Apply the GA and CBR approach to designing an e-training system

Shu-Lin Wang[1], Zhong-Ying Lin[2], Chun-Yu Hsieh[3]
Department of Information Management[1]
National Taichung Institute of Technology [2,3]
National Taichung Institute of Technology, Taichung, Taiwan, ROC
shulin@ntit.edu.tw[1*], linzhongying@gmail.com[2]

Abstract-*With the booming of informational technology and Internet development, it is a trend in knowledge creation and management for enterprises to apply e-learning in their educational training. The purpose of this study attempts to integrate genetic algorithm (GA), case-based reasoning (CBR), and knowledge management (KM) approach to acquire, to classify and to organize the knowledge of maintenance case. At last, the problem-based learning strategy (PBL) had been used for curriculum planning with an aim at designing digital contents for the web-based training platform.*

Keywords: e-training system, case-based reasoning, knowledge management, genetic algorithm.

1. Introduction

In recent years, artificial intelligence (AI) approaches, including inductive learning [1] [2], case-based reasoning (CBR) [3] [4] and neural networks (NNs) [5] [6] [7] [8] [9], etc., have been applied extensively to solve problems in different fields. Among them, CBR is an instrument applicable for supporting decision-making. It can be used to solve complicated and dynamic problems in decision-making and, as a result, "cases" become its principal basis of reasoning to make reuse of existing knowledge possible [10] [11]. This approach is similar to the process of people solving problems in real life – finding in their previous experience situations closest to problems presently encountered and use them as the foundation to conduct further reasoning. It has been exercised in various domains over the past years to help people solve problems [12] [13] [14] [15]. Results of many studies show that through this approach human experience and knowledge can be converted and stored in computers for future needs. Consequently, this approach is therefore a feasible strategy for developing online maintenance training systems.

However, the question of how to cluster and organize these data effectively to build them into applicable training materials to place in e-training systems for enterprises become the key to success for e-training systems that employ CBR as its foundation. This study adopts the genetic K-means algorithm (GKA) Krishna and Narasimha Murty (1999) introduced by combining genetic algorithm and K-means clustering algorithm, and replaces the crossover procedure in genetic algorithm with K-means. The chief target is to accelerate the rate of convergence when searching for the global optimum as well as to achieve optimized clustering of maintenance cases throughout the entire clustering process.

In the end, this study applies airplane engine maintenance and fault diagnosis training as an example to establish an e-training system, in the purpose that the approach can support learners to conduct independent e-training activities.

2. System method

2.1. Case-based reasoning

The so-called case-based reasoning (CBR) means use of precedents as the basis of analogical reasoning to make reuse of existing knowledge possible [10]. The chief consideration in such a reasoning approach is human reasoning modes in reality – how they find in their experience situations closest to problems they encounter, make certain modification to solve the problem at hand, and, finally, store the problem-solving experience and knowledge systematically for later use when coming across problems again [18] [19]. Detailed description of the 4 steps of the reasoning cycle mechanism (Figure 2.1) [12] is as follows:

1. RETRIEVE ：Discovers the most similar case from the case base
2. REUSE：The use finds case question solution and knowledge
3. REVISE：The revision system outputs result
4. RETAIN：Already succeeded the solution case method and the knowledge stores in the case base, made the new case for the next reference

The next chart is CBR cycle:

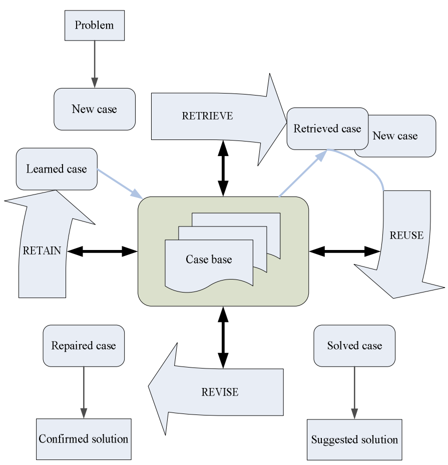

Figure 2.1 CBR cycle [18]

To set up the example of airplane engine maintenance and fault diagnosis training in this study, the maintenance experience and knowledge of airplane engineers are extracted to establish a case database to provide teaching materials in the e-training system applying case-based reasoning thus developed as the foundation. To make it easier to systematically organize and cluster relevant case knowledge for application in case-based reasoning and help learners training in maintenance through a methodical approach, this study adopts the "neatest-neighbor (NN) approach" Kolodner proposed in 1993 to calculate the similarities between any two cases according to the NN principle.

2.2. Representation of Case Knowledge

This study uses "cases" to represent the basic knowledge elements of airplane engine maintenance. Relevant maintenance training materials are collected through interviews with specialists to build the case knowledge base. Materials for e-training include three parts: (1) description of case characteristics, (2) causes of malfunctions, and (3) malfunction diagnosis procedures.

