Cost estimation model for building projects using case-based reasoning

Sae-Hyun Ji, Moonseo Park, and Hyun-Soo Lee

Abstract: The case-based reasoning (CBR) method can be an effective means of utilizing knowledge gained from past experiences to estimate cost in construction. It has also been observed that CBR can enhance the accuracy of construction cost estimates. However, there are challenges related to the process of retrieving knowledge and information that still need to be addressed. One challenge is the computation of similarity and another is the assignment of the attribute weight values. To address these challenges, this paper develops a CBR cost estimate model for building projects using a Euclidean distance concept and genetic algorithms. Consequently, it was found that this model can enhance the accuracy of cost estimation and act as a basis for further research on the fundamentals of the case-based reasoning method.

Key words: case-based reasoning, cost, estimate, genetic algorithm.

Résumé : La méthode du raisonnement par cas peut représenter un moyen efficace d'utiliser les connaissances acquises des expériences passées pour estimer le coût de la construction. Il a également été remarqué que la méthode du raisonnement par cas peut améliorer la précision des estimations des coûts de construction. Toutefois, certains défis reliés au processus de récupération des connaissances et de l'information doivent toujours être abordés. L'un d'eux est le calcul de la similitude et l'autre est l'assignation des valeurs de poids d'attribut. Pour aider à relever ces défis, le présent article développe un modèle d'estimation des coûts par la méthode du raisonnement par cas pour les projets de construction utilisant le concept de la distance euclidienne et des algorithmes génétiques. Il a été découvert que ce modèle peut améliorer la précision de l'estimation des coûts et servir de base pour une future recherche sur les fondements de la méthode de raisonnement par cas.

Mots-clés : raisonnement par cas, coût, estimation, algorithme génétique.

[Traduit par la Rédaction]

Introduction

One of the purposes of estimation (e.g., of cost, schedule, or risk) is to persuade key decision-makers whether to initiate or continue a project. However, references regarding effective methods for persuading decision-makers date back to the time of the ancient Greeks. Aristotle proposed that an audience is more likely to be persuaded by a speaker whose characteristics they understand (ethos). More recently, Stiff and Mongeau (2002) observed that an audience will be more inclined to understand and be persuaded by a more familiar and esteemed source. 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. Even estimators who are not well known could be more persuasive if they had trust-worthy project data and used them with similar objectives, which is the principle of the case-based reasoning method. More exactly, data of case-based reasoning are composed of trustworthy (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 CBR method regardless of speakers' reputation or background and as opposed to "black box" machine learning algorithms, such as neural networks.

Often applied to construction cost estimation, CBR estimation generally relies on identifying and comparing similar past cases within 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; Chua and Loh 2006; Yi 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 to handle 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.

Received 21 September 2009. Revision accepted 14 February 2011. Published at www.nrcresearchpress.com/cjce on 17 May 2011.

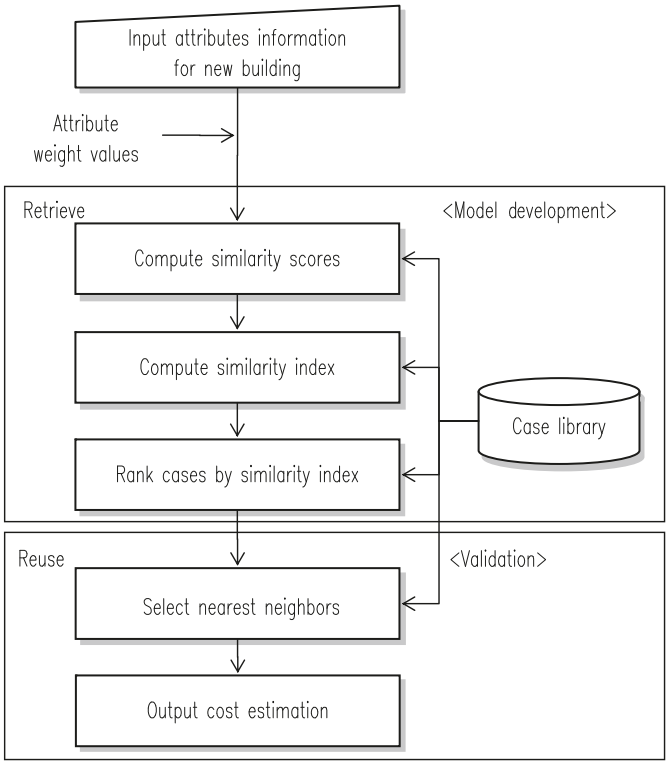
S.-H. Ji, M. Park, and H.-S. Lee. Department of Architecture, Seoul National University, San 56-1 Shinrim-dong, Seoul, Korea.

Corresponding author: Moonseo Park (e-mail: mspark@snu.ac.kr).

Written discussion of this article is welcomed and will be received by the Editor until 30 September 2011.

To address these challenges, this paper develops a CBR cost estimate model for building projects. This model utilizes the Euclidean distance concept for similarity measuring and genetic algorithms for attribute weight assignment. Moreover, we try to improve the explanatory power of case distribution by approximating the case data to a standard normal distribution to mitigate the negative effects of output distortion provoked by the sudden change of data features. 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 assignment method, based on genetic algorithm optimization, are introduced (Fig. 1). The proposed model was developed based on these two concepts using Microsoft 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 the accuracy of the cost estimation for industry practitioners as well as acting as a basis for further research on the fundamentals of case retrieval.

Fig. 1. Process of the case-based reasoning (CBR) cost estimation model.



Preliminary research

Case-based reasoning

Instance-based methods, such as CBR, store a set of training examples that are generalized when a new instance must be classified (Burkhard 2001). Each time a new query instance is encountered, its relationship to the previously stored examples is examined to assign a target function value for the new instance. The basic idea behind CBR is the hypothesis that similar problems have similar solutions. An aim of constructing a case-based system is to use the notion of similarity that best fits with this hypothesis (Burkhard 2001). Generally, the CBR problem-solving process has four steps: (1) retrieve, (2) reuse, (3) revise, and (4) retain (Aamodt and Plaza 1994). Broadly applied across industries, case-based reasoning has been utilized for medical knowledge discoveries (Funk and Xiong 2006; Park et al. 2006; Dussart et al. 2008; Zhuang et al. 2007), managerial decision support (Sun et al. 2003; Ahn et al. 2006), healthcare management (Huang et al. 2007), educational application (Han et al. 2005), and diagnostics of power transformer faults (Qian et al. 2008).

