© Nottingham University Press
Proceedings of the International Conference on Computing in Civil and Building Engineering
W Tizani (Editor)

Similarity measurement method of case-based reasoning for conceptual cost estimation

Sae-Hyun Ji, Moonseo Park, Hyun-Soo Lee, and You-Sang Yoon
Seoul National University, Korea

Abstract

Case-based reasoning (CBR) method—which utilizes the knowledge gained from past experiences—can be viewed as an effective method for construction cost estimation. Even though decision-makers or users have not experienced all the cases of a database, this method makes it possible to retrieve the knowledge of similar cases, which are familiar to them, for new experiences. Thus, it is likely that the parties involved (i.e., the persuadees) will be more inclined to accept the estimation of a CBR method. Hitherto the CBR method having been applied already in construction cost estimation, generally estimations are rely on identifying and comparing the similar past cases with scope reflecting parameters. However, there are challenges related to the retrieval process that still having been discussed. One is the computation of similarity that becomes the most important in the retrieval process. And the other is how to assign the attribute weight values that make the most similar case identified by an index of the corresponding features. To address these issues, this research attempts to develop CBR in terms of a similarity measuring method based on the Euclidean distance concept and attribute weight value assigning method using genetic algorithm. For this purpose, we suggest the CBR model of building cost estimation and verify its effectiveness. Consequently, this research can provide a means of enhancing the accuracy of the cost estimation as well as a basis for the case retrieving fundamental, is relevant to both industry and academia.

Keywords: cost, case-based reasoning, genetic algorithms, similarity

1 Introduction

One of the purposes of estimation (e.g., of cost, schedule, or risk) is to persuade key decision-makers whether or not to initiate or continue a project. In this context, the case-based reasoning (CBR) method—which utilizes knowledge gained from past experiences—can be viewed as an effective method for estimation in construction Because, even though estimators who are not well-known this method could have their influence on persuasion increased, if they would have trust-worthy project data and use some of them with similar objectives which is the principle of case-base reasoning method. More exactly, data of case-based reasoning are composed of trust-worthy (i.e. actually implemented) project data; and the method makes it possible to retrieve the knowledge for new experiences based on similarity. Accordingly, decision-makers or users can experience indirectly all the cases in a database. Thus, it is likely that the parties involved (i.e., the persuadees) will be more willing to trust the estimation of a case-based reasoning (CBR) method regardless of speakers' reputation or background, and as opposed to "black box" machine learning algorithms such as neural networks. Often applied for construction cost estimation, CBR estimation generally relies on

identifying and comparing similar past cases with scope reflecting parameters (Ellsworth 1998; Hendrickson 2000). It has also been observed that CBR methods can increase the accuracy of construction cost estimates (Karshenas and Tse 2002; Chou and Loh 2006; Yi et al 2006; An et al. 2007). However, there are challenges related to the retrieval process that still need to be addressed. One issue is the computation of similarity, which is particularly important during the retrieval process. The effectiveness of a similarity measurement is determined by the usefulness of a retrieved case in solving a new problem. Therefore, establishing an appropriate similarity function is an attempt at handling the relationships between the relevant objects associated with the cases (Pal and Shiu 2004). A second challenge is how to assign the attribute weight values that enable the most similar case to be identified by an index of corresponding features. Nevertheless, most previous studies have not examined these two issues in detail.

In order to address these challenges, this paper develops a CBR cost estimate model for conceptual stages. This model utilizes the Euclidean distance concept for similarity measuring and genetic algorithms for attribute weight assigning. The research process is as follows. First, the scope of the cost model is defined as limited to the initial project stages (specifically budgeting) because early cost estimates are integral to an owner's decision to initiate construction projects and whether or not administrative organizations decide to participate (Seeley 1997). Then, data are collected with the assistance of a public housing company in Korea and converted into cost information and feature (attribute) data. Subsequently, a similarity measure method based on the Euclidean distance measuring concept, and an attribute weight assigning method based on genetic algorithm optimization, are introduced. The proposed model is developed based on these two concepts using an MS-EXCEL program. Finally, the model's effectiveness is validated by comparing it with models suggested in previous research. Consequently, this research can provide a means of enhancing accuracy of the cost estimation for industry practitioners as well as acting as a basis for further research on the fundamentals of case retrieval.

