Structural Technology Research on Symptom Data of Chinese Medicine

Aziguli[1,2], YuanYu Zhang[1,2], YongHong Xie[1,2*], Yang Xu[1,2], YuJia Chen[1,2]
1. School of Computer and Communication Engineering, University of Science and Technology Beijing (USTB), Beijing 100083, China
2. Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing 100083, China
*E-mail: hardmoon7@Hotmail.com

Abstract—**Traditional Chinese Medicine (TCM) symptoms are the basis of the diagnosis and differentiation. Analyzing TCM symptoms is significant for discovering the knowledge of TCM. Different doctors of TCM prefers using different terms for the same symptom, which is not conducive to the standardization of TCM knowledge and hinders the heritage of TCM. This paper presents a solution to structure the symptoms of TCM by constructing two lists, standard TCM symptom list and synonym list), which take standardization rules of TCM symptoms into account. In addition, the algorithm of Chinese literal similarity with the parameters fine-tuned, is applied in field of TCM. Experimental results have shown the effectiveness of the proposed solution.**

Keywords-TCM symptoms; Chinese literal similarity; Synonym; Structure

I. INTRODUCTION

Symptom of TCM is the basis of the diagnosis and differentiation. However, different doctors describe the same symptom with different terms. Therefore, the structural treatment of traditional Chinese medicine could contribute to the discovering of TCM, such as the association analysis between symptoms and syndromes[1].

According to the theory of TCM, "syndrome factors differentiation system"[2] is a cognitive process of identifying disease location and disease nature of syndrome elements, as well as differentiating syndrome types by analyzing syndrome, including symptoms and signs and related data.

Syndrome differentiation is based on a variety of symptoms and signs, referred to as the "syndrome", and the pathological nature, known as "syndrome element", is identified by syndromes. the syndromes are composed of the location of disease and syndrome key factors[3].

A variety of complex logical relationships exist among symptoms, and one symptom could logically derive many symptoms. Zhang Qiming[4] presented a hypothesis of symptom unit (independent symptoms) in the book of the TCM symptom research, with 399 independent symptoms as the core, and descriptive phrases (descriptive symptoms) and derived phrases (derived symptoms) as the basis for classification. Such as headache could derive migraine and occipital pain, and migraine may derive migraine and distending pain, migraine and stabbing pain, migraine and paroxysmal pain. TCM symptoms choose the most appropriate as rectifying name from a set of derived symptoms, and the remaining as alias (synonyms)[5]. If diarrhea is the rectifying name formal name-, then loose stool is the synonym(alternative). Many studies[6, 7, 8] proposed that the symptoms of a variety of clinical manifestations of the terminology should be split, and some researches[9, 10] about resolutions to simply split complex symptoms could be summed up for TCM symptoms or its symptoms with modifiers.

II. ANALYSIS THE SYMPTOMS IN TRADITIONAL CHINESE MEDICAL

It is obvious that symptoms of tongue and the pulse usually conform to certain rules, and the syndrome factors are finite. While the symptoms of other parts are not in accordance with those rules, and not enumerable. Therefore, the TCM symptoms are divided into two categories and processed respectively. In addition, the fine-tuned Chinese literal similarity algorithm is adopted to calculate the similarity between two symptoms.

A. *Analysis of Symptoms of Special Parts*

The special parts indicate the pulse and the tongue. Because the symptom factors of pulse and tongue are finite, it's available to construct the symptom elements list about pulse and tongue by referring to the diagnostics of traditional Chinese medicine and other authoritative books. When cope with the symptom of pulse or tongue, the first step is to extract the disease site and the modifiers in the symptom, then use the list of symptom elements to match the remaining exactly. If there exist some unmatched, then use the combination of the symptom elements to represent the unmatched above. In most cases, this method is feasible to structure the symptoms of pulse and tongue by disease site, symptom elements, and modifiers. Symptom elements are description of symptoms, or the part of symptom that are irrelevant to diseased sites. Besides, symptom elements provide the nature of symptoms, so symptoms of the same kind share common elements[11].

B. *Analysis of General Symptoms*

As this paper researches on the symptoms of medical records which is processed artificially, the symptoms are most made up of phrases with insufficient contextual

978-1-5090-6704-6/17/$31.00 ©2017 IEEE

information. In addition, since there is no large-scale of Chinese medicine corpus, it makes no sense to use statistical methods. If there are any identical Chinese characters in two symptoms, symptoms will be very likely to be similar. As a result, it is meaningful and feasible to calculate the similarity between two symptoms based on the theory of character matching. Due to the differentiation between Chinese and English, some similarity methods for English phrases are unsuitable in this condition.