In an experience-oriented industry, such as construction, knowledge and assessments of previous projects are essential for resolving reoccurring problems. For that reason, the case-based reasoning method is gaining recognition as a decision-making tool for the construction industry. Recently, many studies in the construction domain related to CBR have been conducted for construction cost estimation (Yau and Yang 1998; Karshenas and Tse 2002; Yi 2006; An et al. 2007; Doğan et al. 2008; Chou 2009; Koo et al. 2010*a*, 2010*b*), international market selection (Ozorhon et al. 2006), decision-making support (Chua et al. 2001; Morcous et al. 2002; Chua and Loh 2006; Dikmen et al. 2007), planning and (or) scheduling (Tah et al. 1998; Yau and Yang 1998; Ryu et al. 2007; Koo et al. 2010*a*), safety hazard identification (Goh and Chua 2010), and predicting the outcome of litigation (Arditi and Tokdemir 1999). Most of these researches emphasized the case retrieval method, which is the kernel of CBR. In this context, we again analyzed the aforementioned literatures and then summarized these according to discipline, objective, number of cases for model building, number of attributes, attributes weighting method, and similarity function, as shown in Table 1.

Similarity concept in case-based reasoning

The concept 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 approach is measuring case similarity by computing the distance between the cases. The other approach is related more to the representational or 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 neighbors 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).

Table 1. Summary of case-based reasoning applications.

Researcher	Discipline	Objective	Number of cases for model building	Number of attributes	Attributes weighting method	Similarity function
Yau and Yang (1998)	Construction management	Cost and duration estimation for a building project	60 (hypothetical)	10	Manual	$\sum_{n=1}^{i} \varpi_n \left\lvert \frac{a_n(x_a)-a_n(x_b)}{a_n(x_b)} \right\rvert \times 100$, $\left\lvert \frac{a_n(x_a)-a_n(x_b)}{a_n(x_b)} \right\rvert \times 100 \leq 10\%$ then 0, else 1
Arditi and Tokdemir (1999)	Construction management	Construction litigation outcome prediction	102	38	Gradient descent, manual; ID3, feature counting	$\sum_{n=1}^{i} \varpi_n \left\lvert \frac{a_n(x_a)-a_n(x_b)}{a_n(x_b)} \right\rvert \times 100$ $\left\lvert \frac{a_n(x_a)-a_n(x_b)}{a_n(x_b)} \right\rvert \times 100 \leq P\%$ then 0, else 1
Han et al. (2005)	Education	Development of a case-based tutoring system	20	7	Manual	Not described
Ozorhon et al. (2006)	Construction management	International market selection	600	19	Feature counting, ID3; gradient descent, manual	Not described (weighted feature counting)
Doğan et al. (2006)	Construction management	Cost of structural system estimation	29	8	GA, feature counting, gradient descent	$\sum_{n=1}^{i} \varpi_n \frac{\min\left(a_n(x_a), a_n(x_b)\right)}{\max\left(a_n(x_a), a_n(x_b)\right)}$
Dikmen et al. (2007)	Construction management	Bid mark-up estimation	95	33	Gradient descent, weighted gradient descent	$\sum_{n=1}^{i} \varpi_n \left\lvert \frac{a_n(x_a)-a_n(x_b)}{a_{n,\max}-a_{n,\min}} \right\rvert$
Ahn et al. (2006)	customer relationship management	classification of customers' buying behavior for a specific product	980	14	Genetic algorithms	Not described (weighted average of Euclidean distance)
Ryu et al. (2007)	Construction management	Construction schedule generation	Not described	12	Manual	$\sum_{n=1}^{i} \varpi_n \left\{ 1 - \sqrt{\frac{\left(a_n(x_a)-a_n(x_b)\right)^2}{(a_{n,\max}-a_{n,\min})^2}} \right\}$
An et al. (2007)	Construction management	Construction cost estimation	540	9	AHP; Feature counting, gradient descent	$\sum_{n=1}^{i} \varpi_n \left\lvert \frac{a_n(x_a)-a_n(x_b)}{a_n(x_b)} \right\rvert \times 100$, $\left\lvert \frac{a_n(x_a)-a_n(x_b)}{a_n(x_b)} \right\rvert \times 100 \leq P\%$ then 0, else 1
Huang et al. (2007)	Healthcare management	Chronic disease diagnosis and treatments supporting	15 751	11	Manual	$\frac{\sum_{n=1}^{i} \varpi_n Z_n}{\sum_{n=1}^{i} \varpi_n}$, if $a_n(x_a) < a_n(x_b)$ then $Z_n = \frac{a_n(x_a)}{a_n(x_b)}$ else $\frac{a_n(x_b)}{a_n(x_a)}$
Qian et al. (2008)	Electrical power	Power transformer fault diagnosis	798	7	Feature counting	$\sum_{n=1}^{i} \varpi_n \left\{ 1/\left(1 + \sqrt{\frac{\left(a_n(x_a)-a_n(x_b)\right)^2}{(a_{n,\max}-a_{n,\min})^2}} \right) \right\}$
Doğan et al. (2008)	Construction management	Cost of structural system estimation	24	8	Decision tree (3 types)	$\sum_{n=1}^{i} \varpi_n \frac{\min\left(a_n(x_a), a_n(x_b)\right)}{\max\left(a_n(x_a), a_n(x_b)\right)}$

Table 1 (*concluded*).

Researcher	Discipline	Objective	Number of cases for model building	Number of attributes	Attributes weighting method	Similarity function
Chou (2009)	Construction management	Pavement maintenance project cost budgeting	300	7	AHP	$\sum_{n=1}^{i} \varpi_n \frac{\min\left(a_n(x_a), a_n(x_b)\right)}{\max\left(a_n(x_a), a_n(x_b)\right)}$
Koo et al. (2010*a*)	Construction management	Cost and schedule estimation	100	14	Feature counting, ANN coefficient of multiple regression analysis	$\sum_{n=1}^{i} \varpi_n \frac{\min\left(a_n(x_a), a_n(x_b)\right)}{\max\left(a_n(x_a), a_n(x_b)\right)}$
Koo et al. (2010*b*)	Construction management	Cost estimation	23	27	GA, feature counting	$\sum_{n=1}^{i} \varpi_n \left\lvert \frac{a_n(x_a) - a_n(x_b)}{a_n(x_b)} \right\rvert \times 100$, $\left\lvert \frac{a_n(x_a) - a_n(x_b)}{a_n(x_b)} \right\rvert \times 100 \leq P\%$ then 0, else 1

Note: $a_n(x_a)$, ath attribute nth value of compared case; $a_n(x_b)$, ath attribute nth attribute value of problem case; $a_{n\,\max}$, the highest ath attribute nth attribute value in case library; $a_{n\,\min}$, the lowest ath attribute nth attribute value in case library.

