Advances in Intelligent Systems Research, volume 153

International Conference on Applied Mathematics, Modeling and Simulation (AMMS 2017)

Research on Traditional Chinese Medicine Case Retrieval Method Based on Machine Learning

Aziguli WULAMU, Yan Xu, Dezheng Zhang* and Daole Li

[1]School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China

[2]Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing 100083, China

*Corresponding author

***Abstract*—In recent years, the wave of artificial intelligence began to rise. In the field of traditional Chinese medicine, decision support system[1] is still in the initial stage, which is a lot of room for development. Therefore, it is very important to establish a practical TCM case-based reasoning system on the basis of TCM case database. Based on the existing case and user feedback data learning statistical model, this paper presents a method of traditional Chinese medicine case representation based on key symptoms. On this basis, the paper designs a Chinese medicine case retrieval strategy combining machine learning technology, and improves the retrieval effect under the premise of guaranteeing the simplicity and flexibility of the model.**

Keywords-decision support system; case reasoning; case retrieval; machine learning

I. Introduction

Case representation is the first problem that needs to be solved in case-based reasoning. So far, there are many researches on knowledge representation in academia such as frame representation[2], first-order predicate logic representation[3], production rule[4], semantic network[5] and so on. In the aspect of case representation, this paper designs a case representation method based on the key symptoms which can help to improve the efficiency and accuracy of case retrieval.

Case retrieval is the key of case-based reasoning. There are regular case search methods such as nearest neighbor[6], inductive indexing[7], knowledge guidance, and template retrieval[8]. Whether the inference results are accurate largely is decided by the quality of the retrieved cases. In case retrieval, the strategy of case retrieval will be proposed. The method uses the prescription information in the case of traditional Chinese medicine, and improves the retrieval effect under the premise of ensuring the simplicity and flexibility of the model.

II. Case Representation

In the case-based reasoning system, the case representation is the first step. A case can be seen as the set of questions to be solved, a solution to the goal and the final answer. When building a case database, the presentation of the case must be designed on a case-by-case basis.

A. *Original Structure of Medical Records*

The medical cases cited in the article, are 40,000 old Chinese medicine cases provided by the China national 10th Five-year project.

TABLE I. THE ORIGINAL STRUCTURE OF MEDICAL RECORDS

Field	Table Column Head
Medical number	The only number of each medical case
Medical case title	Usually "doctor name + rule"
Summary of medical records	A brief introduction to medical records
Doctor name	
Patient name	
Patient sex	
Patient age	
Treatment time	
Patient complaints	The patient's own description of the symptoms
Pulse	Patient's pulse performance
symptom	That is commonly referred to as "symptoms"
Tongue coating	The performance of the patient's tongue
Dialectical analysis	The doctor's analysis of the patient's condition
Chinese medicine diagnosis	Doctors for the diagnosis of patient type
Governance is the rule of law	The role of doctors to be reached
Chinese medicine decoction	The name of the prescription
Chinese medicine prescription	The specific composition of the prescription

B. *Case Representation*

In this paper, according to the Chinese medicine description of the patient's specific performance, we then extract the more representative of the key symptoms, and replace the symptoms with the key symptoms to represent the patient's characteristics.

Based on the original structure of the medical case, we have selected some necessary information in this case, and the medical case is as follows:

- Key symptoms: According to the theory of traditional Chinese medicine diagnosis, the patient information which is a decisive role for the diagnosis in Chinese medicine cases can be expressed by a ternary structure (x, s, m). X is the set of patients' symptom, s represents the expression of the patient's tongue, m is the pulse of the patient, and the length is represented by the letter k.

- Decoction prescription: it is expressed by the binary structure (f,c), the set of decoction is represented by the letter f, in which the number of elements is no more than three. C represents a collection of drugs.

In summary, the case based on the key symptoms is represented as a ternary structure (z,f,c).

