Case Retrieval Algorithm Based on Structure Matrix

Lichcuan Gu*, Qingyan Guo, Mengru Cao, and Dengliang Zhang
School of Information and Computer, Anhui Agricultural University, Hefei 230036, Anhui, China
Email: gulichuan@hfut.edu.cn, qyanguo@hfut.edu.cn; mengrucao@ahau.edu.cn; denglz@163.com
*Corresponding author

***Abstract*—Case retrieval is the focal stage of Case-Based Reasoning systems whose quality is determined by the speed and accuracy of retrieva1. In this work, we aim at developing a better Case retrieval algorithm by using vector model and propose its new case retrieval Algorithm based on Structure Matrix, wich is derived to learn the kernel matrix for capturing the relations between the case structure units based on matrix iterative analysis. The experimental comparison of similarity shows that using of the structural information of the case, the accuracy of their experimental results has a general increase, when opposed to some of the existing similarity measure. The same learning algorithm based on matrix iteration method compared to other methods, have higher accuracy (5% to 8%), and required less training documents, the computational cost is smaller.**

***Index Terms*— Case-Based Reasoning; Similarity Measure; Matrix Learning; Case Retrieval**

I. INTRODUCTION

A case-based Reasoning (CBR) system remembers previous experiences or episodes called cases and use them to assist in obtaining a solution to a current problem. The premise of case-based reasoning is that once a problem has been solved, it is often more efficient to solve the next similar problem by starting from the known solution rather than by repeating all the reasoning that was necessary the first time [1] [2].

Traditional CBR consists of four steps as follows: retrieve the most similar cases, reuse existing knowledge of previous cases to solve new problem, revise suggested solutions and retain useful parts of this experience for future problem solving. Among them, the case retrieval is the key activity in CBR cycle, because the most similar case is suggested in the activity and also it is the prerequisite of case adaptation. Furthermore, the remaining operations of adaptation and evaluation will succeed only if the retrieved cases are relevant [2] [3]. The retrieval of relevant cases is closely related to, and dependent upon, the Similarity measuring. After a similarity evaluation of cases, the system will get a preliminary list of the cases being most similar to the new problem among the case base. This list of cases is arranged by descending similarity scores. If the similarity score of a past case is under the threshold, the case will be eliminated from the list. Then, the system or user can decide which case is the most similar and best by further analysis.

For rapid retrieval, the memory structure of case base often can be categorized into (1) associative retrieval and (2) hierarchical retrieval, and (3) the hybrid model. In associative retrieval, each attribute is independent of others. The approach is suitable for multiple retrieval but takes longer time [1] [3]. The attributes of the hierarchical method are organized in a conceptual structure. Due to this kind of memory structure of the case base, the time of retrieval can be reduced enormously. The only shortcoming is failing to retrieve extensive cases. Hybrid retrieval is a combination of associative and hierarchical retrieval. The problem of case retrieval in CBR has been the major subject of much research since 1966. Researchers have developed several case indexing approaches for use in the case retrieval stage. Commonly used similarity measuring measures include nearest neighbor algorithm, Tversky contrast matching function, improved Tversky matching method, and the distance metric method or nearest neighbor algorithm, multi-parameter similarity calculations, Weber, calculation method, the local similarity, object-oriented case. Similarity measuring method based on fuzzy set similarity, etc. [2] [3] [4]. Although these methods are unique, innovative but their actual effects are very limited, the accuracy is not ideal, high computational cost and the database of solved cases are large.

Many case-based reasoning systems need a large database of cases for coverage of a wide variety of problem instances. Typically, large case libraries are necessary for good problem coverage and quality solutions. But large databases may cause degradation in efficiency, especially if the case matching function to determine similarity is computationally expensive. A lot of CBR research is devoted to the organization and indexing of the case library to make case retrieval as quick and accurate as possible. [5] [6]

Hence, On the basis of the expression, storage and structure of the analyzed cases with similar characteristics with the vector space, we propose a new similarity measure based on vector model that considers structure of case and support retrieval process from case base. An experiment is conducted on Case base of the pear black heart disease prediction system developed by the Lichuan Gu as a data source. The experiments show that the

doi:10.4304/jmm.8.6.712-719

retrieval algorithm proposed in this paper can effectively capture the relation of case structure units, significantly improve the accuracy of similarity measure. Further experimental analysis showed that compared with other learning methods, retrieval approach based on structural matrix is potentially useful in many case-based reasoning systems, especially those with computationally expensive case matching and large case libraries.

II. Related Work

A. CBR Retrieval

The ultimate goal of the CBR system is to find the appropriate conclusion set C for each new case. The most important work in the CBR system is to comparing the characteristics between relatively new cases and case base. Cognitive psychology studies have shown that human experts face solving the problems by using of the experience and knowledge of past memories in their own minds, memories handled with similar problems, to make appropriate changes to handle new situations. Case-Based Reasoning is the human analog of this analogical reasoning thinking, the reasoning process has some characteristics of the human experience and reasoning. In solution based on the case, the problem dealt with in the past, is described as cases consisting of the problem characteristics set, solving program elements, stored in the system's case base, known as the source case. Treat currently facing problems or situations as the target case [1].