Genetic algorithms

Inspired by the processes of biological evolution, genetic algorithms (GAs) provide an approach for learning methods by generating successor hypotheses through iterative mutation and crossover (Mitchell 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). In GAs, the hypothesis fitness function is the criterion for ranking potential hypotheses; therefore, GAs can be used to evaluate all members of a population. This has been demonstrated in past research. For example, Ahn et al. (2006) introduced a genetic algorithm to simultaneously optimize the number of neighbors and the weight of attributes whose fitness function is a set weight to maximize the outputs. Moreover, genetic programming is useful for satisfying owners' needs, such as information retrieval (Kraft et al. 1997), medical feedback learning (Lopes 1997), robot control, and the recognition of objects in visual scenes (Mitchell 1997).

Model development

As previously addressed, 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; Ahn et al. 2006; Ryu et al. 2007; Qian et al. 2008) and direct similarity measuring concepts (Yau and Yang 1998; Arditi and Tokdemir 1999; Ozorhon et al. 2006; An et al. 2007; Dikmen et al. 2007; Huang et al. 2007; Doğan et al. 2008, Chou 2009; Koo et al. 2010*a*, 2010*b*). The first group of methods applies arithmetic summation of the weighted similarity scores of each input's attributes. Then, the distance is divided by the attribute range for standardization. This process is based on the assumption of a linear relationship between the two cases, so all problems must be in the case-base range. As well as, Burkhard (2001) and Qian et al. (2008) used a modified fractional function as a weighted similarity measurement equation without mathematical or statistical proof. However, it should be noted that the distribution of all features cannot be represented by this fractional function. Also, to normalize the output value to [0, 1] range, all of these similarity functions use the different degree between the maximum and minimum or make the higher value between the pair; and is used as a denominator. These methods have an assumption that the distribution of attributes is linear. In particular, the formula employed by Ryu et al. (2007) calculates the difference between two coordinates of an attribute as a relative distance to the corresponding data range [0, 1]. Then, attribute similarity is calculated by subtracting this from one. Therefore, these relevant functions cannot reflect the sudden change of case distribution (i.e., feature shift). Moreover, they 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.

On the other hand, CBR similarity computation is generally described by forming of multiplication between an attributes difference of objects and its weight. Thus, the assignment of weights is also important to complete the similarity function. Regarding this, previous approaches have adopted several methods for weight value assignment. Yau and Yang (1998) determined the weight values and adjusting factors heuristically. Arditi and Tokdemir (1999) used a feature counting and gradient descent method. An et al. (2007) introduced an analytical hierarchy process. Doğan et al. (2006) compared the performance of three optimization techniques, feature counting, gradient descent, and genetic algorithms. Koo et al. (2010*a*) introduced weighting methods by applying coefficients of multiple regression analysis, and artificial neural network. Yet, despite new CBR retrieving methods being continuously introduced, most researches adapted the weighting methods of others. In this context, the challenge of determining the best method for the assignment in CBR still needs to be addressed.

Model scope

Performance and overall project success are often measured by how well the actual cost compares to the early cost estimates (Oberlender and Trost 2001). Initial cost estimates are the basis for the release of funds for further studies of estimates and become the marker against which all subsequent estimates are compared (Smith 1995). In other words, initial estimation produces a forecast of the probable cost of a future project before execution. Thus, the development of the proposed cost model is focused on preparing a budget for the initial stages of construction projects.

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, covering 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, the data was only normalized historically. This index is classified according to 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 should 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. To determine the impact factors, the following process is followed. First, based on previous research review and expert interview, a pool of cases is constructed. Next a comprehensive analysis of the sample cases is performed to reduce the number of factors. Finally, the remaining factors are confirmed by the experts. Consequently, 13 representative attributes are extracted and entered into the case library (Table 2). These 13 attributes are used to assign the weight values of cases and to measure case similarity.

In this research, the similarity of numeric scale data is calculated based on Euclidean distance. Generally, similarity of nominal scale data is assigned to the logic values of "true" (i.e., a perfect match) or "false" (i.e., not a perfect match), or to the defined degree values of similarity between each possible pair of attributes (Kolodner 1993). However, the former local monotonicity axiom-based method is likely to yield a deficient explanation of the relationships for the dichotomized hard data. On the other hand, the latter similarity matrix-based method can be intuitive, particularly in terms of selecting one out of over three feature types of nominal data. For example, in Table 2, roof type ($X10$, flat or not) and hallway type ($X11$, hall or not), which are the binary data among the nominal types, are assigned the similarity score of one for "true" and zero for "false." However, structure types ($X12$) can be dichotomized into reinforced concrete (RC) wall type and RC column type. Nonetheless, it cannot be clearly defined how these can be distinguished.

To solve this problem, this research introduces an alteration function of structure type index (RC) based on analysis of research data. This function transforms this dichotomy-hard nominal data (structure type, $X12$) to numerical scale data. This representative is derived from measuring the degree of the differences of quantity of vertical formwork per unit plan area, as defined in eq. [1]. In the case of an RC column type, a decrease in vertical formwork quantity is expected when the RC wall type is substituted by other methods such as brick, block, or drywall. Utilizing this index, the structure type (RC) attribute is converted to the numerical scale.

$$[1] \quad \text{Structure type index } (X_{SI}) = \frac{\text{Quantity of vertical formwork } (\text{m}^2)}{\text{Unit floor area } (\text{m}^2)}$$

Indeed, the structure type index can be used for discriminating the structure type RC wall or RC column. Based on the mean of the index of each type, a discriminating function $\mu_{\text{RC column}}(X_{SI})$ can simply be induced that is associated to each structure type index X_{SI} (eq. [2]). To be exact, the mean value of the structure type index X_{SI} of the research data for RC indoor wall type buildings is 4.78. On the contrary, the mean X_{SI} value of the RC column type buildings is 2.38. By connecting these two figures linearly, a discrimination function can be deduced after converting the X_{SI} of RC column and X_{SI} of RC wall into the [0, 1] range. This $\mu_{\text{RC column}}(X_{SI})$ value represents the grade of membership of X_{SI} to the RC column structure type. Thus, the figure can be used for retrieving similar cases from the dataset as one of numerical attributes for the CBR cost model. This value represents the grade of membership of X_{SI} to the RC column structure type when similar cases of the target case are retrieved from the dataset using the CBR model.

$$[2] \quad \mu_{\text{RCcolumn}}(X_{SI}) = \begin{cases} 1 & X_{SI} < 2.83 \\ 2.451 - \dfrac{X_{SI}}{1.95} & 2.83 \le X_{SI} \le 4.78 \\ 0 & 4.78 < X_{SI} \end{cases}$$

Table 2. Configuration of case features.

