Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2013, 5(12):587-593

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

Ontology case retrieval based on SPA and asthma diagnosis application

Hongcan Yan, Huifang Wang, Shufen Zhang and Li Liu

Science College, Hebei United University, Tangshan, China

ABSTRACT

Case retrieval is a key process of Case-based Reasoning (CBR) system, and it is also one of hot research field of artificial intelligence. The calculation method of case similarity is the key technology of case retrieval. The set pair analysis(SPA) is applied to calculate the similarity of two sets between the target case and source case of case base, and the case retrieval model is built by the property and contact degree. The problem of multi-word synonyms is solved by the semantic extension of ontology, so the accuracy of the retrieval uncertainty is improved effectively. Experimental results on Asthma diagnosis show that ontology based on set pair analysis case retrieval model not only reduces the amount of computation, but also improve the recall and precision. All this provides technical basis for medicine diagnosis and treatment the next step.

Key words: Case-based Reasoning (CBR); Attribute Connection Degree; Similarity Calculation; Set Pair Analysis; Asthma diagnosis application; Chinese medicine Drug compatibility

---

INTRODUCTION

Case-based[1-2] Reasoning(CBR) is an important reasoning technology for problem solving and learning in the field of artificial intelligence in recent years, and it is an analogical reasoning model through accessing to the knowledge base for solving similar problems in the past, so as to get the current problem solution. Case-based reasoning generally goes through four stages which are case retrieval, case reuse and case modification and case storage. In order to solve the problem case, the first thing you need to search the cases similar to the given problem from the case base, then reuse the information and knowledge of retrieved cases to get the suggested solution, if the suggested solutions fail or are not satisfying, you need to modify it, and put the revised case as a new case in the case base.

Case retrieval is a critical step in Case-based reasoning, in the link of the similarity evaluation, the most commonly used method is by weighted Hamming distance[3] and Euclidean distance inverse function [4] to calculate the similarity of two cases. The similarity measure methods only consider the situation when the attribute is determined, or simply treat the fuzzy attributes with the deterministic attributes equally while not considering the difference and opposite factors between the cases. By this measure to calculate the similarity for cases search, case matching when dealing with complex problems, it is not only inefficient, but also it will affect the accurate judgment of acquired information, because it is difficult to achieve new knowledge reasoning according to the existing knowledge, and also can not to process the problems when the information is uncertain in the process of reasoning, thereby lowering the quality of decisions.

Set pair analysis[5] (SPA) is a system and mathematical analysis that making on the certainty and uncertainty and the interaction between certainty and uncertainty in the two sets of set pairs in a certain context, it studies the certainty and uncertainty of problems from three aspects which are the identity, discrepancy and contradistinction, and introduce the connection degree formula: μ=a +bi+cj to describe the uncertainty caused by fuzzy, random, and incomplete information untidily, then analyzes the uncertainty factors. Since the set pair analysis considers the different and opposite factors between the systems, making the judgment of access to information more objective, thus it is widely used [6-7].

In order to improve the efficiency of case retrieval and the decision-making quality, this article will use the set pair analysis theory and method in the assessment of similarity and build a case retrieval model based on set pair analysis, fully considering the uncertain information in the case. The application of medical cases of TCM Asthma on Case-based reasoning fully proves that this method makes the recall and precision of case retrieval improved effectively.

2. CALCULATION OF CONNECTION DEGREE FOR THE ATTRIBUTES OF CASES

Set pair analysis is a kind of system analysis method to deal with uncertainty problems, the core idea of SPA is as follows. First the set pair for two relative sets in an uncertainty system is constructed; then its properties are analyzed and calculated by means of the identity, discrepancy and contradistinction, namely I, D and C; finally the connection degree of the set pair can be established according to I, D and C. So the basis of SPA is set pair, and its key is connection degree.