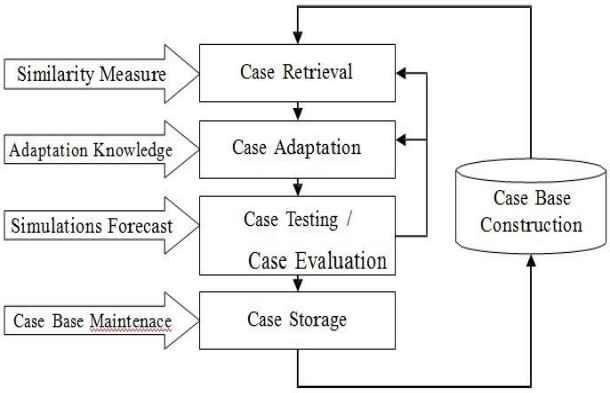

Figure 1. System overview of CBR process

A CBR approach to solve a new problem typically involves 4 processes. Fig. 1 illustrates the case-based reasoning process [1].

1. Case matching and retrieval: The new problem is compared to the library of past cases, and the most similar case (or cases) is retrieved. A set of relevant problem descriptors needs to be defined to match cases. A similarity measure is typically used to compare cases, and some sort of similarity threshold is needed to select the best cases.

2. Case reuse: The information and knowledge in cases retrieved are used to solve the new problem.

3. Case revision: If necessary, the solution is revised and adapted. Revising the solution generated by the reuse process is necessary when the solution proves incorrect. This also provides an opportunity to learn from failures.

4. Case retention: The new problem and its solution are (optionally) retained as a new case, depending on how useful it is expected to be in solving future problems. This step involves deciding what information to retain, how to retain it, and how it should be indexed for future retrieval.

Among the 4 steps, the main component of CBR is case retrieval. Case retrieval is the process of identifying or retrieving previous cases that are similar to the problem case and can be adapted to provide a solution to the problem case, or in general assist the CBR system in achieving its goals. This process starts with the problem case or problem description. Before past cases can be retrieved, situation assessment must be completed on the current problem to determine the entire context of the problem in a vocabulary that the CBR system can understand. In this process the current problem case must be" flushed out", or in other words all available information must be extracted in order to totally quantify the problem. The problem case is the starting point for situation assessment, but other knowledge acquisition techniques such as the use of assumptions, interpolation and guided discovery can assist in complete situation assessment. For example, assuming that the ambient temperature for a process is similar to the ambient temperature recorded at a local weather station could be a good assumption. Depending on the nature of the case-base, situation assessment could introduce additional complexity to the CBR process. For example, a conversational CBR system could require additional initial development in the form of a natural language interpreter to convert dialogue to text and then to a format meaningful to the CBR system. In a structural CBR approach the attributes are represented in a simple feature value matrix that is generally easy to use, manipulate and maintain [6].

In essence, a good assessing similarity between cases is a key success of CBR. In the mean time, this retrieval process is directly related to the structure of knowledge base. Thus, both case retrieval process and knowledge base construction must be designed to accord.

Thus, we propose the use of an approximate, inexpensive, feature-based distance metric (like Euclidean distance) to filter a small number, say k, of potential matches, using the nearest neighbor rule [7]–a more accurate matching function is then used to do the final ranking. The feature-based similarity metric is expected to approximate a correct, objective, and usually expensive matching method. By weighting the features with SVM, the accuracy of the similarity metric is improved, and hence fewer potential matches needs to be filtered to ensure retrieval of true matches.

B. Sector Space Model

The vector space model (vector space model and referred to the VSM) is a common document model. Words construct a high-dimensional space, each word in the space is one dimension, and the document is seen as a vector in this space [8].

$$x=dx(1), dx(2), \ldots, dx(n)T \quad (1)$$

In the vector space model, the text refers to a variety of machine-readable records. D (Document) feature (Term, t) is referred as basic linguistic unit which appear in the document D, can represent that the content of the document, constituted by words or phrases. The text can use the feature set to represent as d (T_1, T_2... T_n), of which Tk is the feature, 1 <= for k <= N. For example, if a, b, c, d are four characteristics of a document, this document can be expressed as d (a, b, c, d).

Text with n features, usually give each feature items some certain weight that shows importance. D = D (T_1 W_1; T_2, W_2... T_n, W_n), denoted by D = D (W_1, W_2... W_n), we called it D vector of the text. Where Wk is the weight of the T_k, 1 <= for k <= N. In the above example, suppose a, b, c, d weights were 30, 20, 20, 10, then the text of the vector is D (30, 20, 20, 10).