2.3. GKA clustering process of case knowledge

In this study, when a new case is imported into the system, the CBR mechanism will compare it with all the existing cases in the case database and calculate to produce the match with the highest similarities. For retrieval of case knowledge, this study first employs the GKA approach to cluster all the cases and then establish the retrieve interface (shown as Figure 3.1) according to genetic clustering results to provide easier access for learners to obtain related maintenance case knowledge to conduct training. Detailed GKA clustering process is describes as following.

The purpose of adopting genetic algorithm in combination with K-means algorithm to perform clustering is mainly to make case knowledge more explicitly to allow users easy to retrieval the learning contents of digital learning. The following is the step-by-step description of complete algorithmic process:

Step 1: Data Input

The data inputs are similarity matrixes of every two cases.

Step 2: Parameter Setting

There are 534 cases in total in this study. n = 534 and each case is given a characteristic vector of d. When all the cases are divided into k clusters, the related parameters of the objective function are defined as follows:

1. $\{x_i,\ i = 1, 2, \ldots, n\}$: represent cases
2. x_{ij} : the number j characteristic value of x_i

3. When i =1, 2, …, n and k= 1, 2, …, K, then :

$$w_{ik} = \begin{cases} 1, & \text{if } i\text{th pattern(case) belongs to } k\text{th cluster,} \\ 0, & \text{otherwise.} \end{cases} \qquad (1)$$

4. Matrix $W = \left[w_{ij} \right]$ with the following characteristics:

$$w_{ij} \in \{0,\ 1\} \text{ and, } \sum_{j=1}^{K} w_{ij} = 1 \qquad (2)$$

5. Define centroid in number k cluster, $c_k = (c_{k1},\ c_{k2},\ \cdots,\ c_{kd})$:

$$c_{kj} = \frac{\sum_{i=1}^{n} w_{ik} x_{ij}}{\sum_{i=1}^{n} w_{ik}} \qquad (3)$$

6. Define the in-cluster differences in number k cluster:

$$S^{(k)}(W) = \sum_{i=1}^{n} w_{ik} \sum_{j=1}^{d} (x_{ij} - c_{kj})^2 \qquad (4)$$

and Total within-cluster variation (TWCV):

$$S(W) = \sum_{k=1}^{K} S^{(k)} = \sum_{k=1}^{K} \sum_{i=1}^{n} w_{ik} \sum_{j=1}^{d} (x_{ij} - c_{kj})^2 \qquad (5)$$

7. The objective function is to locate $W^* = \left[w_{ik}^* \right]$, or square-error measure (SEM). Therefore, the objective function can be defined as minimizes $S(W)$ as shown in the following equation:

$$S(W^*) = \min \ \{S(W)\}$$

Step 3: GKA clustering process [16]

1. Coding: The "string-of-group-number encoding" proposed by Jones and Beltramo (1991) is used in this paper. In this research every gene represents a case. All the cases are numbered to become a group of strings of s_w, with a cluster number given at random between 1 ~ K, K being the highest cluster number. Case coding is as shown in Figure 2.2 below:

X_1	X_2	X_3	X_4	X_5	X_6	…	X_n
3	2	1	2	1	3	…	K

Figure 2.2 Case code

2. Initialization: A P (0) is created randomly to be the initialized cluster. Each chromosome will be given a cluster number at random during the initialization process. The cluster numbers will be allotted from 1, 2. …, K.

3. Setting of the Fitness Function: In the "σ-truncation mechanism" adopted, the fitness function is used as the performance indicator to evaluate the quality of the chromosomes in the genetic algorithm. The fitness value is used as the basis of reproduction of a new generation in the selection stage of genetic algorithm. The fitness function is as shown below:

$$F(s_w) = \begin{cases} g(s_w), & \text{if } g(s_w) \ge 0 \\ 0, & \text{otherwise} \end{cases} \qquad (6)$$

s_w: The target of this study is to find the minimized problem of the objective function; therefore,

$$f(s_w) = -S(W),$$
$$g(s_w) = f(s_w) - (\overline{f} - c \cdot \sigma)$$

$\overline{f}$: Average;

σ : Standard deviation of $f(s_w)$;

c : Constant, the value range is between 1 and 3.

4. Selection: The Roulette wheel selection approach is adopted to determine which chromosomes will be retained for the next generation. Selection is conducted according to the fitness value of the chromosome. The roulette is divided in accordance with the fitness values of the chromosomes of the clusters. As a result, chromosomes with a larger fitness value have a higher probability of getting selected for the next generation. The definition of the probability of a chromosome getting selected is as follows:

$$P(s_i) = \frac{F(s_i)}{\sum_{j=1}^{N} F(s_j)} \qquad (7)$$

$F(s_i)$ is the fitness value of s_i.