TABLE II. EXAMPLE FOR THE CASE EXPRESSION

Field	value
Case number	12989
Key symptoms 1	1
Key symptoms 2	0
...	...
Key symptoms k	1
Decoction	SiWu Soup
Prescription	Angelica

Splitting z into fields whose number is k, and each field has a value of 0 or 1, which represents whether the patient has this symptom. The meaning of each symptom field is determined by the prescription (each prescription corresponds to a different key symptom).

C. Case Building

According to the original medical case, the machine learning method is used to labeling, and then a case representation based on the key symptoms is constructed.

We define z for the key symptoms, the maximum value of $P(z\mid f)$. To select the key symptoms of the prescription f, we only need to calculate the probability of all symptoms under the conditions of use f, and then sort these probabilities from large to small, at last we can select the top k greatest probability of medical records.

In addition, Chinese medicine holds that there may be a link among the different symptoms. So we have to establish a relationship diagram among the symptoms.

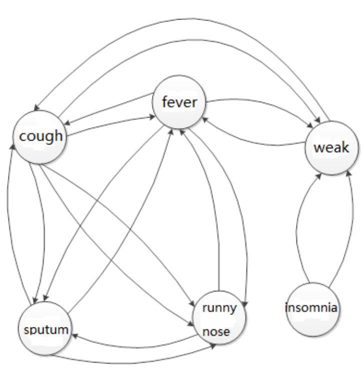

FIGURE I.SOME APPLICABLE SYMPTOMS OF XIAOYAOSAN.

In the figure, the node represents the symptom, and the weight of the edge is the conditional probability p from the node Z_i to the node Z_j. The calculation method of p is as follows:

$$p=\frac{count(Z_i,Z_j)}{count(Z_i)} \tag{1}$$

The numerator indicates the frequency of the occurrence of Z_i and Z_i, and the denominator indicates the total frequency of the occurrence of Z_i .

Finally, on the basis of this model, we use the classical $PageRank$ [9] algorithm to calculate the probability of each symptom. The critical symptom we need is the symptom of the top k greatest probability.

III. CASE RETRIEVAL BASED ON LEARNING RANKING

In case-based reasoning system, case retrieval[10] is absolutely indispensable. Case retrieval is the most relevant case of finding new problems in a case base by a certain method. Case retrieval is divided into the following three steps: feature recognition, initial matching, and best selection. This chapter revolves around these three steps. Based on the case representation method introduced in the previous chapter, a case retrieval method combining learning ranking is designed.

A. The Overall Pprocess

In this paper, a retrieval method based on template retrieval and nearest neighbor[11] is designed, and the prediction model is constructed by learning user behavior. The overall process is as follows:

- According to the characteristics of the new patient q_0, the relevant prescriptions are retrieved. The retrieved set of prescriptions is expressed as F_0;
- Remove the f from the retrieved set of prescriptions F_0, whose frequency is less than the threshold T (=3) removal, and get prescription collection $F_1=\{f_0,f_1\cdots f_N\}$;
- Calculate the probability that the prescription f_i applies to the patient q_0 ,which is recorded as $P(f_i)$, and the probability of not being applied is $P(\bar{f}_i)$.
- Calculate the probability of the case d_j, applying to the patient q_0, under the condition that the patient applies a certain prescription of f_i, which is recorded

as $P(d_j \approx q_0 | f_i)$ .Obviously the value of $P(d_j \approx q_0 | \bar{f}_i)$ is 0.

- Sort the relevant descending order according to the size of $P(d_j \approx q_0)$. The final result are the top k greatest probability, where:

$$\begin{aligned} &P(d_j \approx q_0) \\ &= P(d_j \approx q_0 | f_i) \bullet P(f_i) + P(d_j \approx q_0 | \bar{f}_i) \bullet P(\bar{f}_i) \\ &= P(d_j \approx q_0 | f_i) \bullet P(f_i) \end{aligned} \tag{2}$$

B. Feature Recognition and Initial Matching

In case based reasoning system, the index of key symptom agent should be built to ensure the retrieval efficiency. When searching, we first find the corresponding prescription according to the symptom (index), and then match the case containing the prescription according to the name of the prescription. Establishing inverted index between symptoms and prescriptions. An example of the structure is shown in Figure 2:

