approximate to the actual results, albeit completely accurate results were not yielded owing to the fact that this prediction is equivalent to multi-label prediction, i.e., considering the drugs, treatment and diagnosis as labels. Thus, for a patient, the drugs selected, approach to treatment and diagnosis essentially constitute human-made factors.

References

1. Zhang LW, Duan CL, Xiong ZW and WU H: Study on the application of naive Bayesian methods in identifying syndrome in TCM. J Inner Mongolia Univ 38: 568-571, 2007.
2. Peng JF: Syndrome Element Differentiation Methodology based on Data Mining Technology (unpublished PhD thesis). Hunan University of Chinese Medicine, 2007.
3. Witten IH and Frank E: Data Mining: Practical Machine Learning Tools and Techniques. Vol. 3, 2nd edition. China Machine, Press, Beijing, pp105-106, 2005.
4. Zhang FF, Chen CH and Jiang L: Brain functional connection research based on complex network. Fuza Xitong Yu Fuzaxing Kexue 8: 18-23, 2011.
5. Qin XH and Guan YJ: Viruses spread of influenza AH1N1 based on complex networks. Statistics and Information Forum 25: 86-90, 2010.
6. Yildirim MA, Goh KI, Cusick ME, Barabási AL and Vidal M: Drug-target network. Nat Biotechnol 25: 1119-1126, 2007.
7. Zhou HJ: Application of complex network theory in gene regulatory networks. J Chongqing Univ Sci Technol 11: 141-144, 2009.
8. Fang Z, Li YZ, Xiao JM, Li GB, Wen ZN and Li ML: Complex network-based random forest algorithm for predicting the impact of amino acid mutation on protein stability. Chem Res Appl 23: 554-558, 2011.
9. Doukas C and Maglogiannis I: Enabling human status awareness in assistive environments based on advanced sound and motion data classification. In: Proceedings of the 1st international conference on PErvasive Technologies Related to Assistive Environments, Athens, Greece, Jul 2008. http://dx.doi.org/10.1145/1389586.1389588.
10. Doukas C and Maglogiannis I: Human Distress Sound Analysis and Characterization using Advanced Classification Techniques. In: Artificial Intelligence: Theories, Models and Applications. 5th Hellenic Conference on AI, SETN 2008, Syros Greece, October 2008. Darzentas J, Vouros GA, Vosinakis S and Arnellos A (eds). Springer-Verlag GmbH, Berlin, pp73-84, 2008.
11. Liu H and Huang ST: A fuzzy method to learn text classifier from labeled and unlabeled examples. J Harbin Inst Technol 11: 98-102, 2004.
12. Ma G, Zhu L, Yan G and Chen D: Kernel Method for Building Fuzzy Classifiers. In: Proceedings of the The sixth world congress on intelligent control and automation, 2006. Vol 6. WCICA, pp4307-4311, 2006.
13. World Wide Web Consortium: (W3C). Semantic Web Activity. http://www.w3.org/2001/sw/. Accessed Jun 4, 2008.
14. McGuinness DL and Harmelen FV (eds): OWL Web Ontology Language Overview: W3C Recommendation 10 February 2004. http://www.w3.org/TR/owl-features/. Accessed Jun 4, 2008.
15. Wang CH, Nan LL and Ren YP: Research on the text clustering algorithm based on latent semantic analysis and optimization. In: Proceedings of the Computer Science and Automation Engineering (CSAE), 2011 IEEE International Conference. Vol 4. IEEE, pp470-473, 2011.
16. Zhang XG, Huang GJ, Cao LH and Guo HT: Web services filtrate technologies based on latent semantic analysis. Comput Eng 34: 39-41, 2008.
17. Ishii N, Murai T, Yamada T and Bao Y: Text classification by combining grouping, LSA and kNN. In: Proceedings of the Computer and Information Science, 2006 and 2006 1st IEEE/ACIS International Workshop on Component-Based Software Engineering, Software Architecture and Reuse. ICIS-COMSAR 2006. 5th IEEE/ACIS International Conference on. IEEE, pp148-154, 2006.
18. Jiang Z and Lu C: A latent semantic analysis based method of getting the category attribute of words. In: Proceedings of the Electronic Computer Technology, 2009 International Conference on. IEEE, pp141-146, 2009.
19. He ZL and Wang CH: Application of matrix singular value decomposition (SVD) in latent semantic information retrieval. Mod Comput 6: pp21-23, 2011.
20. Hu M, He Y and Li J: Fault diagnosis method based on LSA and SVM. In: Proceedings of the Information Engineering and Computer Science, 2009. ICIECS 2009. International Conference on. IEEE, pp1-4, 2009.
21. Sun JT, Zhang QY and Yuan ZT: A Junk Mail Filtering Method Based on LSA and FSVM. In: Proceedings of the Fuzzy Systems and Knowledge Discovery, 2008. FSKD '08. Fifth International Conference on. Vol 3. IEEE, pp111-115, 2008.
22. Xuan Y and Zhu Q: Research on tag semantic retrieval in social tagging system based on LSA. Libr Inf Serv 55: 11-14, 2011.
23. Zhang HP and Yu HK, Xiong DY and Liu Q: HHMM-based Chinese Lexical Analyzer ICTCLAS. In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistic. pp1231-1235, 2003.
24. Chu Keming and Li Fang: LDA model-based news topic evolution. Computer Applications and Software 4, 2011. DOI:10.3969/j.issn.1000-386X.2011.04.002.
25. Jing S, Meng F and Li WL: The analysis of the themes based on LDA model. Acta Automatica Sinica 35: 1586-1592, 2009.
26. 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. J Biomed Inform 58: 260-267, 2015.