Definition 1: Set pair connection degree: Analysis the characteristics of the set of H

according to the needs of problem W and get N characteristics which has identity in the s features, on the P a feature on the contrary, and in the rest of the F = N - s - P features, they are neither opposite nor identity, in other ways the nature are uncertain, then let the ratio:

S/N is the identity of the two sets under the question of W, which is short for identity degree;

F/N is the discrepancy of the two sets under the question of W, which is short for discrepancy degree;

F/N is the contradistinction of the two sets under the question of W, which is short for contradistinction degree;

Using the formula $u(w) = \frac{S}{N} + \frac{F}{N}i + \frac{P}{N}j$ to express the connection degree of set pairs H, and i is the uncertainty coefficient of discrepancy, which has different values in [−1,1] in different conditions or sometimes may be considered as a marker of discrepancy only; j is the uncertainty coefficient of contradictory, which has value of −1 or sometimes maybe considered as a marker of contradictory only. So $u(w) = \frac{S}{N} + \frac{F}{N}i + \frac{P}{N}j$ can be rewritten as:

$$u = a + bi + cj \qquad (1)$$

where a+b+c=1。

In order to introduce set pair analysis theory into case retrieval, we give the following definitions:

Definition 2: Case set pair: The certain mapping relationship between question case q and the each case in the case base constitute case set pairs.

And use the formula (q,p) to express the case set pair constituted by question case q and some case q.

Definition 3: Attribute set pair of case set pair: the attribute value between question case q and case q about the same attribute constitutes the attribute set pair of case set pair.

For example, assuming that there are n attributes related to the case pair (q,p), respectively: $x_1, x_2, \ldots x_n$, the attribute value of case q and case p on the n attributes is respectively: $x_{q1}, x_{q2}, \ldots x_{qn}$ and $x_{p1}, x_{p2}, \ldots x_{pn}$, then $(x_{q1}, x_{p1}), (x_{q2}, x_{p2}) \ldots (x_{qn}, x_{pn})$ are all attribute pair of case pair (p,q).

Definition 4: Case connection number of attributes: the expression of connection degree between each pair of attributes.

Use the expression $u_l = a_l + b_l i_l + c_l j_l$ to express the case connection number of Case pair（q，p）in the attribute of number l, for there is only one attribute value, so there only one item exists in the expression, for example ,the two case are identical in the attribute of number l, the attribute connection number are written: $u_l = a_l$ and $a_l = 1$, if they are discrepant, then written: $u_l = b_l i_l$ and $b_l = 1$.

Then, when calculating the attributes connection number of cases with the absence of attribute values, we consider they are discrepant according to the theory of set pair analysis, thus the problem of uncertainty information in the reasoning process is disposed effectively.

In this paper, we take medical cases of TCM Asthma for empirical research, and choose the attribute characteristics of case from the case base, assuming that the case is associated with n attributes: $x_1, x_2, \ldots x_n$, then compare the n attributes of question case q with that of each case q in the case base to determine each attribute set pair number of case pair (p,q):

$$u_l = a_l + b_l i_l + c_l j_l \quad , \quad i_l \in [-1,1], \quad j_l = -1, \quad l = 1,2 \ldots n \qquad (2)$$

Where a_l express the identity degree between question case q and case q on the attribute of number l; b_l express the discrepancy between them, c_l express the contradistinction degree between them and $a_l + b_l + c_l = 1$, so the set pair connection numbers between the question case q and case p on the n attributes are $u_1, u_2, \ldots u_n$ respectively, and this is very important part of the case similarity calculation.

3 BUILDING THE RETRIEVAL MODEL OF ONTOLOGY CASE

3.1 THE STRUCTURAL STORAGE OF CASE

At present, most of the CBR systems use the static frame to describe the case, and there exists many shortcomings such as [1,7,8] : difficult to extend, have disadvantages in reconfigurable and learning, there is difficult to extend, reconfigurable and disadvantages such as difference of learning. But ontology [9]as a tool of knowledge modeling can describe concepts in semantic and knowledge level, it expresses the relationship between concepts within the related territory and determines the concept of mutual recognition, and it can be reused, shared and also can semantic extensible, so it has a very good application in the field of knowledge expression[10-12].Although in a CBR system, the previous experience sets(cases) are the main source of knowledge, but in practice, we integrate the specific knowledge of case into the general domain knowledge mode (expressed in ontology), and some scholars have tried[13-15].In order to realize the semantic extension of keywords of in the link of case retrieval and improve the retrieval recall of case, medical cases in this article will be stored by the means of ontology structure. To calculate the contact degree by weighted average method to reduce the influence of weight, but still cannot remove the disadvantages of artificial custom weight coefficient, in order to reduce human preference, this article will case attributes in ontology knowledge base level as a reference, reflects the important degree of attribute in a layer of than the next layer of attribute importance.

