A Novel Diagnosis Method for Paediatric Common Diseases using Case-Based Reasoning

Lin ZHANG*, Deqing ZHANG

School of AnHui SanLian University, Institute of Computer Engineering, HeFei, AnHui, 230601, China

Abstract — **Severe challenges and difficulties are faced in paediatric medical care, where paediatric diseases do not receive timely diagnosis and treatment leading to unnecessarily prolonged pain for sufferers and anxiety and disruption to families. The paper aims at paediatric diseases diagnosis function not available currently and follows: i) the principle of dynamic evolution process of children's common diseases using Case-Based Reasoning method, ii) Inverse Document Frequency (IDF) concept of Information theory to design: a) the Case organization, b) Case retrieval and c) Case update in CBR technology to: iii) provide the diagnosis of children's with common diseases effectively.**

Keywords- Case-based reasoning; pediatric diseases; text frequency.

I. INTRODUCTION

Our country is facing the severe challenges of pediatric medical care difficult, pediatric disease cannot receive timely diagnosis and rescue, not only do the children suffer pain, parents are very worried too. It not only affects the normal life order of a family, but also influences the working order of the parents. As the "second-child" policy open, our country will usher in the new birth peak, therefore, research on children's common diseases diagnosis problem has the time urgency and important practical significance.

Case-Based Reasoning (CBR) method originated in 1982. Roger Schank, the professor of Yale University said in his book "Dynamic Memory" in 1982: It is a branch of artificial intelligence(AI), which is a kind of AI methods based on empirical knowledge (Case)[1-9]. CBR constructs and forms a rich body of case library, and uses the analogy understanding way which is based on case reasoning strategies and mimic human decision-making process in problem solving mechanisms to effectively solve the problems of unstructured and the lack of knowledge[10]

II. THE PEDIATRIC COMMON DISEASEDIAGNOSIS METHOD FRAMEWORK BASED ON CASE-BASED RESONING

The case-based reasoning technology is introduced in the process of the pediatric common disease diagnosis methods, and the framework based on Case-Based Reasoning is put forward, which is shown in Figure 1:

In the pediatric common disease diagnosis method based on case-based reasoning, firstly, we should give the case the standard description according to the pediatric common disease assessment index. And then, reflect the health of the pediatric in the form of feature vectorsry; finally, search the matched case which is close to the new issues in the case library. If we find the same case or similarity cases within the threshold range, the knowledge of the old case will be directly reused. Otherwise, assessment program will be revised according to the most similar cases to form the new case, which will be saved into the case library.

III. THE KEY TECHNOLOGIES OF PEDIATRIC COMMON DISEASE DIAGNOSIS METHOD BASED ON CASE-BASED RESONING

Case-based reasoning process includes four key technologies, which are case knowledge indication, case retrieval, case reuse / case modification and case study.

A. *Case Knowledge Indication*

With the social and economic development and the improvement of physical therapy, medical institutions and local communities at all levels of health services in our country have established a large number of pediatric common disease diagnosis records. Moreover, due to differences in data type itself, for example, some of them are numeric data, some of them are Boolean data, and some of the record type are text data, therefore, the data is difficult to compare on the same platform.

Here, we use feature vectors to represent the case of knowledge. Due to different types of data stored in a different way, and mostly unstructured data, we first establish a children's common diseases diagnosis table, and integrate the appearance of various kinds of children's common diseases and break it down for each option. For example, table 1 is a children's common diseases appearance case table, which data is researched from the community hospital in developing economic district in HeFei in AnHui province.