2 Preliminary Research

2.1 Case-based reasoning

CBR came from research into cognitive science: the work of Schank and Abelson in 1977 is widely held to be the origins of CBR. They propose that our general knowledge about situation is recorded in the brain as script that allows us to set up expectations and perform inferences (Watson 1997). CBR highly regarded as a plausible high-level model for cognitive processing. It was focused on problems such as how people learn a new skill and how humans generate hypotheses about new situation based on their past experiences (Pal and Shiu 2003). The processes of CBR can be seen as a reflection of a particular type of human reasoning that they generally solve the problems encounter with an equivalent of CBR. In these days, many studies in the construction domain related to CBR have been conducted for cost estimation purposes (Karshenas and Tse 2002; An et al. 2007; Yi 2006; Chou 2009), international market selection (Ozorhon et al. 2006), decision-making support (Chua et al. 2001; Morcous et al. 2002; Chua and Loh 2006), planning and management (Yau and Yang 1998; Tah et al. 1998), scheduling (Ryu et al. 2007), and predicting the outcome of litigation (Arditi and Tokdemir 1999). This method uses conceptually straightforward approaches to approximate real-valued or discrete-valued target functions. Its' learning algorithms consists of simply storing the presented training data. When a new query case encountered, a set similar related cases is retrieved from memory and used to classify the new query case (Mitchell 1997). In this context, establishing the computation of similarity can be a key issue for the whole CBR process.

2.2 Similarity concept in case-based reasoning

The meaning of similarity always depends on the underlying context of a particular application, and it does not convey a fixed characteristic that can be applied to any comparative context. In CBR, there are two major retrieval approaches (Liao et al. 1998). One is measuring case similarity by computing the distance between cases. The other approach is related more to the representational/indexing structures of the case which is more suitable for text-based case applications. On closer examination of the distance computation approach, the most common type of distance measure is based on the location of objects in Euclidean space (i.e., an ordered set of real numbers), where distance is calculated as the square root of the sum of the square of the arithmetical differences between two corresponding objects (Pal and Shiu 2004). In this respect, the nearest neighbours of an arbitrary case—which is the most basic algorithm for the description of relation between two cases—are defined as the standard Euclidean distance (Mitchell 1997).

2.3 Genetic algorithms

Genetic algorithms (GAs) are search algorithms based on the mechanics of natural selection and natural genetics. Genetic algorithms are iterative procedures which maintain a population of candidate solutions to optimize a fitness function (Sanchez et al 1997). Having been established as a valid strategy for problems requiring efficient and effective searching, GAs are used for widespread applications in business, scientific, and engineering circles, as they provide simplicity in computation and are powerful in their search for improvement (Goldberg 2006). GAs are used to search a space of candidate hypotheses to identify the best hypothesis. The best hypothesis is defined as the one that is the optimized value to the predefined numerical measure at hand, which is called hypothesis fitness (Mitchell 1997).

3 Model Development

As previously mentioned, in order to retrieve the most similar case, a similarity function should be employed and defined. This function can be used to distinguish how similarity is measured between two cases. In the literature, previously proposed similarity measuring functions are dichotomized into distance based similarity measuring concepts (Burkhard 2001; Ryu et al. 2007; Qian 2009) and direct similarity measuring concepts (Ozorhon et al. 2006; An et al. 2007; Chou 2009). Existing methods apply arithmetic summation of the weighted similarity scores of each input's attributes. As well, they obtain the distance divided by the attribute range for standardization that is based on the assumption of a linear relationship between the two cases, and all problems must be in the case base range. Thus, these methods are not supported by the Euclidean geometric aspect. Consequently, these techniques often lack explanation and are incomputable when the target case exists outside of the case base range. Such disadvantages can especially be found in the approaches of Burkhard (2001) and Qian (2009), who apply a fractional function of the relation between distance and similarity without mathematical or statistical proving. However, it should be noted that the distribution of all features cannot be represented by this fractional function.

In addition, previous approaches adopt several methods for weight value assigning. Yau andYang (1998) determined the weight values and adjusting factors heuristically. Arditi and Tokdemir (1999) used a feature counting and gradient decent method. An et al. (2007) introduced an analytical hierarchy process. Yet, despite new methods being continuously introduced, the challenge of determining the best method for the assignment of attributes' weight value in CBR still needs to be addressed.