C. *The Chinese Literal Similarity Algorithm*

Wang[12] first proposed the problem of literal similarity algorithm. Zhu[15] proposed the similarity recognition algorithm based on semantic morpheme, which needs to split words into morphemes. However, the expression of TCM is in the form of semi classical Chinese, and the result of the segmentation algorithm for semi classical Chinese segmentation is unsatisfactory, so this method is not eligible for TCM symptoms. Wu[16] proposed the weighted similarity algorithm based on the Chinese vocabulary and structure characteristics, and combined the principle of "end focus" and "end weight" (In the literal sense, the more the morpheme is on the back, the greater the role it play in the concept of the expression), which is obvious in the TCM symptoms, and this paper is based on this idea.

The morphological similarity between two words (ω_l, ω_r) contains two indicators, the number of the same characters and the positions of the same characters of the two words. Firstly, statistic the number of the same characters of each word. Secondly, according to positions of the same Chinese characters in the words, compute the weight in each word, and the similarity calculation of the two words is shown as follows:

$$\begin{aligned} sim(\omega_l, \omega_r) = \alpha * sim_\alpha(\omega_l, \omega_r) \\ + \beta * PosCoef * sim_\beta(\omega_l, \omega_r) \end{aligned} \tag{1}$$

$$sim_\alpha(\omega_l, \omega_r) = 1/2 * \left(\frac{\|S\|}{\|\omega_l\|} + \frac{\|S\|}{\|\omega_r\|} \right) \tag{2}$$

$$sim_\beta(\omega_l, \omega_r) = 1/2 * \left(\sum_{i=1}^{\|S\|} \frac{\omega_l(\beta_i)}{\sum_{m=1}^{len(\omega_l)} \omega_l(e_m)} + \sum_{j=1}^{\|S\|} \frac{\omega_r(\beta_j)}{\sum_{n=1}^{len(\omega_r)} \omega_r(e_n)} \right) \tag{3}$$

$$PosCoef = \min\left(\frac{\omega_l}{\omega_r}, \frac{\omega_r}{\omega_l} \right) \tag{4}$$

Where S is a set of the same Chinese characters in two words (ω_l, ω_r), S that represents the number of elements in the set; w the total number of Chinese characters in the word w; PosCoef indicates the position coefficient, which represents the structure factor of the similarity computation, and its value is the minimum value of the ratio of the total number of each word.

The algorithm follows the Wang's hypothesis: 60% represents the impact of the same number of characters on the similarity; 40% indicates the effect of the position of the same characters in each word, that is α=60%, β=40%. However, 60% and 40% are empirical values, and they don't have theoretical basis and statistical support[17]. Therefore, Logistic Regression algorithm is adopted to get the parameters based on the TCM data in this paper.

D. *The Tuning of the Similarity Algorithm*

Firstly, construct a list of standard symptoms and synonyms referenced to the TCM knowledge. In this list, standard symptoms are defined as one of its own synonyms. Secondly, construct a 0-1 matrix, whose columns are consisted of the standard symptoms which get from rows are made up of the synonyms, Where the "1" is on behalf of its line of words and their columns are synonymous with the word. On the contrary, "0" is not synonymous with the word. Thirdly, randomly select 8000 of the synonyms from the matrix for 2×3 cross validation.

The parameters (α, β) of the similarity algorithm(Eq. (1)) are obtained with logistic regression. With different combinations of and in the experiment, the accuracy, recall and f1-score of the mentioned algorithm are obtained which are shown in Fig.1, which reports performance of the 2 folds as well as that of the average with different parameters. Surprisingly, the average f1-score of fold1 and fold2 is over 70%, which indicates that the parameters (α, β) of the similarity algorithm taking the average values of fold1 and fold2 are eligible. Therefore, the average values of α and β are adopted in following practical processing.

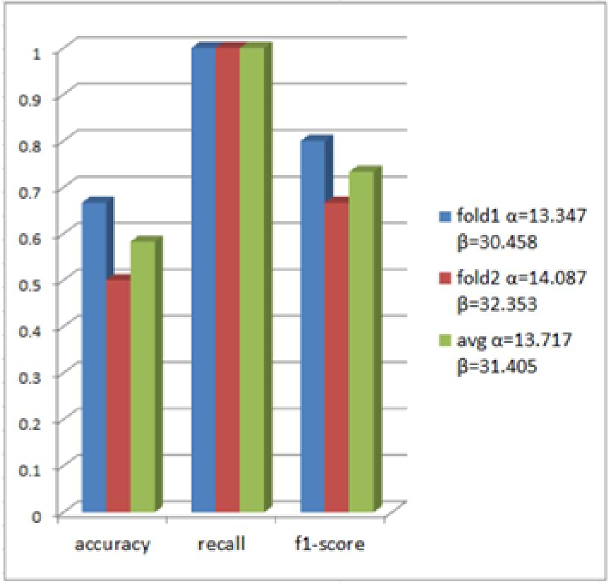

Figure 1. The performance of similarity algorithm with different parameters

III. STRUCTURE THE SYMPTOMS IN TRADITIONAL CHINESE MEDICAL

The experimental data from the Program of the TCM clinical experience, inheritance the academic ideology in