The steps are:

(1) Calculate the chromosome's fitness value $F(s_i)$;

(2) Calculate the selection probability of chromosome (i): $P(s_i) = \dfrac{F(s_i)}{\sum_{j=1}^{N} F(s_j)}$;

(3) By random selection, create a random number between 0 and 1.

(4) When the accumulated value of the selection probability of a chosen chromosome is higher or equal to the random number created, select this chromosome.

(5) Repeat steps (3) and (4). When the number of returned chromosomes is equivalent to the number of clusters, stop selection.

5. Mutation: Mutation is for preventing from falling in local optima when searching for the optimum through genetic algorithm. The GKA mutation approach is described below:

(1) Set $s_w(i)$ as the number x_i case and use the Euclidean distance to calculate the distance between x_i and the cluster core: $d_j = d(x_i, c_j)$;

(2) Among the chromosomes, select a position as the mutation point. The selecting approach is:

$$p_j = \Pr\{s_w(i) = j\} = \frac{c_m d_{\max} - d_j}{\sum_{i=1}^{K}(c_m d_{\max} - d_i)} \qquad (8)$$

c_m: a constant large than 1.

$$d_{\max} = \max_j \{d_j\}$$

6. K-means: The "one-step K-means" algorithm, called "K-means operator (KMO) for short, is adopted to replace the crossover procedure in genetic algorithm. In this study, to begin, S is deemed as a set of strings and the two following steps are taken to produce strings of a new generation:

(1) Calculate the position of the cluster center according to $c_{kj} = \frac{\sum_{i=1}^{n} w_{ik} x_{ij}}{\sum_{i=1}^{n} w_{ik}}$;

(2) Calculate the distance between each case and the center of each cluster, and then re-cluster in accordance with the distance of each case to the cluster center.

During this process, there may be certain cases in S that are not assigned to any cluster. Under such circumstances, eliminate these illegal strings through the selection procedure in the next genetic evolution process.

7. Termination: There are two conditions to termination:

(1) Terminate the evolution when it reaches the set maximum number of generations.

(2) Terminate the evolution when no better results are produced after 20 generations.

Step 4: Category Definition

After evaluation by specialists, give each cluster an appropriate category name that clearly indicates the meaning of each cluster.

3. System Implementation

3.1. System architecture

In this study, case-based reasoning (CBR), Knowledge management and the GKA approach are integrated to obtain, categorize and organize case knowledge into teaching materials. The e-training system adopted chiefly includes the following function modules: (1) a case knowledge base, (2) a learning evaluation database, (3) a knowledge acquisition module, (4) a problem-based learning resource module, (5) an online test module, (6) a teaching material and learning management module, and (7) supplementary learning tool. Detailed descriptions of the function modules and the platform operation modes are as follows:

(1) Case Knowledge Base: This system uses a case knowledge base to store case knowledge, which includes "case attribute indicators and pertaining description vocabulary" and "case solution knowledge"

(2) Learning Evaluation Database: This database stores a fault diagnosis item bank to support the online test module and records the result of each test taken by each student.

(3) Knowledge Acquisition Module: The chief purpose of this module is to obtain case knowledge to

establish and maintain the case database.

(4) Problem-Based Learning Resource Module: The main function of this module is to provide simulated training for identification of faulty configurations. It displays the topology of the relevant attribute weight of each case after case-based reasoning to allow learners to have a clear picture of the relations between cases and train repeatedly maintenance engineers' ability to identify fault characteristics and diagnose fault causes. The system function mode is as shown in Figure 3.1, Figure 3.2 and Figure 3.3.

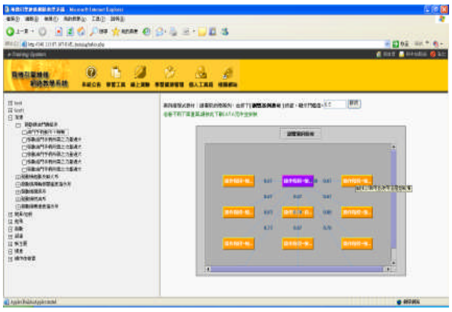

Figure 3.1 the interface of case knowledge retrieve

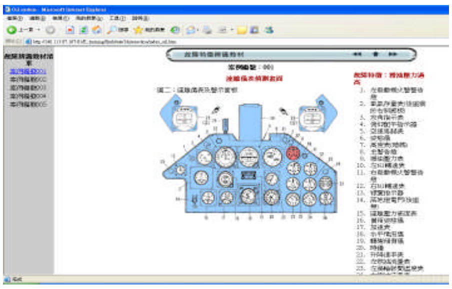

Figure 3.2 the interface of simulated training for identification of faulty configurations

Figure 3.3 the interface of training for diagnose fault causes and repair method

(5) Online Test Module: Available in this module are a fault diagnosis test item bank, fault diagnosis online tests, and a questionnaire to survey the feasibility of this platform.