3.1 Data analysis

All the case study data are actual cost data supplied by a public enterprise established by the Korean government. This data is used to construct the case base of the proposed CBR model. The data of 164 apartment buildings from 15 housing complex projects in Korea are utilized and organized into 164 cases. This data covers cost data from different construction users (104 cases from 2005, 28 from 2007, and 32 from 2009). The Korean government's historical cost index (KICT 2009) was used to normalize this data to year 2009. Although the data should be normalized in terms of escalation, regional location, and system specification, we only get the data normalized historically. This index is classified by 16 types of facilities that are officially announced every month. Due to Korea's relatively small territory, there is little point in normalizing the data for regional location and system specification.

Case storage is an important aspect of designing CBR systems in that it should reflect the conceptual view of what is represented in the case and take into account case characteristics (Watson 1997). Therefore, the potentially useful and predictive features of cases should be determined and extracted before building a CBR model. Through a comprehensive analysis of building drawings (see Table 1), 13 representative attributes are extracted and entered into the case library. These 13 attributes is used to assign the weight values of cases and to measure case similarity.

Table 1. Configuration of case features

Features	Feature type	Measurement scale
(X1) Number of households	Numeric	Integer
(X2) Gross floor area	Numeric	Real number
(X3) Number of unit floor households	Numeric	Integer
(X4) Number of elevators	Numeric	Integer
(X5) Number of floors	Numeric	Integer
(X6) Number of piloti with household scale	Numeric	Integer
(X7) Number of households of unit floor per elevator	Numeric	Real number
(X8) Height between stories	Numeric	Integer
(X9) Depth of pit	Numeric	Real number
(X10) Roof type	One of a list	Flat or Inclined (1 or 0)
(X11) Hallway type	One of a list	Hall or Corridor (1 or 0)
(X12) Structure type (RC)	One of a list	RC wall or RC column
(X13) Cost	Numeric	Real number

3.2 Optimization using GAs for assigning weight value

GAs are used to search a space of candidate hypotheses to identify the best hypothesis. The best hypothesis is defined as the one that is the optimized value to the predefined numerical measure at hand, which is called hypothesis fitness (Mitchell 1997). In order to make a hypothesis fitness function, this research assumes that the project cost of a specific case can be formulated by appropriately weighting its attributes.

$$C_i = \omega_1 X_{i1} + \omega_2 X_{i2} + \omega_3 X_{i4} + \cdots\cdots + \omega_1 X_{it} \quad (1)$$

Let C_j, ω_i, and Xi denote the cost of the jth case project, the weight value of ith attribute, and ith attribute value of jth case. When this relationship is expanded to a set of general cases, it is described by the matrix formula below.

$$\begin{pmatrix} X_{11} & \cdots & X_{i1} \\ \vdots & \ddots & \vdots \\ X_{j1} & \cdots & X_{ji} \end{pmatrix} \begin{pmatrix} \omega_1 \\ \vdots \\ \omega_i \end{pmatrix} = \begin{pmatrix} C_1 \\ \vdots \\ C_j \end{pmatrix} \tag{2}$$

Then, searching the optimal value of ω_i is conducted by minimizing the sum of the square root of the distance (i.e., Euclidean distance) between each side of the equation. This is because the solution that satisfies all the above equations probably does not exist, or, the equation is unsolvable. Thus, let D_j represent the distance, and then the hypothesis fitness function is defined as follows.

$$min \sum_{n=1}^{j} \sqrt{D_n^2} \quad , \quad s.t. \begin{pmatrix} C_1 \\ \vdots \\ C_j \end{pmatrix} - \begin{pmatrix} X_{11} & \cdots & X_{i1} \\ \vdots & \ddots & \vdots \\ X_{j1} & \cdots & X_{ji} \end{pmatrix} \begin{pmatrix} \omega_1 \\ \vdots \\ \omega_i \end{pmatrix} = \begin{pmatrix} D_1 \\ \vdots \\ D_j \end{pmatrix} \tag{3}$$

Before optimizing this function, normalization must first be conducted. Normalization, which converts raw values to the standard scores, requires selecting values that span one range and representing them in another range. The previously identified 13 types of attribute data, which all have a different data range, are converted to a scale of zero to one by applying a statistical standardization process. Based on the assumption that the data are approximated by the normal distribution, which is supported by the Central Limit Theorem, the distribution of the data is converted to a standard normal distribution, which has a mean of 0 and a standard deviation of 1. Its probability density function of the standard distribution $f(x \mid \mu, \sigma^2)$ is written by Eq. 6 (let μ and σ represent the sample mean and its standard deviation). This feature range assigning concept can mitigate or prevent a sudden feature shift which could distort the accuracy of the similarity measure.