Features	Feature type	Measurement scale (converted scale)
(*X*1) Number of households	Numeric	Integer
(*X*2) Gross floor area	Numeric	Real number
(*X*3) Number of unit floor households	Numeric	Integer
(*X*4) Number of elevators	Numeric	Integer
(*X*5) Number of floors	Numeric	Integer
(*X*6) Number of piloti with household scale	Numeric	Integer
(*X*7) Number of households of unit floor per elevator	Numeric	Real number
(*X*8) Height between stories	Numeric	Integer
(*X*9) Depth of pit	Numeric	Real number
(*X*10) Roof type	One of a list	Flat or inclined (1 or 0)
(*X*11) Hallway type	One of a list	Hall or corridor (1 or 0)
(*X*12) Structure type (RC)	One of a list	RC wall or RC column (structure type index)
(*X*13) Cost	Numeric	Real number

Note: RC, reinforced concrete.

Optimization using GAs for assigning weight value

Genetic algorithms are used to search a space of candidate hypotheses to identify the best hypothesis. The best hypothesis is defined as the optimized value of the predefined numerical measure at hand, which is called *hypothesis fitness* (Mitchell 1997). 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.

$$[3] \qquad C_j = \omega_1 X_{ji} + \omega_2 X_{j2} + \omega_3 X_{j3} + \cdots + \omega_i X_{ji}$$

Let C_j, ω_i, and X_{ji} 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 (eq. [4]).

$$[4] \qquad \begin{pmatrix} X_{11} & \cdots & X_{1i} \\ \vdots & \ddots & \vdots \\ X_{j1} & \cdots & X_{ji} \end{pmatrix} \begin{pmatrix} \omega_1 \\ \vdots \\ \omega_i \end{pmatrix} = \begin{pmatrix} C_1 \\ \vdots \\ C_j \end{pmatrix}$$

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 ω_i represent the distance, and then the hypothesis fitness function is defined as follows.

$$[5] \qquad \min \sqrt{\sum_{m=1}^{j} D_n^2}$$

$$\text{when} \begin{pmatrix} D_1 \\ \vdots \\ D_j \end{pmatrix} = \begin{pmatrix} C_1 \\ \vdots \\ C_j \end{pmatrix} - \begin{pmatrix} X_{11} & \cdots & X_{1i} \\ \vdots & \ddots & \vdots \\ X_{j1} & \cdots & X_{ji} \end{pmatrix} \begin{pmatrix} \omega_1 \\ \vdots \\ \omega_i \end{pmatrix}$$

Before optimizing this function, normalization must first be conducted. Normalization, which converts raw values to standard scores, requires selecting values that span one range and representing them in another range. Although Koo et al. (2010*a*) tried to standardize the attributes' weight values into the [0, 1] range by dividing the weight values by its own maximum value, this attempt was limited by the possibility of distortion provoked by feature shift. To overcome this problem, the previously identified 13 types of attribute data, which all have a different data range, are converted to a scale of 0 to 1 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. The 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). Consequently, this feature range assignment concept can resolve the incomputable problem when the target exists outside of the case-based range; and mitigate or prevent a sudden feature shift that could distort the accuracy of the similarity measure.

$$[6] \qquad f(X|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(X-\mu)^2}{2\sigma^2}\right]$$

Using the "MS-EXCEL STANDARDIZE" and "NORMSDIST" functions, the data are converted to standardized values. Then, the cumulative probability density of each value of attributes is computed. All the values are represented by a new score of 0 to 1. 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 0 to 1 for the adjusting cell (attribute weight) range, 0.1 for the crossover rate, and 0.05 for the mutation rate.

Similarity measure

The computation of similarity is an important issue for the case retrieving process. Establishing an appropriate similarity function is an attempt to handle the hidden relationships among 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 an equation. An arbitrary case x can be described by the feature vector as follows:

$$[7] \qquad [a_1(x), a_2(x), a_3(x), \cdots a_n(x)]$$

where $a_n(x)$ denotes the value of the nth attributes of case x and w_n denotes the weight of the attributes of nth case. Then, the weighted distance between the two case x_a and x_b is defined as DIS (x_a, x_b), as seen in eq. [8] (Pal and Shiu 2004, eq. [8]).

$$[8] \qquad \mathrm{DIS}(x_a, x_b) = \sqrt{\sum_{n=1}^{i} \varpi_n{}^2 [a_n(x_a) - a_n(x_b)]^2}$$

A good similarity measure should take the concept of invariance into consideration (Perner 2002). As an example of similarity measures for images, despite they are rotated, translated and different in scale, images can be considered similar. In particular, scale invariance can be obtained by normalization that reduces the influence of energy (Perner 2002). In this respect, the normalization of attribute values has a significance that can mitigate a measurement distortion that has originated from different data range.

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

$$[9] \qquad x_a = x_b \rightarrow \mathrm{SIM}(x_a, x_b) = 1 \quad \text{and} \quad \mathrm{DIS}(x_a, x_b) = 0$$

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

$$[10] \qquad \mathrm{SIM}(x_a, x_b) = 1 - \mathrm{DIS}(x_a, x_b) = 1 - \sqrt{\sum_{n=1}^{i} \varpi_n{}^2 [a_n(x_a) - a_n(x_b)]^2}$$

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 (SI) is defined as below. Accordingly, this function has high explanatory power supported by the Euclidean geometric aspect.

$$[11] \qquad \mathrm{SI}(\%) = \left[1 - \sqrt{\frac{\sum_{n=1}^{i} \varpi_n{}^2 [a_n(x_a) - a_n(x_b)]^2}{\sum_{n=1}^{i} \varpi_n{}^2}} \right] \times 100$$

Model validation

So far, this research has introduced an attribute weight assignment 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 assignment and similarity scoring.