The case is a special form of knowledge representation, is the basis and premise of the case-based reasoning, and its representation determines the conversion of real-world problems to the case, and have a great impact on the efficiency of reasoning. Usually the case is available to be described by the diverse style below:

$$ASE=<I, A, C, J, S> \quad (2)$$

where: I={I_1, I_2, ⋯, I_m} is a finite set, indicating the basic information of the case; A={A_1, A_2, ⋯, A_n} is a finite on-empty, indicating the case has a variety of characteristics, thus, feature sets; C={C_1,C_2,⋯,Co}is a finite non-empty, indicating that set of conclusions aroused from the feature set A, judging set J is a text description, usually written by the analysis's, to illustrate the adequacy of the conclusions set C; S={S_1,S_2, ⋯ ,Sq} is a finite set, indicating that the conclusion set C solution. When there are some similarities (structure, grammar, etc.) between the source case and target case, the reasoning process depends on this similarity. The basic process of case-based reasoning is: When you encounter a new problem, the system retrieve in original case base the according to the key features, identify one or a group closest to the unknown candidate case, reuse this candidate solution of the case. If you are not satisfied with the solution to the case of this candidate case, this case can be modified to adapt to the problem. Last modified case will be saved as a new case in the library so that when next encountering similar problems they will serve as reference.

In the vector space model, the two text content of S_{im} (D_1, D_2), between D_1 and D_2, is commonly indicated by using the cosine of the angle between vectors.

The formula is:

$$Sim(D_x, D_y) = \cos\theta \frac{\sum_{i=1}^{n} \varpi_{xi} \times \varpi_{yi}}{\sqrt{(\sum_{i=1}^{n} \varpi_{xi}^2)(\sum_{i=1}^{n} \varpi_{yi}^2)}} \quad (3)$$

Among them, respectively, ϖ_{xi} , ϖ_{yi} indicate the features weights of text Dx and Dy K, 1<=i<=n.

During automatically classification, usually adopt a similar approach to calculate the correlation of documents and certain categories to be classified. Such as characteristics of the text Dx is Dx1, Dx2, Dx3, Dx4, weights were 30, 20, 20, 10, the characteristics of category C1, is a, c, d, e, weights were 40, 30, 20, 10, the vector Dx is(30, 20, 20, 10, 0) C1 vector should be expressed as C1 (40, 0, 30, 20, 10), calculated according to the equation, the Relevance of text Dx and category C1 is 0.86.

The most widely used method is the k-NN algorithm [6] in the case-based reasoning technology. It assumes that all cases correspond to the points in the n-dimensional space Rn. A close neighbor of a case is defined according to the standard Euclid distance, thus regarding any case as the feature vector. For case expression model from for multi-type (2), the feature set can be used as a retrieval index. Therefore, the VSM model can be used as reference, in the case retrieval of case-based reasoning system, using the following method to measure the similarity between the two cases:

$$Sim(d_i,d_j) = \sum_{l=1}^{n} d_{i(l)}^{T} \bullet M \bullet d_{j(l)} \quad (4)$$

where: n is the number of different characteristics in the case; dx, dy, respectively are the Case CASEx, CASEy 's matrix in the case base space. To eliminate the difference of the number of characteristics in each case, for the matrix, united each column vector (vector of the same structural unit), that is $\sum d_{x(l,k)}^2 = 1 \bullet M$, M is an $M \times M$ matrix, where is used to describe the relevance and weights of the similarity between cases. In this paper, we called it structure matrix.

Case retrieval is to find one or more target case similar to the source case, which is operated on the base of comparisons the similarity measure is a key step of case retrieval. The similarity between the cases is defined according to the similarity between the properties. The similarity between the target case and source cases are divided into four types: semantic similarity, structural similarity, objectives similarity and individual similarity [5].

The two cases are analogous, firstly, the semantic similarity should be meeting, and structural similarity helps the initial search to Retrieval the analog source case. Problems solving is always to achieve a certain goal, a greater role in the source case should be given priority to the target to achieve the target case. After the initial search, it is necessary to first consider the case with the goal of individual similar or containment relationship of the source case, the analogy of the individual, so that we can take advantage of some information to solve the overall problems [6]. As the correlation between the cases of the case base, so the feature space constructed by the case base of case-based reasoning system is not orthogonal space. In this paper, matrix to describe this relationship between the cases, put forward a way on the basis of mathematical transforming to analyze and obtain

this relationship by the learning algorithm, and give specific matrix iteratively earning algorithm.

III. A TWO-PHASE CASE RETRIEVAL METHOND

Consider a case-based reasoning system with a database consisting of N cases, and a set of query instances Q. we here assume that N and Q are drawn from the same distribution of instances, each of which is represented by a set $\hat{W}$ of numerical features. Our Case Retrieval algorithms can be split into two phases. The first step is to remove redundant features of Cases, and the second is similarity measuring.