In the reference [7], it calculates the connection degree by weighted average method to reduce the influence of

weight, but it still cannot remove the disadvantages which weight coefficient is made artificially, in order to reduce human preference, this article will take the level of case attributes in ontology knowledge base as a reference, the attribute in a layer is more important than the attribute in next layer.

3.2 CALCULATION OF CASE SIMILARITY

Assume the weight of the attributes $x_1, x_2, \ldots x_n$ related to case pair （q，p）is $w_1, w_2, \ldots w_n$ respectively, combine formula (1) and (2),we can get the case similarity of case pair （q，p）,it can be written:

$$Sim(p,q) = \sum_{l=1}^{n} a_l w_l + \sum_{l=1}^{n} b_l w_l i + \sum_{l=1}^{n} c_l w_l j \quad i \in [-1,1] \quad j = -1, \quad l = 1,2\ldots n$$
$$= A + Bi + Cj \qquad (3)$$

where $A = \sum_{l=1}^{n} a_l w_l$, $B = \sum_{l=1}^{n} b_l w_l i$, $C = \sum_{l=1}^{n} c_l w_l$

Given the right value i , we can calculate the value of case similarity between question case and each case in the case base, then store these cases in the target case base according to the similarity value sequence, the case whose similarity value is greater than the predetermined threshold of case is the goal we are looking for case study.

In some cases, especially in the medical cases, for the fuzzy description of ancient medicine basis, the differences of human understanding greatly affect the certainty of the data, so the subjective understanding of similarity should be considered. The general principle is: for identity degree A, the bigger, and the better; for discrepancy C and contradistinction degree B, the lower, the better.

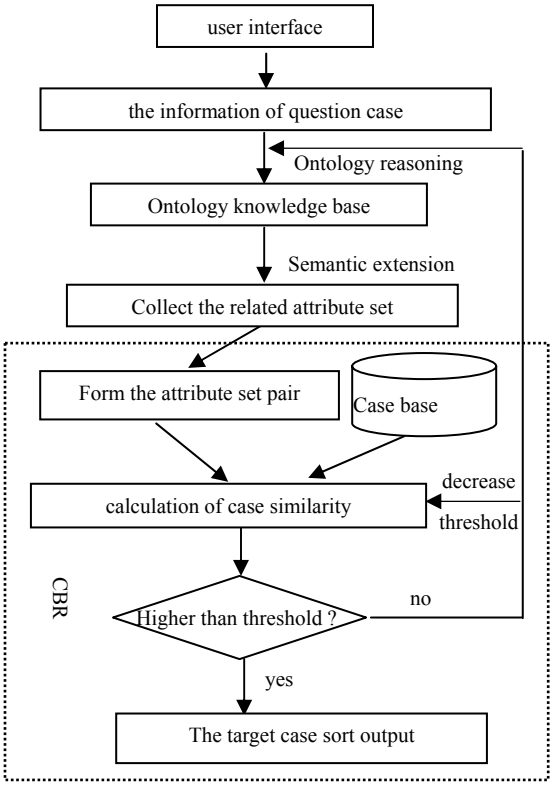

Figure1 illustrative diagram of case retrieval system

3.3 THE SYSTEM ARCHITECTURE OF CASE RETRIEVAL MODEL

Figure 1 shows the extension in the case-based reasoning retrieval words, tectonic attribute sets, the calculation of attribute connection degree, the process of case similarity calculation, the retrieve steps as follows: the system extract the attribute sets related to the question case after the expansion of righteousness or synonym by ontology knowledge base according to the information that the users put in, and then construct the attribute set pair of case pair, using formula (3) to calculate the case similarity.