First, 20 cases were randomly selected from the case base and excluded from the case base of the CBR model. The profile of these 20 cases is shown in Table 3. 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 (the nearest neighbor, One-NN) approach and the ten higher rank similarity score case base (ten nearest neighbors, Ten-NN) approach are utilized for estimating the building cost. Simultaneously, nine different types of CBR models, which are dependent on the combination of weight value assigning methods and similarity functions, are defined. S_0, S_1, and S_2 denote the similarity function proposed in this research, the arithmetic summation based function, and the fractional function based similarity measure function, respectively. W_0, W_1, and W_2 refer to the weight value assignment method proposed in this research, feature counting, and the method utilizing the standardized coefficient of multiple regression analysis (CMRA), respectively (Table 4). The weight value of each attribute was computed using the genetic algorithm optimization process and the SPSS linear regression function (Table 5). The absolute error ratio (AER) is defined as below (C_A and C_E denotes actual cost and estimated cost, respectively) to evaluate the effectiveness of the system and is compared to other counterparts.

$$[12] \qquad \mathrm{AER}(\%) = \begin{cases} C_A - C_E > 1, & [(C_A - C_E) - 1] \times 100 \\ \text{Otherwise}, & [1 - (C_A - C_E)] \times 100 \end{cases}$$

As summarized in Table 6, it is identified that different cases are retrieved according to applying both weight value assignment and similarity scoring methods. On closer examination, the impact of similarity measuring methods on cost

Table 3. Profile of cases for model validation.

Case	*X*1	*X*2	*X*3	*X*4	*X*5	*X*6	*X*7	*X*8	*X*9	*X*10	*X*11	*X*12	Cost
Case 1	7	735	1	0.5	7	—	2	2.9	5.2	—	1	2.38	404 340 643
Case 2	7	990	1	0.5	7	—	2	2.9	5.2	—	1	2.38	547 492 099
Case 3	14	1508	2	1	7	—	2	2.9	5.2	—	1	2.57	1 006 531 964
Case 4	26	2188	2	1	14	2	2	2.9	8.7	1	1	4.7	1 134 424 960
Case 5	24	1994	2	1	13	2	2	2.8	9.36	—	1	5.1	1 223 354 254
Case 6	38	3185	4	2	10	2	2	2.9	8.7	1	1	4.25	2 038 465 772
Case 7	50	3753	8	1	7	6	8	2.9	5.2	—	—	2.22	2 045 693 136
Case 8	63	4531	5	1	13	2	5	2.8	5.85	1	—	4.74	2 591 614 705
Case 9	40	4448	4	1	11	4	4	2.8	5.85	1	—	4.47	2 894 608 265
Case 10	60	4703	4	1	15	—	4	2.9	8.7	1	1	5.84	2 924 845 777
Case 11	24	2635	2	1	12	—	2	2.8	9.36	—	1	5.98	2 951 709 985
Case 12	54	4472	4	2	15	4	2	2.8	5.85	—	1	5.4	2 983 510 552
Case 13	44	4908	4	2	12	4	2	2.8	5.85	—	1	4.29	3 110 324 443
Case 14	48	5320	4	1	13	4	4	2.8	5.85	1	—	5.42	3 334 662 539
Case 15	56	6189	4	1	15	4	4	2.8	8.6	1	—	6.65	3 682 360 556
Case 16	50	7199	4	2	14	2	2	2.9	8.7	—	1	4.11	3 806 177 268
Case 17	50	7115	4	2	14	8	2	2.8	5.85	—	1	4.37	4 196 711 507
Case 18	50	7121	4	2	14	4	2	2.8	5.85	—	—	4.68	4 398 362 914
Case 19	58	8218	4	2	15	—	2	2.8	5.85	—	1	4.75	4 660 579 804
Case 20	52	5811	4	2	14	4	2	2.9	8.7	—	1	4.65	4 707 437 631

Note: *X*1, number of households; *X*2, gross floor area; *X*3, number of unit floor households; *X*4, number of elevators; *X*5, number of floors; *X*6, number of piloti with household scale; *X*7, number of households of unit floor per elevator; *X*8, height between stories; *X*9, depth of pit; *X*10, roof type; *X*11, hallway type; *X*12, structure type.

Table 4. Model combinations for the validity test.

	Weight value assigning methods		
Similarity measuring methods	Genetic algorithm (W_0)	Feature counting (W_1)	CMRA (W_2)
Euclidean distance base (S_0) $\left[1-\sqrt{\frac{\sum_{n=1}^{i}\varpi_n{}^2\left(a_n(x_a)-\mathrm{a}_n(x_b)\right)^2}{\sum_{n=1}^{i}\varpi_n{}^2}}\right]\times 100$	S_0W_0	S_0W_1	S_0W_2
Arithmetic summation (S_1) $\left[\sum_{n=1}^{i}\varpi_n\left\{1-\sqrt{\frac{\left(a_n(x_a)-\mathrm{a}_n(x_b)\right)^2}{(a_{n,\max}-\mathrm{a}_{n,\min})^2}}\right\}\right]\times 100$	S_1W_0	S_1W_1	S_1W_2
Fractional function base (S_2) $\left[\sum_{n=1}^{i}\varpi_n\left\{1/\left(1+\sqrt{\frac{\left(a_n(x_a)-\mathrm{a}_n(x_b)\right)^2}{(a_{n,\max}-\mathrm{a}_{n,\min})^2}}\right)\right\}\right]\times 100$	S_2W_0	S_2W_1	S_2W_2

estimate error rate is more influential than weight value assignment methods. The average AER according to weighting value methods is 11.60% / 8.17% (One-NN / Ten-NN), whereas the similarity is 0.27% / 1.47% (One-NN / Ten-NN).

As summarized in Table 7, it is identified that different cases are retrieved according to applying both weight value assignment and similarity scoring methods. More precisely, when testing the model's effectiveness in respect to the weight assignment method, it was determined that the mean percentage of error of the proposed genetic algorithm optimization (W_0) model is lower than all its counterparts in terms of both the One-NN and Ten-NN approaches. Moreover, this method yielded better results regardless of the combination of similarity scoring methods. Models that utilize the proposed attribute weighting method (W_0) with the One-NN and Ten-NN approaches have AERs of 9.01% to 10.66% and 10.78% to 11.59%, respectively. On the other hand, when the feature counting method (W_1) based models are used with the One-NN and Ten-NN approaches, the AERs are 20.77% to 22.63% and 18.24% to 20.85%, respectively, while the stand-

Table 5. Weight values applying genetic algorithm, feature counting, and CMRA.

Attributes	Genetic algorithm (W_0)	Feature counting (W_1)	CMRA (W_2)
($X1$) 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

Note: CMRA, coefficient of multiple regression analysis.

Table 6. Analysis of averages of absolute error ratios (public apartment projects).