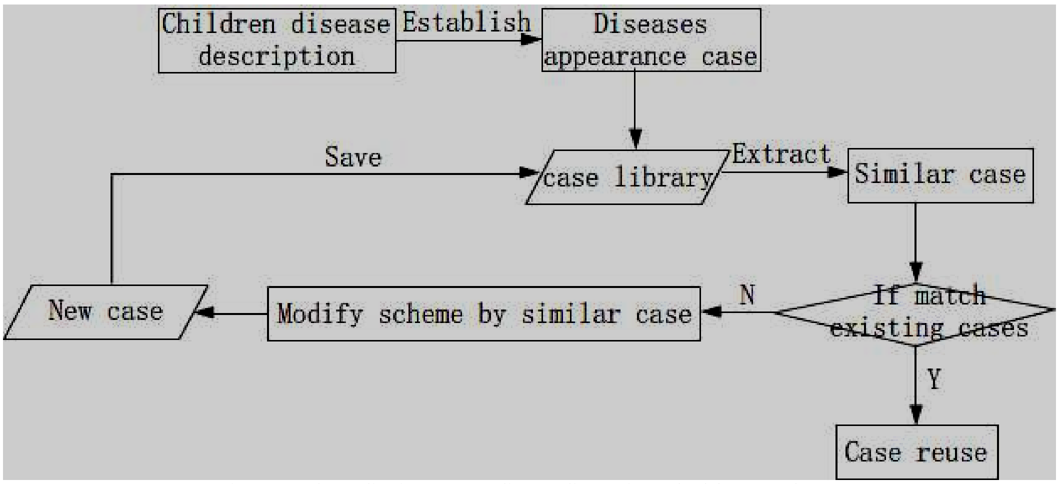

Figure 1. The pediatric common disease diagnosis method framework

TABLE 1. ONE COMMUNITY HOSPITAL'S CHILDREN'S COMMON DISEASES APPEARANCE TABLE

1.	(1) loss of appetite (2) Nausea and vomiting (3) abdominal distension (4) yellow and white tongue coating
2.	(1) fever (2) cough (3) throat have phlegmy
3.	(1)cough (2)Breathing (3) the lungs have noise
……	……
n.	(1) loss of appetite (2) Diarrhea (3) palms and soles hot (4) mental burnout

By integrating and breaking down the data in children's common diseases appearance table, we can establish children's common diseases statistics table shown in Table 2. In the statistics tabel, we will divide the appearance which is diversification (such as grey thick armour's colors) or has the appearance of the numerical value(such as the degree of fever) into multiple appearance content, which is shown as follows:

TABLE 2. CHILDREN'S COMMON DISEASES APPEARANCE STATISTICS TABLE

Appearance Number	Appearance content
1	Loss of appetite
2	Nausea and vomiting
……	……
6	Fever(37.5℃-38℃)
7	Fever(38℃-38.5℃)
8	Fever(>38.5℃)
……	……
86	Gray thick armor disease with gray color
87	Gray thick armor disease with yellow color
88	Gray thick armor disease with brown color
89	Gray thick armor disease with black color
……	……

Assuming that there are 357 appearances in children's common diseases appearance statistics table, and each appearance indicators can be represented by a 357-dimensional feature vector X, X=(x1, x2, ..., x357), in the vector X, if the i-th appearance indicators do not appear, then xi=0, otherwise xi=1.

It is easy to find that if the frequency of a appearance indicators' occurrences in all cases is lower (for example, Gray thick armor disease, the rash, etc.), which means the appearance indicators for the children's common diseases diagnosis is typical. Thus, when making the case retrieval, its weight should be bigger. Conversely, if the frequency of a health indicators' occurrences in all cases is higher(for example, cough, loss of appetite, etc.), and these health indicators is difficult to judge the pediatric disease situation, then its weight to make the case retrieval should be smaller. So all appearance indicators which appeared in the cases are set to 1 is unreasonable. This is similar to the information theory Inverse Document Frequency (IDF).

So-called IDF, in a nutshell, assuming that the number of a keyword w's occurrences in a web page is Dw, then the greater Dw is the smaller w's weight is. Vice versa, IDF=log (D/Dw), where D is the number of all pages [11].

Here, we can learn from the concept of IDF to set the right value of appearance indicators. Assume that the number of a appearance indicators i's occurrences is Di, then the weight of Di is log (D/Di), so D is the total number of the cases in the cases library.

Suppose that there were 100,000 cases in the case library, 50,000 cases among which have cough phenomenon, and 800 cases have skin rash phenomenon, namely D=100000, D117=50000, D118=800, then the weight for cough appearance index is log (D/ D117)=log(2)=1, and the weight for skin rash index is log(D/D118)=log(125)=6.96. So, we can get all appearance indicator's weight, and then we can establish the weighted children's common diseases appearance statistics table shown in Table3 as follows:

TABLE 3. WEIGHTED CHILDREN'S COMMON DISEASES APPEARANCE STATISTICS TABLE

Appearance number	Appearance content	Weight
1	Loss of appetite	0.94
2	Nausea and vomiting	1.32
……	……	……
6	Fever(37.5℃-38℃)	3.43
7	Fever(38℃-38.5℃)	2.98
……	……	……
86	Gray thick armor disease with gray color	6.34
87	Gray thick armor disease with yellow color	6.96
88	Gray thick armor disease with brown color	7.13
89	Gray thick armor disease with black color	7.02
……	……	……

Then we can obtain the appearance indicators weight vector $Y=(y1,y2,\cdots\cdots,y357)^T=(0.94,1.32,\cdots\cdots,1.72)^T$.

Thus, the knowledge of the case can be simply shown by the following 3-tuple: Case = (CA, FV, DT), in which, CA represents the property of classification, including the name, sex, age, the time of establish the case, the geographical area of life and so on, which be used to classify the case (by time, age or geographic classification, etc.); FV represents the appearance indicators' feature vector (the feature vector only have 0 and 1), while carrying case retrieval, we just need to do the matrix multiplication of feature vectors and appearance indicators' weight vector and constitute a weighted feature vector; DT represents health diagnosis and treatment recommendations, these diagnoses and recommendations can also be used to establish tables through the cases, and each diagnosis and recommendation can be shown by corresponding feature vector.