11th Five-year Plan of China, near 19300 copies of medical records. What follows is to introduce the structured process of the symptoms of the tongue and pulse and the partially disassembled moment disease.

A. *Pretreatment Process of Symptoms in Special Parts*

Special parts are the pulse and tongue that has been mentioned above. Due to the data processing of pulse is similar to that of tongue, here present the structural process of the tongue symptoms. As shown in Table 1, extract the tongue symptoms in the medical records, where Ch ID is for medical ID, and tongue coating is for symptoms of tongue.

In the Table 1, it is necessary to extract and encode the words representing disease sites or positions, symptom factors, degree and so on from the Shetai column. For example, “sy317”stands for “thin”, “sy319” represents “greasy”, “wz11” on behalf of “both sides”, “cd5”on behalf of the “a little”. The result of the processing is shown in Table 2.

TABLE I. THE SYMPTOMS OF TONGUE IN MEDICAL RECORDS

Ch ID	Tongue coating
26491	苔两侧薄腻
	Thin and greasy on both sides
17769	苔白根部略厚
	White, thick on the tongue root
19623	苔厚
	Thick

TABLE II. THE STRUCTURE OF SYMPTOMS OF TONGUE

Ch ID	std symtom	modifier
26491	sy317+sy319	wz11
17769	sy313+sy318	wz21; cd5
19623	sy318	

B. *The Preliminary Process of Symptoms in General Parts*

Part of the TCM symptoms (about general parts) is shown in Table 3, Ch ID means the medical record number, Symptom means the symptoms appeared in the medical cases.

Construct a body parts table referenced to the TCM knowledge, and use the hierarchical coding to reflect the subordinate relationship among different parts. In addition, the length of the code is the same for the same layer. For example, upper limb(“bw010701”) is part of limb(“bw0107”), and upper limb includes shoulder(“bw0107011”), arm(“bw0107012”). Then match the symptoms on the list of the standard symptoms on literal exactly. However, only the symptom of “(debility of the legs)” gets the perfect matching result, which is represented by the code “sy1011”.

Further decomposition is necessary for the unmatched result from which extract and encode the words that stand for degree (such as “微(slightly) cd4”), time (such as “下午(afternoon) sj49”), frequency (such as “经常(often) pc121”) and other ingredients (such as “右(right) wz65”, “双(double) wz62”). Then match the remaining with the standard symptoms or synonyms again, and the remaining of “(hypochondrium distension)”, “下肢肿(lower limb swelling)” get the perfect matching results with the coding of “sy142” and “sy983” respectively. While the unmatched result is shown in the column labeled with rest of Table 5.

As shown in Table 5, the columns labeled with “Fist”, “Second” and “Third” are the matching results of the contents in the rest column. The code starting with “sy” represents the matching result that corresponds to the standard symptom. Abandon the result whose disease part is not the same or subordinated to the part of the symptoms in the rest column (such as “心热(heart heat) sy166” whose disease site is the “心(heart)”, and has no affiliation with “手足心(heart(palms and soles)”). Then if the codes of the remaining matching results are the same, the code will be the final result. If the code of the matching results are not consistent, the final result will be the combination of the matching results (such as the result of “(tinnitus and giddiness)” in the sixth row of Table 4.

TABLE III. THE GENERAL SYPMTOMS

Ch ID	Syptom
14799	右胁下胀满
	distension of the bottom right hypochondrium
25850	下午两足跗肿
	tarsus swollen in the afternoon
18338	经常手足心发热
	palm and arch often in calorific
33018	双下肢微肿
	double lower limb slight swelling
21152	耳鸣眼花
	tinnitus and giddiness
31671	腿软
	debility of the legs

TABLE IV. STRUCTURED RESULTS OF GENERAL SYMPTOMS

Ch ID	Std symptom	modifiers
14799	sy142	wz65+wz43
25850	sy1249	wz61;sj49
18338	sy1290+sy1289	pc121
33018	sy983	wz62; cd4
21152	sy955+sy380	
31671	sy1011	

TABLE V. GENERAL SYMPTOMS OF MATCHING RESULTS

rest	First	Second	Third
足跗肿（tarsus swollen）	跗肿（spavin）sy1249	足肿（swollen feet）sy1123	足背肿 (instep swollen) sy1249
手足心发热 (palm and arch in calorific)	足心热（arch in calorific）sy1290	手心热 (palm in calorific)	心热 (heart heat) sy166
耳鸣眼花（tinnitus and giddiness）	眼花（giddiness）sy955	耳鸣 (tinnitus) sy380	

If there is any unmatched result that still exists, statistic the number of the occurrence of the unmatched that do not have proper similar results, and add it to the standard symptom and synonym list whose statistic value is exceed a certain threshold.