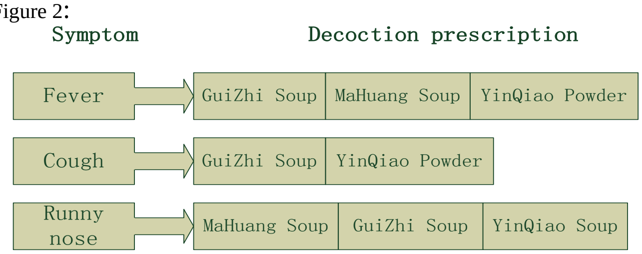

FIGURE II. EXAMPLE OF INDEX STRUCTURE OF “SYMPTOM-PRESCRIPTION”

When a new patient has symptom z_0, and symptom z_0 is the critical symptom of f_1 and f_2, then we need to list all the cases that contain f_1 or f_2, and calculate $P(f_i)$. $P(f_i)$ is calculated as follows:

$$P(f_i) = \frac{count(f_i)}{count(f)} \tag{3}$$

The denominator represents the number of elements in the list of retrieved prescriptions, and the numerator indicates the frequency of occurrence in the list.

The method in this section extracts critical symptoms, just excludes obvious, unimportant symptoms, other remaining symptoms are not all important symptoms. In order to make the matching more in place, we also need to base on this, and choose the top k best cases, complete the final matching of the case.

C. Final Matching Based on Learning Order

In order to complete the final matching, we introduce the Learning To Rank method (LTR) in the system to set parameters，so as to determine the order of the cases, that is to say, to calculate $P(d_j \approx q_0 | f_i)$

Our application is to achieve the goal of assisting physicians to prescribe drugs, that is, to obtain all cases that are similar to the threshold of the target case. Therefore, we adopt the idea of Pointwise [12]. But it raises two questions:

1) How to define the similarity between patients?

2) How to establish the training set of the LTR model?

The similarity among cases can be expressed directly by the similarity of the prescriptions. The prescription is the set of drugs, and the similarity between sets can be measured by the Jaccard coefficient, so the similarity of cases can be expressed as:

$$sim(X, Y) = \frac{\|X \cap Y\|}{\|X \cup Y\|} \tag{4}$$

X and Y represent the prescriptions for two cases, respectively.

In order to build a training set, each patient is required to construct a feature vector. The method of constructing eigenvectors is as follows:

$$t_i = \begin{cases} 0 \ , & a_i = b_i = 0 \\ 1 \ , & a_i = b_i = 1 \\ -1 \ , & a_i \neq b_i \end{cases} \tag{5}$$

Among them, a and b represent two patients who apply the same prescription, and 0 represents two patients don’t have this symptom neither, and the following two are the same. The characteristic vector of this pair of patients is t.

There are two ways the model is built:

1) Learning ranking, regression model, training set: The training set is setting up in the method described above, and the structure of the training set is shown in TABLE III. Training focused data is obtained by combining each two of all cases under the same prescription. The sorting problem is transformed into a regression problem, and $P(d_j \approx q_0 | f_i)$ is represented by the size of similarity. The decision process can be completed by any regression model, and the types of regression models are not limited.

TABLE III. EXAMPLE OF LEARNING RANKING, REGRESSION MODEL, TRAINING SETS

whether the critical symptom 1 is the same	**whether the critical symptom 1 is the same**	**…**	**whether the critical symptom 1 is the same**	**Similarit y**
1	-1	…	1	0.75
-1	-1	…	0	0.01
0	1	…	1	0.55

2) Training Set for Learning Sort Classification Model:

Setting up a train set as shown in TABLE III, where class labels are valued at 0 or 1, respectively representing uncorrelated or related labels. The data of train set is obtained by pairwise covering of all cases under the same prescription. On this basis, the classification model is trained, and the sorting problem is transformed into two element classification problem. The size of $P(d_j \approx q_0 | f_i)$ is used to represent $P(y=1| f_i)$. The class labels are calculated as follows:

$$y_i = \begin{cases} 0, j_i < \alpha \\ 1, j_i > \alpha \end{cases} \quad (6)$$

In the formula, y_i is an class label which represents a positive correlation between the case and the test case; j_i is an Jaccard similarity coefficient which represents a case detection and test case ; α represents a artificial similarity threshold; When the similarity of the detection case and test case is greater than α , that we think they are similar, When the similarity of the detection case and test case is less than α , that we think they are not similar.