	One-NN					Ten-NN				
Weighting method	W_0	W_1	W_2	Mean (S_i is Fixed)	Difference	W_0	W_1	W_2	Mean (S_i is Fixed)	Difference
S_0	9.00	22.60	11.00	14.20		10.80	18.20	11.70	13.57	
S_1	10.00	21.10	10.70	13.93	0.27	11.60	20.90	12.60	15.03	1.47
S_2	10.70	20.80	10.70	14.07		11.20	19.00	11.20	13.80	
Mean (W_i is Fixed)	9.90	21.50	10.80			11.20	19.37	11.83		
Difference	11.60								8.17	

Note: W_0, suggested weighting method; W_1, feature counting; W_2, coefficient of multiple regression analysis; S_0, suggested similarity measuring method; S_1, arithmetic summation; S_2, fractional functions.

ardized coefficient of multiple regression analysis (CMRA) method (W_2) base models have 10.75% to 11.01% and 11.24% to 12.58% AERs, respectively.

In terms of the effectiveness of the similarity measuring method, the proposed similarity measuring method (S_0) based model combined with GA optimized (W_0), feature counting (W_1), and CMRA (W_2), 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. The arithmetic similarity method (S_1) 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 (S_0) 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 algorithm is the most accurate using both One-NN and Ten-NN (Table 5). Additionally, after conducting a one-way ANOVA procedure, a significant difference between these nine models was not detected.

Regarding the applicability issue, we conducted an additional validation of the suggested CBR model for another type of project. For the test, a CBR cost model was developed using data from 129 military quarter projects. Thirteen cases were selected from the case base and excluded from the case base of the CBR model. To prevent selection bias, all case data were arrayed in ascending order based on gross floor area. Then, test cases were extracted from the fifth case and every tenth case thereafter (i.e., 5th, 15th, 25th, …, and 125th). At the same time, we extracted 10 numerical and Boolean scale representatives by analyzing the drawings. Thereafter, the AER values according to the combinations of three similarity and three weighting methods were compared. Based on this test, we evaluated the model's capability and applicability to another type of project. The test results of the military quarter projects were similar to those of the public apartment projects (Table 8).

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 CBR cost model can be an effective means of estimating construction cost, as it is based on identifying the characteristics of cases. Case-based reasoning 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 measures based on 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. Despite the fact that various weight assign-

Table 7. Comparison of absolute error ratios (public apartment projects).

	S_0W_0		S_0W_1		S_0W_2		S_1W_0		S_1W_1		S_1W_2		S_2W_0		S_2W_1		S_2W_2	
Case	One-NN	Ten-NN	One-NN	Ten-NN	One-NN	Ten-NN	One-NN	Ten-NN	One-NN	Ten-NN	One-NN	Ten-NN	One-NN	Ten-NN	One-NN	Ten-NN	One-NN	Ten-NN
1	44.2	49.9	46.3	51.0	44.2	48.4	47.2	60.4	49.7	63.1	38.9	49.3	38.7	50.4	42.4	50.9	38.9	48.3
2	8.3	4.7	8.3	9.8	10.6	4.7	8.3	4.7	8.0	9.8	10.6	4.7	8.3	4.7	8.3	9.8	10.6	4.7
3	9.8	24.1	9.8	40.9	9.8	15.4	9.8	29.3	9.8	41.0	9.8	9.0	9.8	29.3	9.8	40.0	9.8	9.0
4	1.2	23.3	24.0	18.6	2.0	3.9	1.2	17.5	24.0	30.3	2.0	10.3	1.2	17.5	24.0	18.6	2.0	7.3
5	2.4	10.8	58.6	36.8	0.5	9.2	2.4	12.8	18.2	34.7	0.5	23.0	2.4	11.1	18.2	34.9	0.5	9.2
6	4.7	0.2	29.4	10.4	20.1	4.2	4.7	1.3	36.5	2.1	20.1	15.5	4.7	1.3	36.5	3.7	20.1	10.5
7	8.4	4.5	153.5	25.4	9.6	18.4	31.6	4.3	153.5	19.8	9.6	16.3	53.0	5.5	153.5	35.0	9.6	8.7
8	2.2	5.6	2.2	14.4	2.2	6.1	2.2	4.9	2.2	14.4	2.2	8.1	2.2	4.9	2.2	11.6	2.2	7.2
9	4.3	6.7	4.3	7.4	4.3	7.1	4.3	2.7	4.3	9.3	4.3	2.1	4.3	6.7	4.3	10.7	4.3	2.1
10	19.2	3.6	19.2	5.4	19.2	0.0	19.2	1.2	19.2	34.3	19.2	0.4	19.2	1.2	19.2	30.2	19.2	0.4
11	9.4	17.0	30.5	20.9	9.4	29.9	9.4	17.0	30.5	11.7	9.4	19.8	9.4	17.0	30.5	13.8	9.4	24.5
12	5.5	3.4	12.7	9.6	0.1	4.7	5.5	3.4	12.7	7.7	0.1	1.7	5.5	3.4	12.7	11.6	0.1	1.7
13	1.6	0.3	1.6	0.4	1.6	2.7	1.6	3.7	1.6	7.0	1.6	3.2	1.6	1.2	1.6	7.3	1.6	1.3
14	5.1	2.0	5.1	6.8	5.1	2.2	5.1	0.8	5.1	5.0	5.1	2.7	5.1	0.8	5.1	5.0	5.1	2.7
15	0.8	1.8	0.8	2.2	0.8	0.4	0.8	0.1	0.8	0.8	0.8	0.1	0.8	1.8	0.8	0.8	0.8	0.1
16	9.3	8.8	19.1	4.8	14.4	12.1	9.3	8.8	19.1	24.7	14.4	10.8	9.3	8.8	19.1	11.9	14.4	10.8
17	0.8	1.8	0.8	2.2	0.8	0.4	0.8	0.1	0.8	0.8	0.8	0.1	0.8	1.8	0.8	0.8	0.8	0.1
18	1.1	6.1	1.1	25.7	1.1	5.6	1.1	10.8	1.1	19.8	1.1	8.8	1.1	10.8	1.1	18.1	1.1	10.9
19	7.7	14.1	1.7	39.2	1.7	16.8	1.7	14.0	1.7	33.0	1.7	18.3	1.7	13.1	1.7	33.0	1.7	18.3
20	34.3	26.9	23.7	33.0	62.9	41.9	34.3	34.1	23.7	47.6	62.9	47.4	34.3	32.8	23.7	31.5	62.9	47.0
Mean	9.0	10.8	22.6	18.2	11.0	11.7	10.0	11.6	21.1	20.9	10.7	12.6	10.7	11.2	20.8	19.0	10.7	11.2
S.D.	11.4	12.3	34.8	15.1	16.1	13.7	13.0	14.9	34.0	17.3	15.5	14.1	14.6	13.0	33.7	14.5	15.5	13.9

Table 8. Analysis of averages of absolute error ratios (military quarter projects).