A. Attribute Reduction of CBR

Using SVR method to streamline case-based reasoning system and extract rules and knowledge in new reasoning systems. This treatment is aimed at minimize training time, simultaneously ensuring the classification accuracy, and ultimately improve the difficulty and speed of knowledge discovery in reasoning systems, following these steps.

For case-based reasoning system corresponding to the case base set (x_1, d_1),…, (x_n, d_n), $x_i \in R_m$, $i \in \{1, -1\}$ i=1,2,…,n conduct the SVM classifier training, where each sample value is $x_i = (a_{i1},\ldots, a_{ij},\ldots, a_{im})$. The choice of the linear kernel function, using a similar type (3) the classification of functions, such as (5) as follows:

$$f(x') = \operatorname{sgn}\{(\omega \bullet x') + b\} = \operatorname{sgn}\{\sum_{i=1}^{n} \alpha_i d_i (x_i \bullet x') + b\} \qquad (5)$$

The value of component ω_j (j = 1, 2, ⋯ , m) of Equation (5) $\omega = (\omega_1, \omega_2, \cdots, \omega_m)$ represents the weight which corresponding properties a_j determine to x 'category of the sample.

If ω_j values vary close to 0, it can be identified attribute a_j makes no effect in the sample x 'category determination; additionally, if a property is removed, the rate of classification function accuracy basing on new case-based reasoning systems remained unchanged or higher than original one, which indicating that the properties has minimal or no effect on the original classifier learning. If the weight of a property is close to 0, and after being removed, classifier's classification accuracy rate increases, or at least remains the same, thus, the property can be identified as redundant properties which should be removed in the attribute reduction.

Design the following case-based reasoning system ($\hat{W}$, A, F, D, G) condition attribute reduction algorithm.

(1) Conduct training on the training sample set $\hat{W}$, get a form such as equation (5) of the SVM classifier, the classification accuracy rate recorded for the rold.

(2) j = 1.

(3)If $|\omega_j| / (\Sigma\, nj = 1|\omega_j|) > \delta$ (δ is the Lower bound of relative contribution rate pre-specified), the attribute aj corresponding to ω j should be reserved, turn (6); otherwise, to (4).

(4) Remove attribute aj, re-training support vector machine as a form of equation (5), the corresponding accuracy rate recorded as rnew. If rnew ≤ rold, then retain the property, turn (6); Otherwise go to (5).

(5) Remove attributes aj, get a new attribute set A'.

(6) Assign j = j + 1.

(7) If j ≤ n, turn (3); if j = n + 1, end, get attributes A after reduction, the new case-based reasoning system was recorded as ($\hat{W}$, A, F, D, G)'.

The reduction method taking into account two factors: the weight of the properties and classification accuracy changes, only when both are in line with certain conditions (algorithm Step 3 and Step 4), only then the property can be treated as a redundant attribute to be removed; this can ensure that valuable property is not removed probably [9] [10].

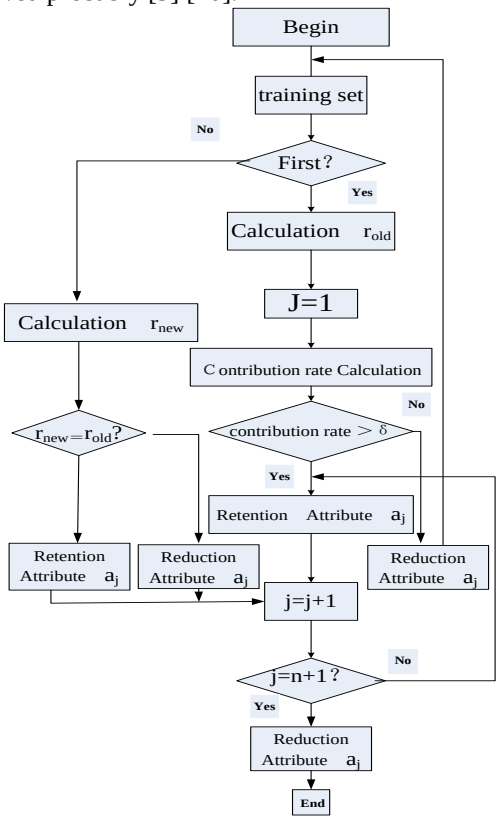

Figure 2. Overview of rules reduction process

B. Matrix-Based Iterative Learning

Such as formula (4) shows, the structure matrix M plays an important role in the case of similarity measurement. The value of structure matrix directly affects the calculation of the similarity between the source case and target. In the case of the introduction of VSM retrieval model, the structure matrix shows the relationship between cases and their contribution weight to the similarity rate of the calculation of case. The relationship between the cases can be directly calculated by the similarity between the cases, such as the calculation of the XML document structure similarity [8], by editing distance. This approach for getting some document similarity calculation is valid, but it is a fixed model to calculate the relationship between the structures and not really representative of the semantic relationship between the structural units can not change with the

variation of organizational form of relationship between organizational form and document structure. Using different ways to organize the same structural relationship, the calculation results may vary greatly, its applicability is very limited. Training to learn this relationship by learning can avoid the limitation of this approach.

