S5-5

A METHOD OF REVISING RETRIEVED SIMILAR CASES IN GA-CBR COST MODELS

Sooyoung Kim[1], Hyun-Soo Lee[2], Moonseo Park[3], Sae-Hyun Ji[4] and Joseph Ahn[5]

[1] Unified course of the master's and the doctor's, Seoul National University, Seoul, Korea
[2] Professor, Seoul National University, Seoul, Korea
[3] Associate Professor, Seoul National University, Seoul, Korea
[4] Ph.D. Candidate, Seoul National University, Seoul, Korea
[5] M.S. Student, Seoul National University, Seoul, Korea
Correspond to *wing195@snu.ac.kr*

ABSTRACT: Early cost estimates are important to decision-making for a construction project. Moreover, the possibility of reducing the project cost is getting less as the project is progressed. Case-based reasoning (CBR), which can be viewed as an effective method for early cost estimating, is widely utilized recently. Early cost estimates using CBR have advantages over the traditional ones as they produce reasonable outputs and self-studying is possible by simply adding new cases. Case-based reasoning is composed of a cycle of retrieve, reuse, revise, and retain process. However, in the majority of research cases, they are focused on how to retrieve the similar cases, instead of revising the cases which is expected to increase accuracy results of cost estimation. This research suggests a method of revising retrieved similar cases in a GA-CBR cost model which is widely studied and utilized for early cost estimating recently. To validate the proposed method, case study is conducted based on Korean public apartment projects.

Keywords: Case-based Reasoning; Cost Model; Genetic Algorithm; Early Cost Estimating; Revise

1. INTRODUCTION

1.1 Background and Objective

Early estimates are critical to the initial decision-making process for the construction of capital projects (Trost and Oberlender, 2003). Moreover, the possibility of reducing the project cost is getting less as the project is progressed (Duverlie and Castelain, 1999). Because of these reasons, early cost estimating is essential process for the successful project achievement. However, it is difficult to accurately estimate construction cost in the early stage of a project due to various uncertainties in construction.

To estimate proper construction cost, it is necessary to compare with actual cost data of the other projects. Case-based reasoning (CBR), which can be viewed as an effective method for early cost estimating, is widely utilized recently. A case-based reasoning solves new problems by using or adapting solutions that were used to solve old problems. Early cost estimates using case-based reasoning have advantages over the traditional ones as they produce reasonable outputs and self-studying is possible by simply adding new cases.

Case-based reasoning is composed of a cycle of retrieve, reuse, revise, and retain process. However, in the majority of research cases, they are focused on how to retrieve the similar cases, instead of revising the cases. In other words, retrieved similar cases are directly used to solve a problem without revising. Consequently, this leads to lower accuracy results of cost estimation.

This research suggests a method of revising retrieved similar cases in GA-CBR cost models which is widely studied and utilized for early cost estimating recently. To validate the proposed method, case study is conducted based on Korean public apartment projects.

1.2 Scope and Methodology

This research is focused on the early stage of cost estimating, which the only limited information of construction can be obtained. It is carried out based on cost data of Korean public apartment projects.

In order to suggest the method of revising cases in GA-CBR models, this research applied the following procedure.

(1) The principle of case-based reasoning and genetic algorithms is examined.

(2) Previous researches of revising cases in case-based reasoning are analyzed.

(3) A new method which can revise retrieved cases in CBR is suggested.

(4) In order to validate the suggested method, this research compares revised cost data with its predecessor based on actual data.

2. PRELIMINARY STUDY

2.1 Case-Based Reasoning

Case-based reasoning is a problem solving approach that it utilizes solutions of past experiences to solve problem. A new problem is solved by finding a similar past case, and reusing it in the new problem situations (Aamodt and Plaza, 1994). Case-based reasoning contains two main assumptions, which are that similar problems have similar solutions and once happened problem tends to come about again (Watson and Marir, 1994). It is widely utilized in the construction field such as construction design, decision making, scheduling and cost estimating, which need to consider past experiences and knowledge..

Case-based reasoning may be described by the following four processes; "the four **RE**s" (Aamodt and Plaza, 1994).