	One-NN					Ten-NN				
Weighting method	W_0	W_1	W_2	Mean (S_i is Fixed)	Difference	W_0	W_1	W_2	Mean (S_i is Fixed)	Difference
S_0	7.75	32.54	12.79	17.69		9.23	31.49	9.92	16.88	
S_1	8.12	58.24	13.29	26.55	8.86	10.77	40.92	14.10	21.93	5.05
S_2	6.75	55.75	12.63	25.04		10.35	41.05	11.73	21.04	
Mean (W_i is fixed)	7.54	48.84	12.90			10.12	37.82	11.92		
Difference	41.3					27.70				

Note: W_0, suggested weighting method; W_1, feature counting; W_2, coefficient of multiple regression analysis; S_0, suggested similarity measuring method; S_1, arithmetic summation; S_2, fractional functions.

ing methods continue to be proposed, there is no consensus on which method is the best.

As an effort to address these challenges, this research developed methods for measuring similarity and assigning weight value for CBR modeling and suggested a CBR cost estimation model for the budgeting of apartment buildings in Korea. After evaluating the model in terms of the similarity scoring and attributes weight assignment 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 counterparts. Moreover, the proposed method can enhance the value of a case-based reasoning method by improving the explanatory power of similarity measurement and by mitigating the output distortion provoked by sudden changes of features. 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 customized and applied to different types of construction and contribute toward the enhancement of cost estimation research.

The results of this research initiated efforts toward developing CBR by suggesting new methods in terms of similarity measure and attributes weighting. It must be noted that this research is based on data from a limited number of cases, and that additional research and testing must be conducted to further validate the model and to generalize the effects of the suggested methods.

Acknowledgment

This research was supported by a grant (R&D06CIT-A03) from the Innovative Construction Cost Engineering Research Center; and a grant from the super-tall Building R&D project (VC-10) funded by the Korean Ministry of Land, Transport, and Marine Affairs.

References

Aamodt, A., and Plaza, E. 1994. Case-based reasoning: Foundational issues, methodological variations and system approaches. AI Communications, **7**(1): 35–39.

Ahn, H., Kim, K.J., and Han, I. 2006. Global optimization of feature weights and the number of neighbors that combine in a case-based reasoning system. Expert Systems: International Journal of Knowledge Engineering and Neural Networks, **23**(5): 290–301.

An, S.-H., Kim, G., and Kang, K. 2007. A case-based reasoning cost estimating model using experience by analytic hierarchy process. Building and Environment, **42**(7): 2573–2579. doi:10.1016/j.buildenv.2006.06.007.

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**(6): 385–393. doi:10.1111/0885-9507.00157.

Burkhard, H.-D. 2001. Similarity and distance in case-based reasoning. Fundamenta Informaticae, **47**: 201–215.

Chou, J.-S. 2009. Web-based CBR System applied to early cost budgeting for pavement maintenance project. Expert Systems with Applications, **36**(2): 2947–2960. doi:10.1016/j.eswa.2008.01.025.

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 (ASCE), **20**(5): 339–350. doi:10.1061/(ASCE)0887-3801(2006)20:5(339).

Chua, D.K.H., Li, D.Z., and Chan, T. 2001. Case-based reasoning approach in bid decision-making. Journal of Construction Engineering and Management (ASCE), **127**(1): 35–45. doi:10.1061/(ASCE)0733-9364(2001)127:1(35).

Dikmen, I., Birgonul, M., and Gur, A. 2007. A case-based decision support tool for bid mark-up estimation of international construction projects. Automation in Construction, **17**(1): 30–44. doi:10.1016/j.autcon.2007.02.009.

Doğan, S.Z., Arditi, D., and Günaydın, H.M. 2006. Determining attribute weights in a CBR model for early cost prediction of structural system. Journal of Construction Engineering and Management, **132**(10): 1092–1098. doi:10.1061/(ASCE)0733-9364(2006)132:10(1092).

Doğan, S.Z., Arditi, D., and Murat Günaydin, H. 2008. Using decision trees for determining attribute weights in a case-based model for early cost estimation. Journal of Construction Engineering and Management, **134**(2): 146–152. doi:10.1061/(ASCE)0733-9364(2008)134:2(146).

Dussart, C., Pommier, P., Siranyan, V., Grelaud, G., and Dussart, S. 2008. Optimizing clinical practice with case-based reasoning approach. Journal of Evaluation in Clinical Practice, **14**(5): 718–720. doi:10.1111/j.1365-2753.2008.01071.x. PMID:19018901.

Ellsworth, R.K. 1998. Cost-to-capacity analysis for estimating waste-to-energy facility costs. Cost Engineering, **40**(6): 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): 238–253. doi:10.1111/j.1467-8640.2006.00286.x.

Goh, Y.M., and Chua, D.K.H. 2010. Case-Based Reasoning Approach to Construction Safety Hazard Identification: Adaptation and Utilization. Journal of Construction Engineering and Management, **136**(2): 170–178. doi:10.1061/(ASCE)CO.1943-7862.0000116.

Goldberg, D.E. 2006. Genetic algorithms in search, optimization and machine learning. Addison-Wesley.

Han, S.-G., Lee, S., and Jo, G. 2005. Case-based tutoring system for procedural problem solving on the WWW. Expert Systems with Applications, **29**(3): 573–582. doi:10.1016/j.eswa.2005.04.026.

Hendrickson, C. 2000. Project management for construction 2. 2nd ed. World Wide Web Publication, http://pmbook.ce.cmu.edu/05_Cost_Estimation.html.

Huang, M.-J., Chen, M., and Lee, S. 2007. Integrating data mining with case-based reasoning for chronic disease prognosis and diagnosis. Expert Systems with Applications, **32**(3): 856–867. doi:10.1016/j.eswa.2006.01.038.

Jrade, A., and Alkass, S. 2007. Computer-integrated system for estimating the costs of building project. Journal of Construction Engineering and Management, **13**(4): 205–223. doi:10.1061/(ASCE)1076-0431(2007)13:4(205).

Karshenas, S., and Tse, J. 2002. A case-based reasoning approach to construction cost estimating. Computing in Civil Engineering: 113–123. doi:10.1061/40652(2003)10.