TABLE IV. EXAMPLE OF TRAINING SET FOR LEARNING SORT CLASSIFICATION MODEL

whether the critical symptom 1 is the same	whether the critical symptom 1 is the same	…	whether the critical symptom 1 is the same	Similarity
1	-1		1	-1
1	1		1	1
-1	-1		-1	-1

IV. EXPERIMENT AND ANALYSIS

In this section we will experimentally verify the effectiveness of the method. The experimental steps are as follows:

A. Remove Useless Prescriptions

First of all, we standardize the medical cases based on prescriptions grouped, and remove the prescription whose frequency of occurrence is too small. Our experiment selects a total of 11458 standardized medical cases and a total of 285 kinds of prescription species. The frequency of occurrence of prescriptions is shown in TABLE V. We don't mark them in the chart one by one, due to the large number of prescriptions. In order to make the experimental data more reliable, we delete prescriptions whose frequency of occurrence is less than 50 each. At last 59 cases of prescriptions remain, and the total number of medical records used is 8296.

TABLE V. FREQUENCY STATISTICS OF PRESCRIPTIONS

Name of prescription	Frequency of occurrence
WuLing San	600
SiWu Soup	460
WuMei Wan	400
…	...
DaHuanglianzi Soup	30
HuangLianjiedu Soup	40

B. Evaluation of the Search Results

In this experiment, we use two case representation (that is, the original medical representation and the key symptoms), then we do horizontal comparison with the results of it. In addition, we compare the three sorting methods(that is, directly sort in $P(f_i)$, predict $P(d_j \approx q_0 | f_i)$ using the regression model then sort in $P(f_i) \bullet P(d_j \approx q_0 | f_i)$ and predict $P(d_j \approx q_0 | f_i)$ with the classification model then sort in $P(f_i) \bullet P(d_j \approx q_0 | f_i)$) of the results vertically.

The evaluation of the results is as follows: Calculate and compare the average similarity of the answer (prescription) part of the first 50 cases and the answer part of the test case. A high average similarity indicates better search results. We use the Jaccard similarity coefficient for the similarity measure of the answer:

$$jaccard(X,Y) = \frac{\|X \cap Y\|}{\|X \cup Y\|} \quad (7)$$

1) Regression model: To change the default, adjust the template as follows.

In the experiment, we use the Ridge Regression model. Ridge regression is a supplement to the least squares method. Combined with practical problems, we define the optimization of the model objectives L_θ as follows:

$$L_\theta = (J - T\theta)^T (J - T\theta) + \lambda \|\theta\|^2 \quad (8)$$

In the formula J represents the Jaccard similarity coefficient of the answer part of the retrieved case and the answer part of the test case; T is the characteristic of the test case , θ is the parameters to be optimized parameters in the model; $\lambda\|\theta\|^2$ is a regularization

2) Classification Model

The model of classification which we used in the experiment is Logical regression. The Logical Regression is not only to predict the "category", but also can get the approximate probability of prediction, which is useful to tasks that need probabilistic to make decision. Combined with practical problems, we defined the optimization of model as follows:

$$L_\theta=\prod_{i=1}^{n}\left(\frac{1}{1+e^{-\theta^T T_i}}\right)^{y_i}\left(1-\frac{1}{1+e^{-\theta^T T_i}}\right)^{1-y_i} \quad (9)$$

The n in the formula represents the size of the training set, the T_i is a feature representation of detected cases, the y_i represents the actual relevance of the detected cases and the test case. We take the threshold $\alpha=0.5$ in this experiment.