During real problem, the object spaces faced in many cases are non-orthogonal space. Correlation exists between the non-orthogonal space in each dimension. This relationship is the basic to analyze the relationship between the spatial object, being concerned by researchers from different areas. A more effective method is: assuming there are mutual reliance on similarity between the entities themselves and object characteristics a [7]. Its basic form is $S_0 = B^{\mathrm{T}} S_j B$ *and*

$$S_j = B S_0 B^{\mathrm{T}} \tag{6}$$

where: B is a matrix composed of a group of objects vectors, So matrix formed by the similarity between each two object; S_j is the similarity matrix formed by the similarity between each two characters.

To avoid enormous amount of calculation by the inversion of the matrix, we can use an iterative approach to solve characteristics similar matrix of S_j in the former equation [11]. We proposed a recursive form of solving the characteristic similarity matrix S_j:

$$\begin{aligned} S_0^{k+1} &= \lambda_1 B^{\mathrm{T}} S_{\mathrm{j}}^k B + L_1^k \\ S_{\mathrm{j}}^{k+1} &= \lambda_2 B S_0^k B^{\mathrm{T}} + L_2^k \end{aligned} \tag{7}$$

Among them, the λ1 and λ2 are two true figures meeting$\lambda\ 1 \le 1/\|B\|\infty$ and $\lambda\ 2 \le 1/\|B\|1$

$$\begin{aligned} L_1^k &= I - diag(\lambda_1 B^{\mathrm{T}} S_1^k B) \\ L_2^k &= I - diag(\lambda_2 B S_0^k B^{\mathrm{T}}) \end{aligned} \tag{8}$$

Dependence assumption (9), (10) was established.

$$S = \sum_{i=1}^{n} B_{(i)}^T \cdot M \cdot B_{(i)} \tag{9}$$

$$M = \sum_{i=1}^{n} B_{(i)} \cdot S \cdot B_{(i)}^T \tag{10}$$

where: S is the matrix of similarity rate between each two cases of a case base (i) is a group of cases, the composition of the matrix on the i-th feature vector in each case the structural unit; M is the similarity rate matrix between each structural characteristics, thus the matrix structure. (9), (10) transform:

$$\begin{aligned} S_{(j,k)} &= \sum_{i=1}^{n}\sum_{u=1}^{m}\sum_{v=1}^{m}(B_{(i\mathrm{X}u,j)} M_{(u,v)} B_{(i\mathrm{X}v,k)}) \\ &= \sum_{u=1}^{m}\sum_{v=1}^{m}(M_{(u,v)} \cdot \sum_{i=1}^{n}(B_{(i\mathrm{X}u,j)} B_{(i\mathrm{X}v,k)})) \end{aligned} \tag{11}$$

$$\begin{aligned} M_{(u,v)} &= \sum_{i=1}^{n}\sum_{j=1}^{r}\sum_{k=1}^{r}(B_{(i\mathrm{X}u,j)} S_{(j,k)} B_{(i\mathrm{X}v,k)}) \\ &= \sum_{j=1}^{r}\sum_{k=1}^{r}(S_{(j,k)} \cdot \sum_{i=1}^{n}(B_{(i\mathrm{X}u,j)} B_{(i\mathrm{X}v,k)})) \end{aligned} \tag{12}$$

$$f(j,k,u,v) = \sum_{i=1}^{n} B(i)(u,j) B(i)(v,k) \tag{13}$$

$$S_{(j,k)} = \sum_{u=1}^{m}\sum_{v=1}^{m}(M_{(u,v)} \cdot f(j,k,u,v)) \tag{14}$$

$$M_{(u,v)} = \sum_{j=1}^{r}\sum_{k=1}^{r}(S_{(j,k)} \cdot f(j,k,u,v)) \tag{15}$$

And (9) is similar, the meaning (vector) of function f (j, k, u, v) is the similarity between the content of the case j, structural unit U, and case k structural unit v a. In the premise to maintain the meaning of function f (j, k, u, v), this function can be any measurements of the similarity between text.