$$f(X \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(X-\mu)^2}{2\sigma^2}\right] \tag{4}$$

Using the "MS-EXCEL STANDARDIZE" and "NORMSDIST" functions, the data are converted to standardized values and then the cumulative probability density of each value of attributes is computed. All the values are represented by a new score of zero to one. Based on the normalized data, the hypothesis fitness function for optimizing attribute weight value is executed. As a result, the GA optimized weight values are assigned using "EVOLVER 4.0" by setting the condition of zero to one for the adjusting cell (attribute weight) range, 0.1 for the crossover rate, and 0.05 for the mutation rate.

3.3 Similarity measure

The computation of similarity is an important issue for the case retrieving process in CBR. An appropriate similarity function needs to be developed to handle the hidden relationships between the objects associated with cases (Burkhard 2001). The most common distance measuring method is based on the location of objects in Euclidean space, in which the distance is calculated as the square

root of the sum of the arithmetical differences between the corresponding coordinates of two objects (Pal and Shiu 2004). Specifically, CBR methods, which use more complex, symbolic representations (e.g., assume instance), can be presented as points in Euclidean space (Mitchell 1997). More formally, the weighted Euclidean distance is defined by the equation. Let an arbitrary case x be described by the feature vector:

$$[a_1(x), a_2(x), \ldots, a_n(x)] \tag{5}$$

where $a_r(x)$ denotes the value of the r^{th} attributes of case x, and w_r denotes the weight of the attributes of the case. Then, the weighted distance between the two cases x_i and x_j is defined as DIS (x_i, x_j) as seen below (Pal and Shiu 2004, Eq. 8).

$$DIS(x_i, x_j) = \sqrt{\sum_{r=1}^{n} w_r^2 (a_r(x_i) - a_r(x_j))^2} \tag{6}$$

Because all the attributes' values are converted to new scores of zero to one applied by the probability density function, as previously mentioned, when the square root of the sum of squares of the weight values is assigned as one ($\sum w_r^2 = 1$), the range of the weighted distance of the two cases can be standardized by [0, 1]. Therefore, the axiom of reflexivity for the distance measure, as well as for the similarity measure SIM, where SIM (x_i, x_j) stands for the degree of similarity between x_i and x_j (Burkhard 2001), comes into existence as follows.

$$x_i = x_j \rightarrow SIM(x_i, x_j) = 1 \text{ and } DIS(x_i, x_j) = 0 \tag{7}$$

Based on this concept, it can be assumed that similarity and distance are in linear inverse proportion to each other. Accordingly, the relation of similarity and distance is defined as below.

$$SIM(x_i, x_j) = 1 - DIS(x_i, x_j) = 1 - \sqrt{\sum_{r=1}^{n} w_r^2 (a_r(x_i) - a_r(x_j))^2} \tag{9}$$

To facilitate the similarity function of Eq. 10, the GA optimized attribute weight values should be converted to the new score which satisfies the sum of squares of all the values being one. Based on the previously discussed concept, the similarity index is defined as follows.

$$\text{Similarity Index } (\%) = \left[1 - \sqrt{\frac{\sum_{r=1}^{n} w_r^2 [a_r(x_i) - a_r(x_j)]^2}{\sum_{r=1}^{n} w_r^2}}\right] \times 100 \tag{10}$$

4 Model Validation

So far, this research has introduced an attribute weight assigning method that deploys genetic algorithm optimization based on statistical normalization, and a similarity scoring method based on

the Euclidean distance concept. As these methods specifically target the case retrieval process, the validity of this method can be evaluated by comparing the results related to the reuse of retrieved cases. As already noted, many other researchers have also suggested and adopted different methods pertaining to these issues. Accordingly, a comparative experiment was designed to test the validity of the proposed CBR cost estimate model in terms of its effectiveness in weight assigning and similarity scoring.