C. Experimental Verification

We divide the cases into 10 copies respectively, for cross-validation,, and calculate the Jaccard coefficient of 10 answers and the answer part of the test cases ,and take the average.

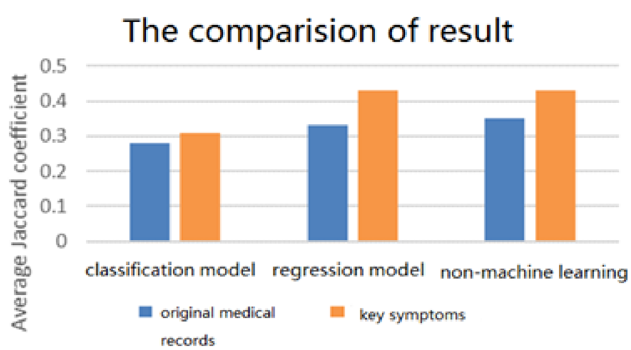

FIGURE III. CASE RETRIEVAL PRELIMINARY EXPERIMENTAL RESULTS

TABLE VI. CASE RETRIEVAL PRELIMINARY EXPERIMENTAL RESULTS STATISTICS

	Classification model	Classification model	No machine learning
Medical original representation	0.28	0.33	0.35
Key symptoms representation	0.31	0.43	0.43

Obviously, the key symptom representation is a better way for representation than that of the original representation of medical records. But there remains two questions: First, two kinds of machine learning with the case retrieval method in the experiment performance is not as direct retrieval, especially the use of classification model of the search method is very bad; Second, each of the average accuracy of these six cases is below 0.45, the overall poor performance.

The main reason for this phenomenon may be that the lack of training data due to the fact that the number of medical records corresponding to some prescriptions is small, so that the optimization of model parameters is not sufficient. So we select at least 200 prescriptions for all medical cases. According to the statistics, there are only six kinds of prescriptions as shown in TABLE Ⅶ.

TABLE VII. THE CORRESPONDING PRESCRIPTION STATISTICS MORE THAN 200 COPIES OF MEDICAL RECORDS

Prescription	Number of medical record
XiaoYao San	243
DiHuang Wan	236
SiJunzi Soup	210
LiuJunzi Soup	207
GuiZhi Soup	204
DaChaihu Soup	200

On the basis of these 6 prescriptions, we carry on the experiment again. The experimental results are shown in Figure 3.

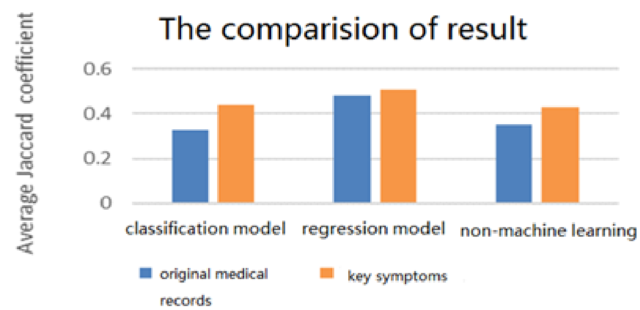

FIGURE IV. SECOND EXPERIMENTAL RESULTS OF CASE RETRIEVAL

TABLE VIII. SECOND EXPERIMENTAL RESULTS STATISTICS OF CASE RETRIEVAL

	Classification model	Classification model	No machine learning
Medical original representation	0.33	0.48	0.34
Key symptoms representation	0.44	0.51	0.45

In the case of sufficient training data, the performance of combined with the machine learning method of retrieval is indeed better than before. And the performance of no direct learning method of machine learning do not has much change.

However, it is not difficult to see that even in the case where the training data has been relatively adequate, the performance of the retrieval method combined with the classification model is still not ideal. In order to solve this problem, we also adjust the parameters in the model (that is, whether the two cases are related to the similarity threshold α), the parameters shown in Figure 4.