Kolodner, J. 1993. Case-based reasoning. Kaufmann.

Koo, C.-W., Hong, T.H., Hyun, C.T., and Koo, K.J. 2010*a*. A CBR-based hybrid model for predicting a construction duration and cost based on project characteristics in multi-family housing projects. Canadian Journal of Civil Engineering, **37**(5): 739–752. doi:10.1139/L10-007.

Koo, C.-W., Hong, T.H., Hyun, C.-T., Park, S.H., and Seo, J.-O. 2010*b*. A study on the development of a cost model based on the owner's decision making at the early stages of a construction projects. International Journal of Strategic Property Management, **14**(2): 121–137. doi:10.3846/ijspm.2010.10.

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., Petry, F.E., Buckles, B.P., and Sadasivan, T. 1997. Genetic algorithms for query optimization in information retrieval: Relevance feedback. Advanced in Fuzzy Systems Applications and Theory, **7**: 155–174.

Liao, T.W., Zhang, Z., and Mount, C. 1998. Similarity Measurement for Retrieval in Case-Based Reasoning System. Applied Artificial Intelligence, **12**(4): 267–288. doi:10.1080/088395198117730.

Lopes, H.S., Coutinho, M.S., and de Lima, W.C. 1997. An evolutionary approach to simulate cognitive feedback learning in medical domain. Advanced in Fuzzy Systems Applications and Theory, **7**: 193–208.

Mitchell, T.M. 1997, Machine learning. McGraw-Hill.

Morcous, G., Rivard, H., and Hanna, A.M. 2002. Case-based reasoning system for modeling infrastructure deterioration. Journal of Computing in Civil Engineering, **16**(2): 104–114. doi:10.1061/(ASCE)0887-3801(2002)16:2(104).

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): 173–182. doi:10.1061/(ASCE)0733-9364(2001)127:3(173).

Ozorhon, B., Dikmen, I., and Birgonul, M.T. 2006. Case-based reasoning model for international market selection. Journal of Construction Engineering and Management, **132**(9): 940–948. doi:10.1061/(ASCE)0733-9364(2006)132:9(940).

Pal, S.K., and Shiu, S.C.K. 2004. Foundations of soft case-based reasoning. Wiley Interscience.

Park, Y.-J., Kim, B.-C., and Chun, S.-H. 2006. New knowledge extraction technique using probability for case-based reasoning: Application to medical diagnosis. Expert Systems: International Journal of Knowledge Engineering and Neural Networks, **23**(1): 2–20. doi:10.1111/j.1468-0394.2006.00321.x.

Perner, P. 2002. Are case-based reasoning and dissimilar-based classification two sides of the same coin? Engineering Applications of Artificial Intelligence, **15**(2): 193–203. doi:10.1016/S0952-1976(02)00020-9.

Qian, Z., Gao, W.S., Wang, F., and Yan, Z.2008 . A case-based approach to power transformer fault diagnosis using dissolved gas analysis data. European Transactions on Electrical Power, **19**(3): 518–530. doi:10.1002/etep.240.

Ryu, H.-K., Lee, H.-S., and Park, M. 2007. Construction planning method using case-based reasoning (CONPLA-CBR). Journal of Computing in Civil Engineering, **21**(6): 410–422. doi:10.1061/(ASCE)0887-3801(2007)21:6(410).

Seeley, I.H. 1997. Quantity surveying practice. 2nd ed. Macmillan Press.

Smith, A.J. 1995. Estimating, tendering and bidding for construction. Macmillan, London.

Stiff, J.B., and Mongeau, P.A. 2002. Persuasive communication. 2nd ed. Guilford Press.

Sun, B., Xu, L.D., Pei, X., and Li, H. 2003. Scenario-based knowledge representation in case-based reasoning systems. Expert Systems: International Journal of Knowledge Engineering and Neural Networks, **20**(2): 92–99. doi:10.1111/1468-0394.00230.

Tah, J.H.M., Carr, V., and Howes, R. 1998. An application of case-based reasoning to the planning of highway bridge construction. Engineering, Construction, and Architectural Management, **5**(4): 327–338. doi:10.1108/eb021086.

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, **13**(2): 143–150. doi:10.1111/0885-9507.00094.

Yi, J. 2006. A study on case-based forecasting model for monthly expenditures of residential building project. Korean Journal of Construction Engineering and Management, **79**(1): 128–137.

Zhuang, Z.Y., Churilov, L., Burstein, F., and Sikaris, K. 2007. Combining data mining and case-based reasoning for intelligent decision support for pathology ordering by general practitioners. European Journal of Operational Research, **195**(3): 662–675. doi:10.1016/j.ejor.2007.11.003.

List of symbols

$a_n(x)$	the value of the attributes of case *x,*
CMRA	Coefficient of multiple regression analysis
C_j	the cost of the *j*th case project
D_j	distance between C_j and $\omega_1 X_{ji}+\omega_2 X_{j2}+\omega_3 X_{j3}+$ ……$+\omega_i X_{ji}$
DIS (x_i, x_j)	the weighted distance between the two cases x_i and x_j
f $(X \mid \mu, \sigma^2)$	probability density function of the standard distribution
One-NN	One nearest neighbor
SIM (x_a, x_b)	stands for the degree of similarity between x_a and x_b
S_0	Euclidean distance based similarity function
S_1	arithmetic summation based function
S_2	fractional function based similarity measure function
Ten-NN	ten nearest neighbors
W_0	the weight value assigning method optimized by genetic algorithm
W_1	the weight value assigning method by feature counting
W_2	the weight value assigning method utilizing the standardized coefficient of multiple regression analysis
ω_n	the weight of the *n*th attributes of the case
X_{ji}	*i*th attribute value of *j*th case
X_{SI}	structure type index
$\mu_{RC\ column}(X_{SI})$	a discriminating function of structure type RC wall or RC column

This article has been cited by:

1. S. Kim, H. Lee, M. Park, W. KimAn Expert System for Construction Decision-Making Using Case-Based Reasoning 89-96. [CrossRef]
2. Sae-Hyun Ji, Moonseo Park, Hyun-Soo Lee. 2012. Case Adaptation Method of Case-Based Reasoning for Construction Cost Estimation in Korea. *Journal of Construction Engineering and Management* **138**:1, 43. [CrossRef]
3. Sae-Hyun Ji, Moonseo Park, Hyun-Soo Lee. 2012. Case Adaptation Method of Case-Based Reasoning for Construction Cost Estimation in Korea. *Journal of Construction Engineering and Management* **138**:1, 43. [CrossRef]