Step 1: The system extracts the key words of background according to the information of question case q ;

Step 2: Through the expansion of righteousness or synonym by the ontology knowledge base, we can get the attribute set related to the question case.

Step 3: Use formula (2) and (3) on each case p in the case base to calculate the similarity of case pair (q,p), if the value is greater than the setting threshold, then put them in the target case base.

Step 4: If the target case base is null, we can decrease the threshold, repeat step 3,when the threshold is small to a certain degree ,there is still no target case, we can extend the attribute sets by using the ontology reasoning, then step 2;if not ,step 5;

Step 5: Rearrange the target case base according to the similarity value from big to small.

4. APPLICATION EXAMPLES AND ANALYSIS

4.1 THE ONTOLOGY FRAGMENT OF CASE AND SIMILARITY CALCULATION

This article combed more than 600 records from the medical cases of TCM Asthma, then extracted relevant attributes, use the Protégé 4.1.0 editor to build the fragment of ontology as shows in figure 2:

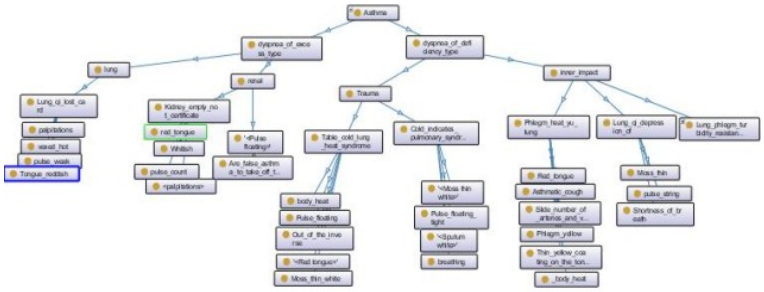

figure 2 Chinese medicine asthma ontology fragment

Data shown in table 1 are part attributes of 12 typical cases selected from case base, including 2 cases as question cases, accompanied by attribute weights obtained by the recommendation of traditional Chinese medicine experts and the reference of ontology structure.

Table 1 cases of TCM Asthma and the attribute weights

cases	asthma	taxiang（color）	hanre	sheti（color）	shetai（color）	shetai（thickness）	maixiang
1	chuancu	tanhuang	zhuangre		taihuang	bo	huashuo
2	chuan		weire	shedan	taibai		fushuo
3	chuanxi	tanhuang	weire		taihong	luehou	xianhua
4	qichuan	wu	ehan	shedan		hou	fushuo
5	qiduan			shehong	taibai		chenruo
6	chuansou	tanhuang	chaore	shehuang		bo	fujin
7	kechuan	tanhuang		shedan	taibai		xihua
8	chuancu	tanbai		shedanqing	taibai	bo	xihua
9	weichuan	tanbai	ehan	shehong	taihong		chenshi
10	chuan	tanbai	weire	shedan	taibai	bo	
question case 1	kechuan	tanhuang	zhuangre		taibai	bo	hua
question case 2	qiduan	tanbai	ehan	shehuang	taihong	hou	fu
attribute weight	0.4	0.15	0.05	0.1	0.05	0.1	0.15

In order to verify that the proposed calculation method of case similarity is of high efficiency, especially the impact of calculation of case similarity when the attribute information absent, respectively using the method of hamming distance and the similarity calculation based on set pair analysis on the question case 1 and question case 2. Set the threshold to $\lambda = 0.8$,the calculation results follows as table 2:

Table 2 experimental results

Source cases	question case 1			question case 2	
number	SPA	Hamming distance		SPA	Hamming distance
1	0.88	0.85		0.25	0.34
2	0.63	0.59		0.62	0.78
3	0.66	0.64		0.7	0.82
4	0.52	0.48		0.69	0.72
5	0.62	0.41		0.46	0.74
6	0.86	0.84		0.68	0.58
7	0.84	0.76		0.6	0.70
8	0.82	0.81		0.5	0.52
9	0.51	0.52		0.46	0.58
10	0.63	0.65		0.42	0.44