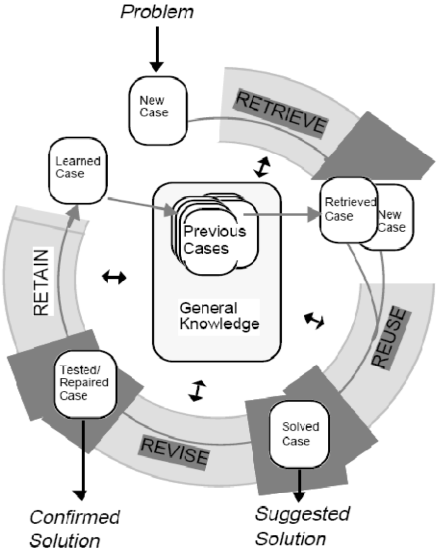

Figure 1. The CBR cycles
(Aamodt and Plaza, 1994)

An initial description of a problem defines a new case. This new case is used to RETRIEVE a case from the collection of previous cases. The retrieved case is combined with the new case - through REUSE - into a solved case, i.e. a proposed solution to the initial problem. Through the REVISE process this solution is tested for success, e.g. by being applied to the real world environment or evaluated by a teacher, and repaired if failed. During RETAIN, useful experience is retained for future reuse, and the case base is updated by a new learned case, or by modification of some existing cases (Aamodt and Plaza, 1994).

2.2 Genetic Algorithm

Genetic algorithms are adaptive heuristic search algorithm premised on the evolutionary ideas of natural selection and genetic (Gen and Cheng, 2000). Genetic algorithms use techniques inspired by evolutionary biology such as mutation and crossover. Mutation is a genetic operator used to maintain genetic diversity from one generation of a population of chromosomes to the next. Crossover is a genetic operator used to vary the programming of a chromosome or chromosomes from one generation to the next.

2.3 Preliminary Research of Revising Method in CBR Models

Most researches for case-based reasoning are focused on retrieve phase. On the contrary to this, few researches have been executed for a revise phase.

Rial et al.(2001) suggested a method for automating the revise phase of CBR systems. It is carried out by a belief revision technique. Belief revision is useful in terms of automating system; however, when the number of rules increases, its computational complexity grows.

Ji, C. Y.(2010) suggests a CBR revision model for predicting the construction cost of multifamily projects. It uses two methodology; one is feature counting and the other is multiple regression analysis. However, feature counting cannot reflect differences among attributes. Multiple regression analysis includes assumptions. One of the important assumptions is that the predictors are linearly independent. Nevertheless, because of non-linear and multicollinearity, the reliability of the revising model is decreased.

3. GA-CBR COST MODEL

Ji, S. H. et al.(2009), Park et al.(2010), and Kim et al.(2010) suggest case-based reasoning cost models utilizing genetic algorithms. The explanation of these models is as follows.

The project cost of a specific case can be formulated by appropriately weighting its attributes.

$$\mathrm{C_i} = \mathrm{X_{i1}W_1} + \mathrm{X_{i2}W_2} + \cdots + \mathrm{X_{ij}W_j} \tag{1}$$

where, $\mathrm{C_i}$: the cost of *i*th case
$\mathrm{X_{ij}}$: the value of *j*th attribute of *i*th case
$\mathrm{W_j}$: the weight of *j*th attribute

When this relationship is expanded to a set of general cases, it is described by the matrix formula below.

$$\begin{pmatrix} \mathrm{X_{11}} & \cdots & \mathrm{X_{1j}} \\ \vdots & \ddots & \vdots \\ \mathrm{X_{i1}} & \cdots & \mathrm{X_{ij}} \end{pmatrix} \times \begin{pmatrix} \mathrm{W_1} \\ \vdots \\ \mathrm{W_j} \end{pmatrix} = \begin{pmatrix} \mathrm{C_1} \\ \vdots \\ \mathrm{C_i} \end{pmatrix} \tag{2}$$

In order to make a range of weights from 0 to 1, assuming that all attributes and costs are normally distributed, they are converted to a standard cumulative normal distribution.

$$\begin{pmatrix} \mathrm{x_{11}} & \cdots & \mathrm{x_{1j}} \\ \vdots & \ddots & \vdots \\ \mathrm{x_{i1}} & \cdots & \mathrm{x_{ij}} \end{pmatrix} \times \begin{pmatrix} \mathrm{w_1} \\ \vdots \\ \mathrm{w_j} \end{pmatrix} = \begin{pmatrix} \mathrm{c_1} \\ \vdots \\ \mathrm{c_i} \end{pmatrix} \tag{3}$$

where, c_i : the cost of *i*th case (standardization), $0\leq c_i \leq 1$
x_{ij} : the value of *j*th attribute of *i*th case (standardization), $0\leq x_i \leq 1$
w_j : the weight of *j*th attribute, $0\leq w_j \leq 1$