Using the following iterative form to solve the structural matrix (structural unit matrix):

$$S_{(j,k)}^{g+1} = \begin{cases} 1, & if \quad j = k \\ \lambda \cdot \sum_{u=1}^{m}\sum_{v=1}^{m}(M_{(u,v)}^{g} \cdot f(j,k,u,v)), & if \quad j \neq k \end{cases} \tag{16}$$

$$M_{(u,v)}^{g+1} = \lambda \cdot \sum_{j=1}^{r}\sum_{k=1}^{r}(S_{(j,k)}^{g} \cdot f(j,k,u,v)) \tag{17}$$

where:

$$\begin{aligned} &\lambda \le 1 / \max\left\{ \max_{j,k}\left\{ \sum\sum f(j,k,u,v) \right\}, \right. \\ &\left. \max_{u,v}\left\{ \sum_{j=1}^{r}\sum_{k=1}^{r} f(j,k,u,v) \right\} \right\} \end{aligned} \tag{18}$$

If S is the initial value of the training set of artificial label value, this paper call it supervised learning.

$$S^O = S^H \tag{19}$$

S^H, a matrix composed by artificial dimension value between the two each documents in the training set Without labeling the training set, that every case could only be similar with its own and not similar with other cases, that is

$$S_{(j,k)}^0 = \begin{cases} 1, if \ j == k \\ 0, if \ j \neq k \end{cases} \tag{20}$$

Clearly, under the conditions of (18), the values of any element in the matrix M and S would be within the range of 0, 1 during the process of iterative computation and the iterative calculation process is convergent.

IV. EXPERIMENTS

This study includes experiments using real data sets which developed by the system case base as data set (the system case base owns the data of Pear black heart old course of the Yellow River in Anhui Province region, from May 1979 to 2008 data, a total of 1038 cases), from the hit rate and run time of the case retrieving, to validate the effectiveness of the proposed method.

A. Retrieval Phase I: Attribute Reduction of CBR

Use MATLAB7. 0 to experiment to achieve the proposed algorithm. Select the Linear Kernel K (x, y) = (x, y), the classification accuracy rate is got by 10 times cross-validation; the training time refers to the CPU time.

TABLE I. PEAR BLACK HEART DISEASE PREDICTION CASE INFORMATION SHEET CHART1

Ŵ											
	1	2	3	4	5	6	7	8	9	10	11
Mn	5	3	1	5	3	3	4	5	1	5	3
Mo	3	5	1	1	3	2	1	2	4	2	4
Mp	2	3	5	1	3	3	2	1	2	3	3
Mq	1	2	3	5	5	4	1	5	1	5	2
Ms	3	2	5	5	3	3	4	1	3	1	1
Mt	3	2	1	4	3	3	4	1	3	1	1
Rr	1	2	3	1	3	4	3	1	1	1	3

Table 2 shows forecasting system case base of Dangshan pear of black heart disease, developed in [4]. There are six factors related to disease: ten-day average temperature, ten day temperature departure, the number of precipitation days of ten-day, ten days precipitation level departure, ten-day sunshine hours, sick fruit rate, that is condition attribute set M = {n, o, p, q, s, t}. A decision r is the prediction of the forecast, target attribute set R = {r}. Mn, Mo, Mp, Mq, Ms attribute values are the same as {1, 2, 3, 4, 5}, where 1 is low, 2 lower, 3 normal, 4 the slightly higher, 5 higher ; Mt = {1, 2, 3, 4}, where 1 is very slight, 2 slight, 3 general and 4 high. Rr = {1, 2, 3, 4}, where 1 is the green safety zone, 2 orange alert, 3 yellow alert, 4 red alert. In experiment, we choose the penalty factor C = 50.

TABLE II. THE RULES OF PEAR BLACK HEART DISEASE PREDICTION CASE BASE

Case base rules											
	1	2	3	4	5	6	7	8	9	10	11
Mn	5	*	*	*	3	3	4	5	*	*	3
Mo	3	5	1	1	3	2	1	2	4	2	4
Mp	2	3	5	1	3	3	2	1	2	3	3
Mq	1	2	3	5	5	*	*	*	1	5	2
Ms	3	*	5	*	*	3	4	1	3	*	*
Mt	3	2	1	4	3	3	4	1	3	1	1
Rr	1	2	3	1	3	4	3	1	1	1	3

Note: * indicates that value ranges among desirable properties of any value.

Eliminate redundant attributes b by using SVM method, conduct training for the new case-based reasoning systems, get 7 initial supporting samples (grade 2, 3, 7, 9, 10, 11). Obtain 10 decision rules by using the method given in previous; the results are given in Table 2.

B. Retrieval Phase II: Similarity calculating.

When retrieval Phase II occurs, the primary Case Base is traversed and the current Case is compared against each Case using the similarity calculation described above. Cases that have both a high similarity and a large score difference are regarded as best. The logic behind choosing the most advantageous case is basic but not trivial to explain.

The case base has classified information, the classification of each case is the unique [12] [13]. We believe that the cases belong to the same classification has a strong similarity. In the experiment, we use classified information to replace the artificial label on the case of similarity between each two.