After statistical analysis, there were 52340 symptoms (repetitive symptoms included) collected from 19300 copies of medical records. After first matching, there were about 5477 unmatched symptoms (without duplicates) counted, and further processing got only 800 results without matching which indicate the effectiveness of this method for TCM symptoms structure.

IV. CONCLUSION

The culture of TCM is extensive and profound. There is not an excellent standardization system, so as a unified law of the structure. The method proposed in this paper can solve the structural problems of traditional Chinese medicine symptoms, yet some results still need to be determined manually. Therefore, there are still a lot of space for further research.

ACKNOWLEDGE

This work is supported by the National Key Technology R&D Program in 12th Five-year Plan of China (No. 2013BAI13B06)..

REFERENCES

[1] G. P. Liu, S. X. Yan, R. W. Zhen, G. Z. Li, F. F. Li, J. J. Fu, J. Z. Association Analysis between Symptoms (Signs) and Syndromes of Chronic Gastritis Based on Associated Density. Journal of Computational Information Systems. 2012, 8(19): 8239-8247.

[2] W.F. Zhu. Diagnostics of TCM [M]. Beijing: People's Medical Publishing House, 1991.1.

[3] R.C. Zhou. Research and application of terminology standardization of Chinese medicine attending information[D]. Fujian University of Traditional Chinese Medicine. 2009.

[4] Q.M. Zhang, B.Y. Liu, Y.Y. Wang. Research on diagnostics of traditional Chinese Medicine[M]. Beijing: Traditional Chinese Medicine Classics Press,2013.

[5] M.X. Liang, X.F. Wang, D. Dong. The basic problems in the traditional Chinese medicine syndrome differentiation standard[J]. World science and technology modernization of traditional Chinese medicine. 2005, 7(3): 18-23.

[6] B.Q.Huang. The necessity of standardization of TCM symptoms[J]. China Journal of Traditional Chinese Medicine and Pharmacy. 2011, 26(3) : 429-432.

[7] Z.Q. Zhang, Y.Y. Wang, G.Z. Gai. The symptoms of TCM symptoms are independent of each other[J]. journal of beijing university of traditional chinese medicine. 2011, 34 (12):797-799.

[8] W.H. Liu, W.F. Zhu. Some problems of the standardization of TCM symptoms [J]. Chinese Journal of traditional Chinese Medicine, 2007, 48(6): 555-556.

[9] X. Zhang, T.P. Chen, W.L. Li. Characteristics and norms of language description of TCM symptoms[J]. Acta Chinese Medicine and Pharmacology: 2011,39(2):1-2.

[10] H.Y.Zhang, J. Zhang, X.Y. Ma. Study on the standardization of terminology in Department of internal medicine of traditional Chinese Medicine[J]. Liao-ning Journal of traditional Chinese Medicine: 2011, 38(6):1032-1033.

[11] Z.G. Wang, Y.Y. Wang. Symptom Elements and Symptom Standardization[J]. World Chinese Medicine. 2012.7(4): 277-278.

[12] Y. Wang, X.B. Wu, C.W. Tu. Computer processing after control specification [J]. Modern Library and information technology, 1993.2.

[13] M.L. Song. Chinese vocabulary and literal similarity principle and the control on maintaining the Chinese vocabulary[J]. Journal of the China Society for Scientific and Technical Information.1996.4.

[14] Y.Y. Wang. Research and application of Chinese phrase similarity calculation method[D]. Hu nan:Changsha University of Science and Technology,2008.

[15] Y.H. Zhu. Research on synonym recognition algorithm in intelligent search engine[D]. Nanjing Agricultural University,2001.

[16] Z.Q. Wu, H.Q. Hou. Recognition of Chinese synonyms by literal similarity[C]. The Fifteenth National Symposium on computer information management: 222-229.

[17] Y.H. Zhu, H.Q. Hou, Y.T. Sha. Comparison and evaluation of two methods of computer aided Chinese synonyms recognition [J]. Journal of The Library Science in china,2002,28(4):81-84.