Attribute weights are optimized to minimize the sum of the absolute value of the distance by genetic algorithms.

$$\begin{pmatrix} c_1 \\ \vdots \\ c_i \end{pmatrix} - \begin{pmatrix} x_{11} & \cdots & x_{1j} \\ \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ij} \end{pmatrix} \times \begin{pmatrix} w_1 \\ \vdots \\ w_j \end{pmatrix} = \begin{pmatrix} d_1 \\ \vdots \\ d_i \end{pmatrix} \tag{4}$$

$$\text{optimizing } w_j \text{ for } \min \sum_{k=1}^{i} |d_k|$$

where, d_i : distance of case *i*th

Similar cases are retrieved from the database by utilizing attribute values and attribute weights. By calculating this, the cost of a target case is estimated.

This method can attain reliable results based on genetic algorithms; however, calculation time is longer than other methods such as gradient descent and multiple regression analysis.

4. THE METHOD OF REVISING CASES IN GA-CBR COST MODEL

In order to revise cases in GA-CBR models, this research analyzes the estimation error arising from difference between a target case and retrieved similar cases.

The estimation error of each attribute value can be explained by the formula below.

$$r_{ij} = w_j(x_{tj} - x_{ij}) \tag{5}$$

where, r_{ij} : the estimation error for *j*th attribute's residual of *i*th case
w_j : the weight of *j*th attribute, $0\leq w_j \leq 1$
x_{tj} : the value of *j*th attribute of a target case (standardization), $0\leq x_{tj} \leq 1$
x_{ij} : the value of *j*th attribute of *i*th case (standardization), $0\leq x_i \leq 1$

When this formula is expanded to all the attributes, it can be described by the formula below.

$$R_i = \sum r_i = \sum w_j(x_{tj} - x_{ij}) \tag{6}$$

where, R_i : sum of estimation error of case *i* (standardization)

Thus, the cost estimation error which is resulted by difference between a target case and retrieved cases is R_i. The cost of retrieved similar cases can be revised to the target case like formula below.

$$c_i' = c_i + R_i = c_i + \sum w_j(x_{tj} - x_{ij}) \tag{7}$$

Where, c_i' : revised cost of case *i* (standardization)

c_i' is converted to real value like below formula.

$$C_i' = (c_i' \times \sigma_j) + m_j \tag{8}$$

Where, C_i' : revised cost of case *i*
σ_j : standard distribution of *j*th attribute
m_j : average of *j*th attribute

The figure 2 represents a process of revising the cases in this research.

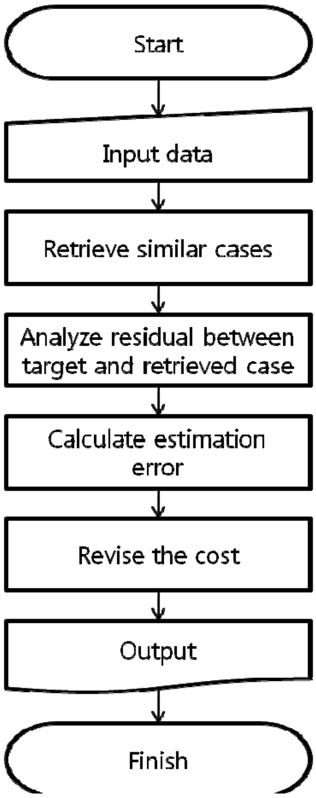
Start
Input data
Retrieve similar cases
Analyze residual between target and retrieved case
Calculate estimation error
Revise the cost
Output
Finish

Figure 2. Process of Revising Cases

5. CASE STUDY

5.1 Data Analysis

The cost data of Korean public apartment projects are distributed from 2006 to 2008. It is normalized to December 2008 with the cost index of Korea Institute of Construction Technology (KICT). Total 76 data is used to build database and remaining ten are left for validation. Seven attributes are extracted by interviewing experts. Table 1 below shows information of the attributes.

Table 1. Information of the Attributes

Attribute	Measure	Range
Number of households	People	5-56
Gross floor area	m^2	546-6,169
Number of unit floor households	EA	1-6
Number of floors	Floors	3-15
Number of elevators	EA	1-3
Number of households of unit floor per elevator	EA	2-4
Number of piloti with households scale	EA	0-6

For assigning attribute weights, Evolver, the software of genetic algorithm application for MS Excel, was used to find an optimal attribute weight. Conditions for optimization by genetic algorithms are as follows. Initial population is 50, crossover rate is 0.05, mutation rate is 0.1 and stopping condition is that trials reach 5,000,000. The Table 2 below represents the attribute weights which were calculated by using Evolver 4.0.