B. Case Retrieval

Case retrieval is the core of CBR, its purpose is to retrieve as few cases as possible which has reference significance to solve the current problems from a large number of cases. Common case retrieval strategy are mainly nearest neighbor strategy, induction indexing strategies, knowledge guide strategy and templates retrieval strategy. This paper uses the nearest neighbor strategy, but the similarity is calculated by the law of cosines, not by the Euclidean distance.

During the knowledge representation of the case, we have established the feature vector of appearance indicators for each case, so we can calculate the angle size between two vectors feature by the law of cosines. Since all weight of appearance indicators are positive, so the cosine between the two feature vectors is 0-1. If the cosine between the two feature vectors are closer to 1, namely the angle between two vectors is smaller, thus the two feature vectors indicated by appearance indicators is closer. Vice versa, the cosine between the two feature vectors are closer to 0, which means the larger the angle between the two feature vectors is, the smaller the velevance of two feature vectors indicated by appearance indicators is.

As we know the cosine of $\angle A$ is: $\cos A = \frac{b^2 + c^2 - a^2}{2bc}$

At this time, if b and c are two vectors starting from A, the above formula can be equivalent to: $\cos A = \frac{<b,c>}{|b|\cdot|c|}$, among which <b,c> represents the inner product of vectors, and |b| and |c| represents the length of vector.

Assuming that the appearance indicators feature vector of case X is (x1, x2, ..., x357), among which xi is 0 or 1, the weighted vector of health indicators is Y = (y1, y2, ..., y357) T, then the weighted feature vector is: (x1, x2, ..., x357) × (y1, y2, ..., y357) T = (x1y1, x2y2, ..., x357y357).

Thus, assuming that the weighted appearance indicators feature vector of two cases A and B are respectively (a1, a2, ..., a357) and (b1, b2, ..., b357), then the cosine of the angle between A and B is:

$$\cos\theta = \frac{a_1b_1 + a_2b_2 + \cdots + a_{357}b_{357}}{\sqrt{a_1^2 + a_2^2 + a_{357}^2}\cdot\sqrt{b_1^2 + b_2^2 + b_{357}^2}}$$

That means, the smaller the two vectors' value of Cos A is, the smaller the similarity degree of the vectors is, on the contrary, the larger the value of CosA is, the closer the two vectors are.

If CosA=1,then,the two vectors completely overlap , that is to say two cases of health indicators are completely the same.

C. Case reuse / Case modification

When a new case appeared, we only need to calculate the cosine of new case and each case in the case library, if a $\cos\theta = 1$, which shows that the current cases and new cases are identical, so the current case can be directly reused. Otherwise, we can descend the order of the case according to the value of $\cos\theta$, and screen the cases which is closer to the new cases, or threshold t, and screened out all cases of $\cos\theta \geq t$ as approximate cases. Because the children's common diseases diagnosis gets involve specialized knowledge. Therefore, it is very difficult to depend totally on the computer to automatically modify the case which requires manual intervention. However, if DT (diseases diagnosis and treatment recommendations) in 3-tuple which represents case knowledge is also represented by feature vectors, we can use the DT feature vectors which are screened approximate cases to do Boolean operations "and" first, and to determine the basic treatment recommendations, and then in order to reduce the degree of human intervention through manual intervention of expert.

D. Case Study

The results of a new case after a manual intervention was not certainly correct, you need practice to verify, just the new case which is proved to be correct can be added to the case library. When a new case is added to the case library, the weight vector needs to be adjusted, and each weight in weight vector is calculated by log (D/Di). While re-adjusting the weight vector, due to the addition of a new case, so D'=D+1; if the i-th appearance indicators appears in the new case, namely xi=1, then Di'=Di+1, otherwise Di'= Di, thus the new weight of i-th appearance indicators is log(D'/Di').

IV. METHOD TEST

In order to verify the accuracy of the children's common diseases diagnosis method based on case-based reasoning put forward in this paper, Visual Studio 2010 is used as the development platform, C# as the development language, and the cases are saved in the SQL Server 2000 to test the accuracy of the method in a simple way. In the process of test, 120 cases are used, including 100 training cases and 20 test cases. The 120 cases are all derived from the case library in which, all data are collected from the community hospital in developing economic district in HeFei in AnHui province,

In the test, firstly, according to the training cases and the corresponding evaluation standard, the SQL Server 2000 is used to establish a children's common diseases appearance statistics table which contains 357 attributes (each attribute for a appearance index) and 120 tuples (each tuple is a feature vector Xi) . The type of each appearance indicators in children's common diseases appearance statistics table is Boolean . After establishing the children's common diseases appearance statistics table, the former 100 tuple(training cases) are used as the foundation to calculate the weight of each appearance indicators in the table, and then a one-dimensional array containing 357 elements is established to hold the weight.