In the experiment, using the similarity as artificial marked similarity as the evaluation of the structure matrix of learning and KNN search results. In the experiment, we randomly selected part of the case as a training case, treated other cases as a test case. Over the training case studying, get structure matrix (the relationship between the structures of the case). Using structure matrix, with each case as a test case, to find whom the most similar to the k (k were taken10, 20, 30, etc.) cases, and by judging the case and manual annotation of the most similar cases comparison to evaluate the accuracy of the similarity search [11]. Specifically, we use the following formula for the evaluation of the results:

$$p(k) = \frac{1}{\gamma}\sum_{i=1}^{r}\frac{\left|Q_{(i,k)} \cap R_{(i,k)}\right|}{k} \tag{24}$$

Of which: k is the number of the closest case to specify a search; r is the total collection in the case; Q (i, k) is k cases which are most similar to the ith case, k cases are obtained on the basis of the similarity calculation method, R (i, k) are k cases most similar to the ith case, which are obtained based on manual annotation information

To study the effectiveness of the S-KMVSM in the similarity calculation problem, the results obtained are compared with those of other similarity calculation methods such as Tversky constancy measurement matching function, this paper use S-Tvers to identify. The dissent method (S-LENG) and weighted nearest neighbor algorithm(S-Neaneig) methods were established by using the same training data for benchmarking as for the similarity calculation of CBR.

The process of choosing the most advantageous Case is one involving similarity calculations. When comparing the values of two features, F1 and F2, a similarity score can be defined as the following:

$$sim(F1, F2) = 1 - \frac{F1 - F2}{FMAX - FMIN} \tag{21}$$

where and are two values for a feature being compared. And represent the minimum and maximum possible values for that feature, respectively. For example, if we are examining two Closest Destination Tendency Features, with values. 5 and. 4, our equation becomes:

$$Sim(F1, F2) = 1 - \frac{0.5 - 0.4}{1.0.0.0} = 0.9(90\% Similar) \tag{23}$$

This process is done for all of the Policy-Defining features, and they are aggregated and weighted as 80% of the total similarity. The final 20% of the weight is applied

to the similarity of Domination Point Ownership Ratios (An Other Notable Feature). This is all combined to form the final similarity between two cases. An example of this equation to define the similarity between cases C1 and C2 is defined thusly:

$$CaseSim(c1,c2) = a_1 \sum_{pdf's} Sim(PDFC1, PDFC2) + a_2 Sim(OWNC1, OWNC2) \quad (24)$$

An important distinction to make about Case comparison is the solutions that are involved. We want to compare our opponent's current strategy to the Losing Solution in each Case in our Case Base. I

TABLE III. THE COMPARISON OF HITTING RATE

Experiment method	Data of testing (%)	Data of experiment
S-Tvers	68.24	56.8
S-Neaneig	67.56	60.7
S-LENG	70.5	63.6
S-KMVSM	79.8	70.2

TABLE IV. THE TESTING TIME OF DIFFERENT DATA SET

Percentage of datas	S-Tvers	S-LENG	S-Neaneig	S-KMVSM
10%	15	30	10	12
30%	31	50	38	26
50%	45	68	49	35
70%	198	382	700	170

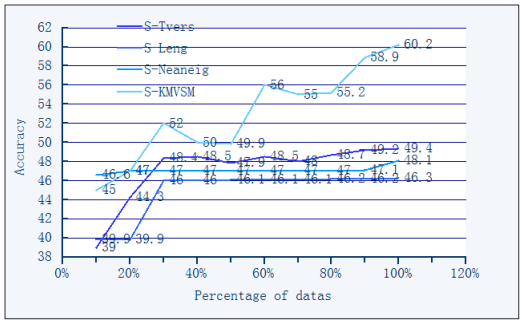

Figure 3. The accuracy of similarity retrieval under different number of training texts

This article use S-Neatening to identify it, Use structure matrix learning method of Matrix iterative learning, identified by S-KMVSM, and training cases are randomly selected from a specified number of cases in the case of systems [4] developed by the library collection, as a training set and repeating the randomly selected five times the average results of the various results shown in Table 4, which can be seen from Table 2 in this article, the hit rate of the proposed method is superior to other methods. With Longer run-time, use multi-case library in a multi-strategy data mining system to reduce the runtime. Thus User's database memory retrieval strategy in some of the old user tasks, these strategies are applied directly to the retrieval of the case base, thus reducing the time for retrieval.

Table 4 and figure 3 shows the times of the four methods under different data rates, from the table we can see that the method of S-KMVSM case retrieval of the similarity measurement has advantage in computational cost. Reduce the computational cost of retrieval, this is mainly own to reducing the impact of miscellaneous cases on the system, especially the larger the data set, the more obvious advantages, improving the computational efficiency by a range of 15% to 20% approximately.