(6) Teaching Material and Learning Management Module: This module chiefly includes acquisition of teaching material knowledge.

(7) Supplementary Learning Tool: This is the area for discussion, idea-sharing and Q&A.

4. Conclusion

Genetic algorithm and the case-based reasoning approach are integrated in this study to classify and organize case knowledge to help learners obtain knowledge effectively. In the end, airplane engine maintenance is used as an example to establish a web-based e-training system. For organizing case knowledge, this study uses GKA to arrange case knowledge. GKA combines genetic algorithm and K-means algorithm for the purposes of accelerating the rate of convergence when searching for the global optimum and promoting the optimized result in the clustering process. Through simple coding of all the cases in the GKA approach, excessive and complex procedures can be avoided when conducting clustering calculation and the clustering results can be optimized to establish the most ideal cluster categorization of case knowledge for e-training system.

Acknowledgements

This research was sponsored by research grants form National Science Council (NSC) of ROC Taiwan (NSC 94-2213-E-025-003).

References

[1] I. Han, J. S. Chandler, and T. P. Liang, "The impact of measurement scale and correlation structure on classification performance of inductive learning and statistical methods", *Expert System with Applications*, vol. 10, no. 2, pp. 209-221, 1996.

[2] M. Shaw, and J. Gentry, "Using an expert system with

inductive learning to evaluate business loans", *Financial Management*, vol. 17, no. 3, pp. 45-56, 1988.

[3] S. M. Bryant, "A case-based reasoning approach to bankruptcy prediction modeling", *International Journal of Intelligent Systems in Accounting, Finance and Management*, vol. 6, no. 3, pp. 195-214, 1997.

[4] P. Buta, "Mining for financial knowledge with CBR", *AI Expert*, vol. 9, no. 10, pp. 34-41, 1994.

[5] J. E. Bortiz, and D. B. Kennedy, "Effectiveness of neural network types for prediction of business failure", *Expert Systems with Application*, vol. 9, no. 4, pp. 503-512, 1995.

[6] J. R. Coakley, and C. E. Brown, "Artificial neural networks in accounting and finance: Modeling issues", *International Journal of Intelligent Systems in Accounting, Finance and Management*, vol. 9, no. 2, pp. 119-144, 2000.

[7] H. Jo, and I. Han, "Integration of case-based forecasting, neural network, and discriminant analysis for bankruptcy prediction", *Expert Systems with applications*, vol. 11, no. 4, pp. 415-422, 1996.

[8] G. Zhang, M. Y. Hu, B. E. Patuwo, and D. C. Indro, "Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis", *European Journal of Operational Research*, vol. 116, no. 1, pp. 16-32, 1999.

[9] Sung-Hwan Min, Jumin Lee, and Ingoo Han , "Hybrid genetic algorithms and support vector machines for bankruptcy prediction", *Expert Systems with Applications*, vol. 31, no. 3, pp. 652-660, 2006.

[10] J. L. Kolodner, *Case-Based Reasoning*, Morgan Kaufmann Publishers, San Mateo, 1993.

[11] A. Aamodt, E. Plaza, "Case-based reasoning: foundations issues, methodological variations, and system approaches", *AI Communications*, vol. 7, no. 1, pp. 39-59, 1994.

[12] J.M. Garrell i Guiu, E. Golobardes i Ribe´, E. Bernado´ i Mansilla, X. Llora` i Fa`brega, "Automatic diagnosis with genetic algorithms and case-based reasoning", *Artificial Intelligence in Engineering*, vol. 13, no. 4, pp.367-372, 1999.

[13] Yonggang Fu, Ruimin Shen,GA based CBR approach in Q&A system, *Expert Systems with Applications*, vol. 26, no. 2, pp.167-170, 2004.

[14] Chaochang Chiu, "A case-based customer classification approach for direct marketing", *Expert Systems with Applications*, vol. 22, no. 2, pp.163-168, 2002.

[15] D.W. Aha, "Tolerating noisy, irrelevant and novel attribute in instanced-based learning algorithm", *International Journal of Man-Machine Studies*, vol. 36, pp.267-287, 1992.

[16] K. Krishna and M. Narasimha. Murty, "Genetic K-Means Algorithm", *IEEE Transaction on System, Man and Cybernetics-Part B:Cybernetics*, vol. 29, no. 3, pp.433-439, June, 1999.

[17] Mary Lou Maher and Andres Gomez de Silva Garza, "Case-Based Reasoning in Design", IEEE Expert, vol. 12, no. 2, pp.34-41, March, 1997.

[18]K.D. Althoff, "Knowledge acquisition in the domain of CBC machine centres: the MOLTKE approach", *EKAW-89*, pp.180-195, July, 1989.

[19] D. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, USA, 1989.