After the preparations all above, we use the later 20 tuple as the test cases, and calculate each vector Angle with the first 100 tuple respectively, and then screen out the cases which meets the threshold. The test shows that the case retrieval accuracy is more than 90%.

V. CONCLUSION

In this paper, case-based reasoning technology is applied to the pediatric common disease diagnosis method to come up with the pediatric common disease diagnosis framework based on case-based reasoning and to provide several key technologies in the reasoning process.

Pediatric common disease diagnosis method based on case-based reasoning in this paper is easy to understand and is relatively simple to achieve. Through experimental verification, the accuracy is also relatively high. However, the test just verifies the accuracy of the method. As for its effectiveness in practical system applications, it still needs to be verified, and the key technologies are still have some problems to be solved.

Firstly, all kinds of illness appearance were integrated and then they were broken down into each evaluated option. This method is simple and is very suitable for most options. However it is not suitable for some indicators with reference data (for example, body temperature, urinalysis performed values, etc.). In these cases, the applicability is not very good. We cannot evaluate the indicator through concrete numerical evaluation, but only through a numerical range. For instance, the body temperature can only be evaluated through not fever (36.5-37.2), low fever (37.5-38), fever (38-38.5), high fever (>38.5),but not through specific numerical 37 or 38.2 , so the assessment of the accuracy is not enough.

Secondly, when we express the knowledge by feature vector, because one index can be decomposed to multiple items, and only on item can be chosen in the same case (for example, low fever, fever and high fever are all decomposed by body temperature indicators, only one item can be chosen in a case, the other two cannot be selected at the same time), as a result, the feature vector is actually a sparse vector. In addition, the threshold which is mentioned in the case reuse technique needs to be set by the professionals, and it will undoubtedly increase the degree of human intervention. Therefore, in the practical application, how to make sure the effectiveness and simplify the existing algorithm at sparse matrix algorithms by the same time, and how to reduce the degree of human intervention, to improve its efficiency are all one of the direction of the study in the future.

ACKNOWLEDGEMENTS

This research is partially supported in the collection, analysis and interpretation of data by the College s and Universities in Anhui Province Natural Science Foundation of Key Projects with No. KJ2016A894.

REFERENCES

[1] Jiang Li-hong, Liu Bao, "On the Application of Case-based Reasoning to Intelligent Forecasting Support System", Journal of Decision Marking and Decision Support Systems, vol. 6, pp. 63-69, June 1996.

[2] Zhang Jian-hua, Liu Zhong-ying, "Case-based Reasoning and Rule-based Reasoning for Emergency Preparedness Information System", Journal of Tongji University, vol. 30, pp. 890-894, July 2002.

[3] Li Ru, Ren Hai-tao, Liu Kai-ying, Liang Ji-ye, "The Application of CBR in Agriculture Expert System", Computer Engineering and Application, vol. 25, pp. 196-198,204, January 2004.

[4] Li Ru, Ren Hai-tao, Liu Kai-ying, "The Study on Feature Weight Auto- learning Method for Case-based Reasoning", Journal of Shanxi University, vol. 27, pp. 245-248, March 2004.

[5] Li Jian-yang, Zheng Han-yuan, Liu Hui-ting, "Case-based Reasoning Based on Multi-layered Feedforward Neural Network", Computer Engineering, vol. 32, pp. 188-190, July 2006.

[6] Shen Ya-cheng, Shu Zhong-mei. "Research of Case Representation and System Architecture Based on CBR", Journa of Southern Medical University, vol. 27, pp. 1114-1116, July 2007.

[7] Gu Dong-xiao, Li Xing-guo, Research of Case-based Information System Business Process Knowledge Reuse", Journal of Chinese Computer Systems, vol. 28, pp. 1439-1443, August 2007.

[8] Peng Ming-de, Peng Tao, "Application of Case-based Resoning in TCM Case Record Distribution System", The World Science and Technology—Modernization of Traditional Chinese Medicine, vol. 11, pp. 698-701, May 2009.

[9] Li Ji-qiong, Li Xing-guo, Gu Dong-xiao, Feng Shuai, "Case Based Reasoning ISP Knowledge Reuse Method", Computer Engineering, vol. 36, pp. 36-39, January 2010.

[10] Han Min, Shen Li-hua, "Case-based Resoning Based on FCM and Neural Network", Control and Decision, vol. 27, pp. 1421-1424, September 2012.

[11] Wu Jun. Beauty of Mathematics, BeiJing: People's Posts and Telecommunications Publishing, 2012.