V. CONCLUSIONS

These are some ideas as to how retrieval can be made faster, as it quickly becomes the system's bottleneck with a very large Case Base. A logical follow-up to the current implementation, therefore, would involve finding a way to reduce Case Retrieval time. One possible way to do this could involve a means of clustering the cases. Rather than having one large Case Base to span the entirety of, the Case Base can be broken into several related 'chunks'. Retrieval can then take place within a single chunk related to the current problem, rather than the entire Case Base. An issue that would quickly arise would be how to determine if Cases are 'related' and if they should be placed in the same chunk. We developed a methodology for cases similarity measures, building on the techniques from case-based reasoning, and vector space model. This paper research results use the advantage that the has a similarity with case structure, do research on retrieval of new similar case of CBR, give full consideration to the three parameters, attributes, case spatial and attribute weight which is related with retrieval. Come up with a learning algorithm of structural Matrix to automatically capture the structural relationship of cases. The experimental comparison of similarity search shows: the similarity measuring method based on this article, using of the structural information of the case, the accuracy of their experimental results has a general increase, when opposed to some of the existing similarity measure. The same learning algorithm based on matrix iteration method compared to other methods, have higher accuracy (5% to 8%), and required less training documents, the computational cost is smaller.

Another possible way to speed up retrieval time is to take advantage of multi-core machines. Most machines now have at least 2 cores, and if the code could be run efficiently in parallel, significant improvements in retrieval time could be made possible. Hence, establishing a system to monitor residual results from S-KMVSM to advise managers when to retrain the tree in S-KMVSM for maintaining accuracy is a topic for further research. Having a quad-core machine run retrieval efficiently would approach a 4x speedup, which is considerable.

ACKNOWLEDGEMENTS

This work was supported by the Natural Science Foundation of Anhui Province, China (Grant No. 1308085MF89). This work was also supported by the National Natural Science Foundation of china (Grant No. 71271071 and No. 30800663).

REFERENCES

[1] Mario Lenz. "Case-based reasoning: from foundations to applications". *Berlin: Springer*, 1998

[2] A. An, N. Cercone, C. Chan, Integrating rule induction and case-based reasoning to enhance problem solving, *Lecture Notes in Computer Science*, 1997 12 (6) pp. 499-508.
[3] Juan L. Castro, Maria Navarro, Jose M. Sanchez, Jose M. Zurita, "Introducing attribute risk for retrieval in case-based reasoning", *Knowledge-Based Systems*, 24(2), 2011: pp. 57-268.
[4] A. An, N. Cercone, C. Chan, Integrating rule induction and case-based reasoning to enhance problem solving, *Lecture Notes in Computer Science*, 199712(6) 499-508.
[5] Gu Lichuan, Ni Zhiwei, Li Shw. "Forecasting Method on Pear Scab Based on Fusion Reasoning", *CIS2007 proceedings. Harbin, China: IEEE Press*, 2007 pp. 401-404
[6] Li Haifang, Wei Xiaoyan, Chen junjie." Multi-dimensional reduction technique research on case retrieval model in CBR". *Computer Engineering and Applications*, 2008, 44(25) pp. 157-160
[7] Leake, D., Kinley, A., and Wilson, D. Learning to improve case adaptation by introspective reasoning and CBR. *In Proceedings of the First International Conference on Case-Based Reasoning, Berlin. Springer Verlag.* 1995 pp. 229-240
[8] Yingjie Song, Rong Chen, Yaqing Liu." A Non-Standard Approach for the OWL Ontologies Checking and Reasoning". *Journal of Computers*, 2012, 7(10) pp. 2454-2461
[9] Qi Jin, Hu Jie, Peng Yinghong, Wang Weiming, Zhan Zhenfei, "AGFSM: A new FSM based on adapted Gaussian membership in case retrieval model for customer-driven design", *Expert Systems with Applications*, 38(1), 2011 pp. 94-905;
[10] N. Arshadi, I. Jurisica, Data mining for case-based reasoning in high-dimensional biological domains, *IEEE Transactions on Knowledge and Data Engineering* 17 (8) (2005) pp. 1127-1137;
[11] Ping Yan, Yan He, Runzhong Yi, Fei Liu, Guorong Chen, "A Measuring Method for Angular Displacement Based on Correlation Algorithm". *Journal of Computers*, 2012, 7(10) pp. 2376-2382
[12] Md. Tarek Habib, Feature Selection for Fabric Defect Classification Using Neural Network, *Journal of Multimedia*, 2011, 6(5) pp. 416-424
[13] Gleb Beliakov, John Yearwood, Andrei Kelarev, "Application of Rank Correlation, Clustering and Classification in Information Security", *Journal of Networks*, 2012, 7(6) pp. 935-945
[14] H. Ahn, K. -J. Kim, Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach, *Applied Soft Computing*, 2009, 9 (2) pp. 599-607.
[15] A. An, N. Cercone, C. Chan, Integrating rule induction and case-based reasoning to enhance problem solving, *Lecture Notes in Computer Science*, 1997, 12(6) pp. 499-508.

Lichuan GU, male, born on October 1974 in Sichuan, China. He received his B. Sc., M. Sc. degrees in computer science from HeFei University of technology, China, in 1997 and 2005. Currently, he is an Associate Professor at Department of Computer Science of Anhui Agricultural University. He has wide research interests, mainly including machine learning, data mining, pattern recognition and artificial intelligence.