Table 2. Profile of Cases for Model Validation

Case	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	Cost
Case 1	7	735	1	0.5	7	-	2	2.9	5.2	-	1	2.38	404,340
Case 2	7	990	1	0.5	7	-	2	2.9	5.2	-	1	2.38	547,492
Case 3	14	1,508	2	1	7	-	2	2.9	5.2	-	1	2.57	1,006,531
Case 4	26	2,188	2	1	14	2	2	2.9	8.7	1	1	4.7	1,134,424
Case 5	24	1,994	2	1	13	2	2	2.8	9.36	-	1	5.1	1,223,354
Case 6	38	3,185	4	2	10	2	2	2.9	8.7	1	1	4.25	2,038,465
Case 7	50	3,753	8	1	7	6	8	2.9	5.2	-	-	2.22	2,045,693
Case 8	63	4,531	5	1	13	2	5	2.8	5.85	1	-	4.74	2,591,614
Case 9	40	4,448	4	1	11	4	4	2.8	5.85	1	-	4.47	2,894,608
Case 10	60	4,703	4	1	15	-	4	2.9	8.7	1	1	5.84	2,924,845
Case 11	24	2,635	2	1	12	-	2	2.8	9.36	-	1	5.98	2,951,709
Case 12	54	4,472	4	2	15	4	2	2.8	5.85	-	1	5.4	2,983,510
Case 13	44	4,908	4	2	12	4	2	2.8	5.85	-	1	4.29	3,110,324
Case 14	48	5,320	4	1	13	4	4	2.8	5.85	1	-	5.42	3,334,662
Case 15	56	6,189	4	1	15	4	4	2.8	8.6	1	-	6.65	3,682,360
Case 16	50	7,199	4	2	14	2	2	2.9	8.7	-	1	4.11	3,806,177
Case 17	50	7,115	4	2	14	8	2	2.8	5.85	-	1	4.37	4,196,711
Case 18	50	7,121	4	2	14	4	2	2.8	5.85	-	-	4.68	4,398,362
Case 19	58	8,218	4	2	15	-	2	2.8	5.85	-	1	4.75	4,660,579
Case 20	52	5,811	4	2	14	4	2	2.9	8.7	-	1	4.65	4,707,437

(X1) Number of households, (X2) Gross floor area, (X3) Number of unit floor households, (X4) Number of elevators, (X5) Number of floors, (X6) Number of piloti with household scale, (X7) Number of households of unit floor per elevator, (X8) Height between stories, (X9) Depth of pit, (X10) Roof type, (X11) Hallway type, (X12) Structure type

First, twenty cases were selected randomly from the case base, and these were excluded from the case base of the CBR model. The profile of these twenty cases is shown in Table 2. Then, the effectiveness of the suggested cost model was compared (to other models (permutation of three similarity and three weighting methods) in terms of estimation accuracy using the k-nearest neighbor principle. This concept, which is based on the Euclidean distance measure method, involves searching for the k nearest cases to the current input case using a distance measure and then selecting the class of the majority of these k cases as the retrieval case (Pal and Shiu 2004). In this respect, the first similarity score case base (one nearest neighbour, One-NN) approach and the ten higher rank similarity score case base (ten nearest neighbours, Ten- NN) approach are utilized for estimating building cost. Simultaneously, nine different types of case-based reasoning models, which are dependent on the combination of weight value assigning methods and similarity functions, are defined. S0, S1, and S2 respectively denote the similarity function proposed in this research, the arithmetic summation based function, and the fractional function based similarity measure function. W0, W1, and W2 refer respectively to the weight value assigning method proposed in this research, feature counting, and the method utilizing the standardized coefficient of multiple regression analysis (CMRA) (Table 3). The weight value of each attribute was computed using the genetic algorithm optimization process and the SPSS linear regression function (Table 4).

Table 3. Model Combinations for the Validity Test

Similarity measuring methods	Weight value assigning methods		
	Genetic Algorithm (W0)	Feature counting (W1)	CMRA (W2)
Euclidean distance base(S_0) $= \left[1 - \sqrt{\frac{\sum w_T^2 (a_T(x_1) - a_T(x_1))^2}{w_T^2}}\right] \times 100$	S_0W_0	S_0W_1	S_0W_2
Arithmetic summation (S_1) $= \left[w_T \sum 1 - \sqrt{\frac{(a_T(x_1) - a_T(x_T))^2}{(a_{T,max} - a_{T,min})^2}}\right] \times 100$	S_1W_0	S_1W_1	S_1W_2
Fractional function base (S_2) $= \left[w_T \sum 1 / \left\{1 + \sqrt{\frac{(a_T(x_1) - a_T(x_T))^2}{(a_{T,max} - a_{T,min})^2}}\right\}\right] \times 100$	S_2W_0	S_2W_1	S_2W_2