RESULTS

It can be seen from table 2, the method of calculation of case similarity based on set pair analysis method has a better degree of differentiation, for the question case 1,there are 4 cases whose value is greater than the threshold is 4 based on method of set pair analysis while there are only 3 on the method of hamming distance, observing the data in table 1,obviously case 7 is closer to the question 1, which says that the method based on set pair analysis is closer to the truth; for the question case 2, there are 4 cases whose value is greater than the threshold is 0 based on method of set pair analysis while there are 1 on the method of hamming distance, observing the data in table 1, obviously case 7 is not too close to the question case 2,if we take the retrieved case based on the method of hamming distance, it is possible to make a wrong diagnosis.

So we can draw the conclusion that the method based on the set pair analysis is more efficient than the traditional method based on hamming distance.

CONCLUSION

The advantages of case-based reasoning technology is mainly manifested that the complete domain knowledge which isn't needed, and also a lot of complete data isn't needed.

Only the specific case in the past experience is required to solve new problems. It has the function of self-learning. The set pair analysis was applied to the similarity evaluation basing on case-based reasoning, and a new similarity calculation method was proposed. The semantic extension of retrieve properties was achieved, and the case retrieval efficiency was improved through the application in the asthma medical cases of traditional Chinese medicine. Based on set pair analysis in the case of ontology similarity calculation method the advantages are as follows.

(1) The application of ontology effectively handled the problems of polysemy, different words with the same meaning, which has improved the recall ratio of retrieval.

(2) Different or even opposite factors between the systems are taken into consideration in the set pair analysis, which has made the judgment on access to information become more objective and more comprehensive, also has improved the efficiency of retrieval algorithm.

Our focus next will be taking the ancient medical of asthma disease as the research object, in order to explore the medication rule of ancient drug compatibility, and the ontology rule reasoning will be implemented by using the

Jena reasoning engine, which will further improve the recall ratio of case retrieval. It will provide a theoretical basis and technical support for the further research and application of case-based reasoning and the semantic web.

Acknowledgements
This paper was funded by the National Natural Science Fund and Hebei Province Natural Science Fund Project (project number: 61370168, F2014209238).

REFERENCES

[1] WANG Dong,LIU Huailiag,YU Haibin. *Computer Engineering*,**2003**(7):10-12
[2] Cheng- Hang Liu, Long- Sheng Chen , Chun - Xin Hu. *Information Sciences*, **2008** :3347~ 3355.
[3] Eva Armengol, Enric Plaza. *Artificial Intelligence Research and Development*,**2005**,131:239-246.
[4]Yin-Shan Gu，Qiang Hua，Yan Zhan ．Case-base maintenance based on representative selection for 1-NN algorithm[C]．In：Machine Learning and Cybernetics，2003 International Conference on，**2003**：242l～2425
[5] ZHAO Keqin. *Information and Control*. **1995**（6）:16-19
[6] WANG Wensheng,XANG Honglian,DING Jing. *Technical Methods*: 320-323
[7] RUAN Guangce. *Journal Of The China Society For Scientific And Technical Information*, **2012**（10）：1090-1095
[8] LI Fenggang,NI Zhiwei. *Application Research of Computers*, **2010** （2）: 544-547.
[9] Studer R，Benjamins V R，Fensel D． *Data and Knowledge Engineering*，**1998**，25(1-2)：161—197．
[10] JIANG Hongchao,WANG Daliang,ZHANG Dezheng. *Computer Engineering*, **2008**(6):16-19
[11] YAN Hongcan,LI Minqiang,REN Yunli. *Journal Of Tianjin University*, **2009**（5）:272-276
[12] HUANG Fenghua,YAN Luming. *Journal of Computer Applications*, **2013**(3):771-775
[13] XU Guichen.Research on Medical Case-Based Reasoning Based on Ontology[D].Zhejiang University, **2011**（12）.
[14] WANG Haitang,GU Junzhong,YANG Jing. *Computer Era* No.1 ,**2009**:58-60
[15] XIE Hongwei,LI Jianwei. *Application Research of Computers*, **2009**（4）:1422-1424