Table 2. Calculated Weight Values

Attribute	Weight
Number of households	0.0000
Gross floor area	0.8122
Number of unit floor households	0.0000
Number of floors	0.0901
Number of elevators	0.0490
Number of households of unit floor per elevator	0.0000
Number of piloti with households scale	0.0467

5.2 Experiment Design

The model was validated to examine reliability using data of Korean public apartment projects. In the total of 86 cases, ten test cases were selected for validation by using simple random sampling method. These are used to the existing cost model and the revised cost model. These are compared to an average cost of five nearest neighbors. It is assumed that revised cost model can be improved in terms of accuracy when comparing average absolute deviations of accuracy rate are relatively lower than another one. Furthermore, stability of the model can also be examined by comparing standard deviations of absolute deviations.

5.3 Results and Discussions

As shown in Table 3, although it is slightly different depending on individual case, the revised cost model was resulted in an overall lower absolute deviation. Moreover, standard deviation of absolute deviations was also resulted lower than the existing cost model. These results represent that the suggested model is improved in terms of accuracy and stability.

Table 3. Comparison of Absolute Deviation

(Absolute Deviation, %)

Case	Original Cost Model	**Revised Cost Model**
1	1.7	**0.9**
2	13.9	**2.4**
3	31.4	**1.6**
4	2.8	**4.8**
5	1.9	**1.2**
6	24.6	**8.6**
7	4.2	**11.8**
8	3.6	**1.6**
9	6.9	**5.4**
10	5.9	**13.7**
Avg.	9.7	**5.2**
S.D.	10.4	**4.7**

6. CONCLUSIONS

To estimate proper cost using case-based reasoning, revising cases is no less important than retrieving cases. However, accuracy of the existing cost models is decreased due to not considering the revise process.

This research suggests the method of revising retrieved similar cases in GA-CBR cost models. It reflects the estimation error caused by differences between attribute values. Consequently, retrieved cost is revised to proper one.

To verify the method, this research utilized the model which was built and case studied based on Korean public apartment project. It is shown that accuracy and stability are improved compared to the existing model which did not consider revising process.

The proposed method utilized attribute weights and converted attribute values which can be obtained in the retrieve phase in GA-CBR models. Ultimately, this research makes it possible to revise the cost without additional rules or database. As the research is limited to Korean public apartment projects, applying to other kinds of building projects is required. Moreover, improving accuracy and usability by applying other cost estimation methods is necessary.

REFERENCES

[1] Aamodt, A. and Plaza, E., “Case-based reasoning: Foundational issues, methodological variations and system approaches”, AI Communications, Vol. 7(1), pp. 35-39, 1994

[2] Duverlie, P. and Castelain, J. M., "Cost Estimation During Design Step: Parametric Method versus Case Based Reasoning Method", The International Journal of Advanced Manufacturing Technology, Vol. 15, pp. 895–906, 1999

[3] Gen, M. and Cheng, R., Genetic algorithms and engineering optimization. 1st ed. Wiley-IEEE, 2000

[4] Ji, C. Y. et al. “CBR Revision Model for Improving Cost Prediction Accuracy in Multifamily Housing

Projects", Journal of Management in Engineering", Vol. 26(4), pp.229-236, 2010

[5] Ji S. H. et al., "Cost Estimation Model for Building Projects Using Case-Based Reasoning", Submitted to the Canadian Journal of Civil Engineering at oct 21st. 2009

[6] Kim et al., "A Method of Assigning Weight for Qualitative Variables in CBR Cost Model", 8th International Symposium on Architectural Interchanges in Asia, 2010

[7] Park et al., "Schematic Cost Estimation Method using Case-Based Reasoning: Focusing on Determining Attribute Weight", Journal of Korea Institute of Construction Engineering and Management, Vol. 11(4), pp.22-31, 2010

[8] Rial, R.P. et al. "Improving the Revision Stage of a CBR System with Belief Revision Techniques", Computing and Information Systems, Vol. 8, pp.40-45, 2001

[9] Trost, S. M and Oberlender, G. D., "Predicting Accuracy of Early Cost Estimates Using Factor Analysis and Multivariate Regression", Journal of Construction Engineering and Management, Vol. 129(2), ASCE, pp. 198-204, 2003

[10] Watson, I. and Marir, F., "Case-Based Reasoning: A review", the Knowledge Engineering Review, Vol. 9(4), pp. 355-381, 1994