Table 4. Weight Values Applying GA, Feature Counting, and MRAC

Attributes	GA(W0)	Feature counting (W1)	CMRA (W2)
(X1) Number of households	0.0193	0.0833	0.0073
(X2) Gross floor area	0.3613	0.0833	0.5517
(X3) Number of unit floor households	0.1760	0.0833	0.0964
(X4) Number of elevators	0.0041	0.0833	0.0798
(X5) Number of floors	0.2924	0.0833	0.0304
(X6) Number of piloti with household scale	0.0067	0.0833	0.0273
(X7) Number of households of unit floor per elevator	0.0405	0.0833	0.0652
(X8) Height between stories	0.0017	0.0833	0.0544
(X9) Depth of pit	0.0057	0.0833	0.0085
(X10) Roof type	0.0073	0.0833	0.0449
(X11) Hallway type	0.0242	0.0833	0.0311
(X12) Structure type	0.0608	0.0833	0.0030

4.1 Analysis and Results

As summarized in Table 5, it is identified that different cases are retrieved according to applying both weight value assigning and similarity scoring methods. On closer examination, the impact of similarity measuring methods on cost estimate error rate is more influential than weight value assigning methods. The average difference of estimate error rate according to weighting values methods is 11.60% / 8.17 % (One-NN / Ten-NN) whereas the similarity is 0.27% / 1.47% (One-NN / Ten-NN).

As a result, it is identified that different cases are retrieved according to applying both weight value assigning and similarity scoring methods. More precisely, when testing the model's effectiveness in respect to the weight assigning method, it was determined that the mean percentage of error of the proposed genetic algorithm optimization (W0) model is lower than all its counterparts in terms of both the One-NN and Ten-NN approach. Moreover, this method yielded better results regardless of the combination of similarity scoring methods. Models which utilize the proposed attribute weighting method (W0) with the One-NN and Ten-NN approaches have error rates of 9.01% to 10.66% and 10.78% to 11.59%, respectively. On the other hand, when the feature counting method (W1) based models are used with the One-NN and Ten-NN approaches, the error rates are 20.77% to 22.63 and 18.24% to 20.85%, respectively, while the standardized coefficient of multiple regression analysis (CMRA) method (W2) base models have 10.75% to 11.01% and 11.24% to 12.58% error

rates. In terms of the effectiveness of the similarity measuring method, the proposed similarity measuring method (S0) based model combined with GA optimized (W0), feature counting (W1), and CMRA (W2), have error rates of 9.01%, 22.63%, and 11.01% with One-NN, respectively, and 10.78%, 18.24% and 11.70% with Ten-NN, respectively. Whereas, the arithmetic similarity method (S1) based models have error rates of 10.02%, 21.12%, and 10.75% with One-NN, and 11.59%, 20.85%, and 12.58% with Ten-NN, while the fractional function base models have 10.66%, 20.73%, and 10.75% with One-NN, and 11.20%, 18.96%, and 11.24% with Ten-NN. Accordingly, it is difficult to compare the effectiveness of the similarity measuring method of these models, although the suggested method (S0) yields a better result in the case of Ten-NN, regardless of the weighting method. Apparently, the model combining normal distribution based genetic algorithm optimization and the similarity scoring method with the Euclidean is the most accurate with both One-NN and Ten-NN (Table 4). Additionally, after conducting a one-way ANOVA procedure, a significant difference between these nine models was not detected.

Table 5. Analysis of Mean of Error Rate

Weighting method	One-NN					Ten-NN				
	W0	W1	W2	Mean (Si is Fixed)	Difference	W0	W1	W2	Mean (Si is Fixed)	Difference
S0	9.00	22.60	11.00	14.20		10.80	18.20	11.70	13.57	
S1	10.00	21.10	10.70	13.93	0.27	11.60	20.90	12.60	15.03	1.47
S2	10.70	20.80	10.70	14.07		11.20	19.00	11.20	13.80	
Mean (Wi is Fixed)	9.90	21.50	10.80	14.07	11.60	11.20	19.37	11.83	14.13	8.17

W0 : suggested weighting method, W1: feature counting, W2: Coefficient of multiple regression analysis,
S0: suggested similarity measuring method, S1: Arithmetic summation, S2: fractional functions