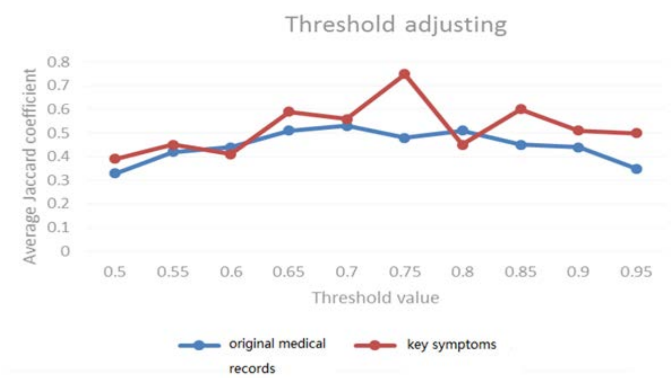

FIGURE V. ADJUST THE PARAMETERS

From the figure, when the value of α is 0.7, the original case representation achieves good results, the accuracy is 0.53; When the value of α is 0.75, the critical symptom representation achieves the best effect with an accuracy of 0.72.That is to say, compared with the accuracy of 0.33 and 0.44 in the previous experiments, the performance of the retrieval method combined with the classification model has been greatly improved by adjusting the parameters.

V. CONCLUSION

In connection with the present situation and the problems of TCM decision support system, the TCM case representation and retrieval were studied. In these two aspects, according to the particularity of Chinese medicine case, an improved method based on machine learning model is designed. The main work is as follows:

- In the case representation, the representation of traditional Chinese medicine cases based on key symptoms, compared to the original representation of medical records, is more helpful in improving the efficiency and accuracy of case retrieval.
- In case retrieval, a similarity comparison method based on machine learning is designed on the basis of case representation based on key symptoms. This method can help to improve the effectiveness of the search under the premise of guaranteeing the simplicity and flexibility of the model by means of the prescription information in the case of traditional Chinese medicine.

ACKNOWLEDGMENT

This work was partially supported by the National Key Research and Development Program of China under Grant 2017YFB1002304 and we would like to express our gratitude to Cao Tianwei for he helps us conduct experiments.

REFERENCES

[1] London S. DXplain: a Web-based diagnostic decision support system for medical students.[J]. Medical Reference Services Quarterly, 1998, 17(2):17-28.

[2] Guo Yanhong, Deng Guishi. A Case Study of Case-based Reasoning (CBR)[J]. Computer Engineering and Applications , 2004,40(21):1-5

[3] Mao Quan, Xiao Renbin, Zhou Ji. Research on Similar Case Retrieval Model Based on Case Feature in CBR[J]. Computer Research and Development ,1997(4):257-263.

[4] Ahn H, Kim K J. Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach[J]. Applied Soft Computing, 2009, 9(2):599-607.

[5] Uninghaus S B, Ashley K D. How Machine Learning can be beneficial for Textual CBR[J]. 1998.

[6] Xiao Shanshan. Research on Assembly Fixture Assembly Technology Based on Case-based Reasoning[D]. Changchun University of Science and Technology,2010

[7] Wu Hao. Research on Hypertension Diagnosis and Treatment System Based on Ontology and Case-based Reasoning[D]. Taiyuan University of Technology,2013

[8] Li Xiaozhan. Research and Application of Case - based Reasoning System for Medical Diagnosis and Treatment Based on Text Mining[D]. Guangdong University of Technology,2011

[9] Li Zhiying, Yang Wu, Xie Zhijun. Summary of Research on PageRank Algorithm[J].ChongQing: computer science,2011,38(10A):185-188

[10] Kang Z, Peng C, Cheng Q. Top-N Recommender System via Matrix Completion[C]// AAAI. 2016.

[11] Mccallum A, Nigam K, Rennie J, et al. A Machine Learning Approach to Building Domain-Specific Search Engines[C]// Sixteenth International Joint Conference on Artificial Intelligence. Morgan Kaufmann Publishers Inc. 1999:662--667.

[12] Learning to Rank[M]. Springer New York, 2014.