5 Conclusion

Construction project cost estimates can be used to persuade management personnel (i.e., owners and decision-makers) to initiate or continue a project. In this context, a case-based reasoning cost model can be an effective cost estimation method, as it is based on identifying the characteristics of cases. CBR methods utilize familiar knowledge to tackle new experiences. However, challenges related to similarity measurement and attributes weight assignment issues still need to be addressed to enhance the reliability of CBR models. Specifically, existing measurement methods are based on arithmetic summation of the weighted features or geometrically unexplainable distance measure based or applied by fractional functions. Thus, similarity cannot be computed when the target case exists outside of the case base range, while limitations of representation exist with fractional functions. As well, despite the fact that various weight assigning methods continue to be proposed, No one knows what method is the best and what might be.

As an effort to address these challenges, this research developed methods for similarity measuring and weight value assigning for CBR modeling, and suggested a CBR cost estimation model for the budgeting of apartment buildings in Korea. After the validation, which evaluated the model in terms of the similarity scoring and attributes weight assigning methods, it was confirmed that these methods can enhance the accuracy of cost estimation. In fact, when combined with the suggested methods (i.e., the Euclidean distance based similarity function and weight values optimized by genetic algorithms), the proposed model was found to be superior to its model counterparts. Ultimately, this research demonstrated that the proposed model can be an effective budgeting tool during the initial project stages, providing the iterative function of cost check and control, which responds to project changes. Finally, although the model was developed and verified using apartment buildings in Korea, the

research findings can also be usefully applied to different types of construction after customization, while also contributing to the enhancement of cost estimation research.

As a result of this research, effort toward developing CBR by suggesting new methods in terms of similarity measure and attributes weighting is initiated. While it must be noted that this research is based on limited case data, and that additional research and testing must be conducted to further validate the model, generalize the effect of suggested methods.

Acknowledgment

This research was supported by a grant (R&D06CIT-A03) from the Innovative Construction Cost Engineering Research Center; and a grant (05CIT-01) from the Construction Technology Innovation Program funded by the Ministry of Land, Transport and Marine Affairs (Government of Korea).

References

ADMODT, A. and PLAZA, E. 1994. "Case-based reasoning: Foundational issues, methodological variations and system approaches", *AI Communications*, 7(1), pp. 35~39.

AHN, H. et al. 2006. "Global Optimization of Feature Weights and the Number of Neighbors that Combine in a Case-Based Reasoning System", *Expert System*, 23(5), pp. 290~301.

AN, S-H. et al. 2007. "A Case-Based Reasoning Cost Estimating Model Using Experience by Analytic Hierarchy Process", *Building and Environment*, 42, pp. 2573~2579.

ARDITI, D. and TOKDEMIR, O. B. 1999. "Using Case-Based Reasoning to Predict the Outcome of Construction Litigation", *Computer-Aided Civil and Infrastructure Engineering*, 14, pp. 385~393.

BURKHARD, H-D. 2001. *"Similarity and Distance in Case-based Reasoning", fundamenta Informaticea,* 47, pp. 201~215.

CHOU, J-S. 2009. "Web-Based CBR System Applied to Early Cost Budgeting for Pavement Maintenance Project", *Expert System with Applications,* 39, pp. 2947~2960.

CHUA, D. K. H. and Loh P. K. 2006. "CB-Contract: Case-Based Reasoning Approach to Construction Contract Strategy Formulation", *Journal of Computing in Civil Engineering*, 20(5), ASCE, pp. 339~350.

CHUA, D. K. H. et al. 2001. "Case-Based Reasoning Approach in Bid Decision-making", *Journal of Construction Engineering and Management*, 127(1), ASCE, pp. 35~45.

DUSSART, C. et al. 2008. "Optimizing Clinical Practice with Case-Based Reasoning Approach", *Journal of Evaluation in Clinical Practice*, 14, pp. 718~720.

ELLSWORTH, R. K. 1998. "Cost-to-Capacity Analysis for Estimating Waste-to-Energy Facility Costs", *Cost Engineering,* 40(6), AACE, pp. 27~30.

FUNK, P. and XIONG, N. 2006. "Case-Based Reasoning and Knowledge Discovery in Medical Applications with Time Series", *Computational Intelligence*, 22(3/4), pp. 238~253.

GOLDBERG, D. E. 2006. *Genetic Algorithems in Search, Optimization and Machine Learning, Addison-Wesley*.

HENDRICKSON, C. 2000. *Project Management for Construction 2.2nd Ed., World Wide Web Publication,* http://pmbook.ce.cmu.edu/05_Cost_Estimation.html.

JARDE, A. and ALKASS, S. 2007. "Computer-Integrated System for Estimating the Costs of Building Project", *Journal of Construction Engineering and Management*, 13(4), ASCE, pp. 205~223.

KARSHENAS, S. and TSE, J. 2002. "A Case-Based Reasoning Approach to Construction Cost Estimating", *Information Technology 2002*, Computing In Civil Engineering, pp. 113~123.

KOLODNER J. 1993. *Case-Based Reasoning, Kaufmann.*

Korea Institute of Construction Technology. 2009. Construction Cost Index July 2009, http://www.kict.re.kr/division/pds_list.asp?dept_code=31200

KRAFT, D. H. et al. 1997. "Genetic Algorithms for Query Optimization in Information Retrieval: Relevance Feedback", Advanced in Fuzzy Systems Applications and Theory, vol. 7, pp.155~174.

LOPEs, H. S. et al. 1997. "An Evolutionary Approach to Simulate Cognitive Feedback Learning in Medical Domain", *Advanced in Fuzzy Systems Applications and Theory*, vol. 7, pp.193~208.

SEELEY, I. H. 1997. *Quantity Surveying Practice 2nd Ed., Macmillan Press*.

MITCHELL, T. M. 1997, *Machine Learning, Mcgraw Hill.*

MORCOUS, G. et al. 2002. “Case-Based Reasoning System for Modeling Infrastructure Deterioration”, *Journal of Computing in Civil Engineering*, 16(2), ASCE, pp. 104~114.

OBERLENDER, G. D. and TROST, S. M. 2001. “Predicting Accuracy of Early Cost Estimates Based on Estimate Quality”, *Journal of Construction Engineering and Management*, 127(3), ASCE, pp. 173~182.

OZORHON, B. et al. 2006. “Case-Based Reasoning Model for International Market Selection”, *Journal of Construction Engineering and Management*, 132(9), ASCE, pp. 940~948.

PAL, S. K. and SHIU, S. C. K. 2004. Foundations of Soft Case-Based Reasoning, Wiley Interscience.

PARK, Y-J. et al. 2006. “New Knowledge Extraction Technique Using Probability for Case-Based Reasoning: Application to Medical Diagnosis”, *Expert System*, 23(1), pp. 2~20.

QIAN, Z. et al. 2008. “A Case-Based Approach to Power Transformer Fault Diagnosis Using Dissolved Gas Analysis Data”, *European Transactions on Electrical Power*.

RYU, H-K et al. 2007. “Construction Planning Method Using Case-Based Reasoning (CONPLA-CBR)”, *Journal of Computing in Civil Engineering*. 21(6), ASCE, pp. 410~422.

SMITH, A. J. 1995. *Estimating, Tendering and Bidding for Construction, Macmillan, London.*

STIFF, J. B. and MONGEAU, P. A. 2002. *Persuasive communication, 2nd edition, Guilford Press*

SUN, B. et al. 2003. “Scenario-Based Knowledge Representation in Case-Based Reasoning Systems”, *Expert System*, 20(2), pp. 92~99.

TAH, J. H. M. et al. 1998. “An Application of Case-Based Reasoning to the Planning of Highway Bridge Construction”, *Engineering, Construction and Architectural Management*, vol. 4, pp. 327~338.

WATSON, I. 1997. *Applying Case-Based Reasoning: Techniques for Enterprise System, Morgan Kaufmann Publishers.*

YAU, N-J. and YANG, J-B. 1998. “Case-Based Reasoning in Construction Management”, *Computer Aided Civil and Infrastructure Engineering*, Vol. 13, pp. 143~150.

YI, J. et al. 2006. “A Study on Case-Based Forecasting Model for Monthly Expenditures of Residential Building Project”, *Korean Journal of Construction Engineering and Management*, 79(1), KICEM, pp. 128~137.

ZHUANG, Z. Y. et al. 2007. “Combining Data Mining and Case-Based Reasoning for Intelligent Decision Support for Pathology Ordering by General Practitioners”, *European Journal of Operational Research*, 195, pp. 662~675.