REFINING THE CASE-BASED REASONING MODEL OF CONSTRUCTION PROCESSES

Darren Graham[1], Simon D. Smith[2] and Martin Crapper[3]

[1]*KPMG (UK) LLP, Saltire Court, Edinburgh, EH1 2EG, UK*
[2, 3] *Institute for Infrastructure and Environment, School of Engineering & Electronics, University of Edinburgh, The King's Buildings, Edinburgh, EH9 3JL, UK*

Case-based reasoning has been shown in previous research to be capable of modelling construction processes with a good degree of accuracy. However, for the technology to become accepted amongst construction practitioners, the accuracy, reliability and especially the efficiency of the modelling methodology must be improved. To this end, a new method of performing one of the most critical functions of a case-based reasoning model, case retrieval, is proposed. This method, known as "cluster-based retrieval" involves removing the cases which are likely to be poor solutions to a problem at an early stage, allowing a focus to be placed on finding a solution from a greatly reduced group of cases. A model based upon real construction data and utilising the "cluster-based retrieval" method have been developed, validated and compared with 'the' model developed in previous research. This comparison aimed to measure any differences in the accuracy, reliability and efficiency as a result of the introduction of "cluster-based retrieval". The model utilising the "cluster-based retrieval" method produced results more efficiently and with more accuracy and reliability than the original model. These results indicate that the use of a "cluster-based retrieval" method in a case-based reasoning model is a 'step in the right direction' to industry acceptance of the technology as a method of modelling construction processes.

Keywords: case-based reasoning, construction planning, estimation.

INTRODUCTION

Many of the processes that are performed in construction projects are stochastic in nature - they are liable to behave randomly over time. This nature makes such processes difficult to control, especially when project planning techniques, such as the critical path method (CPM), are not sufficiently detailed to incorporate the stochastic behaviour of these processes into a project plan. This can result in activities being completed with duration greatly different to that estimated in the project plan, resulting in: a delay in the project schedule, if the activity's duration overruns; a gap in the schedule and resources standing idle, if the activity's duration is less than planned. Either of these situations would result in an increase in the project cost.

[1] darren.graham@kpmg.co.uk
[2] simon.smith@ed.ac.uk
[3] martin.crapper@ed.ac.uk

Graham, D, Smith, S D and Crapper, M (2006) Refining the case-based reasoning model of construction processes. *In:* Boyd, D (Ed) *Procs 22nd Annual ARCOM Conference,* 4-6 September 2006, Birmingham, UK, Association of Researchers in Construction Management, 135-142.

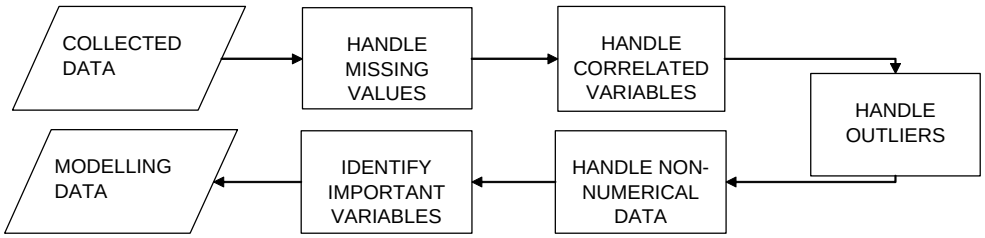

Figure 1 Pre-modelling data analysis framework

A number of models, using a variety of modelling paradigms, have been developed by researchers with the aim of helping construction practitioners to close the gap between the project plan and project reality. Case-based reasoning (CBR) is one of the paradigms which demonstrated a capability for modelling stochastic construction processes to a good degree of accuracy. Graham and Smith (2004) established that CBR is capable of producing estimates of the duration of such processes to within 10% accuracy, with 90% reliability. One drawback of the CBR model developed by Graham and Smith is that it takes 10 minutes for an estimate of process duration to be produced. Thus, the original CBR model, while accurate, is not very practicable.

To make this CBR model practicable and to help the construction industry accept this technological advancement and begin utilising CBR to plan stochastic construction processes there is a need to concentrate on improving the accuracy, reliability and in particular the efficiency of this CBR model.

In a bid to achieve accuracy, reliability and efficiency improvements in the CBR model the research presented in this paper focuses on the refinement of one of the key parts of a CBR model, the case retrieval mechanism. As an example the refined CBR model is developed using data collected from observations of the concrete placement process. This refined model is then compared with the original CBR model to determine if the aims of this research have been achieved.

DATA COLLECTION AND ANALYSIS

The data available for use in this research was collected from four construction projects of varying size that were recently undertaken in United Kingdom. The data set contained: 232 observations of concrete pours; 2050 observations of the cycle a concrete truckmixer follows in performing the concrete placement process; 26 explanatory variables; and 8 dependent variables. This is a substantial set of data, which should help to produce robust models. To put the size of the data set in context a comparison with other data sets used in construction research is as follows: Rowings and Sonmez (1996) used a sample size of 112, whereas, AbouRizk *et al.* (2001) use a sample of 39. Arditi and Tokdemir (1999) used a sample size of only 12.

Prior to developing a model, it is important to ensure that the collected data is of sufficient quality for use in modelling. This was done by following the pre-modelling data analysis framework shown in Figure 1.

The following is a summary of how each step in the framework was performed in this research:

- **Handling missing values:** when performing standard statistical analysis techniques it is important to have a rectangular data set, with no missing values. The data set had a small number of missing values and it was necessary to deal

Table 1 Variables to be used in modelling

Variables to be used in modelling
Date (month) of pour
Travel distance
Type of pour
Number of truck-mixers
Total pour volume
Daily maximum temperature

with them. A number of methods were examined: the variables which contained missing data values were removed; the cases which contained missing data values were removed; the missing data values were estimated using the 'expectation-maximisation' and a 'linear interpolation' method (see Little and Rubin, 1987 regarding these methods). A new dataset was created for each of the methods and a regression model was developed of each dataset to determine which method produced a dataset which is closer to linearity – thus making prediction easier. It was found that the interpolation method produced the best answer.

- **Handling correlated variables:** Explanatory variables can be described as correlated if the have a linear relationship. Thus a model will lose no knowledge of the process it is modelling if one of these correlated variables is not considered. In this research, the Pearson correlation coefficient was calculated between each of the explanatory variables and two pairs were found to be correlated. Thus two variables were removed from the dataset.
- **Handling outliers:** the identification of outliers in datasets is commonly based upon a comparison of the Mahalanobis distance between a data point and all other data points (Santos-Pereira and Pires, 2002); with the value of the Chi-square distribution for a given dataset. No outliers were found in the dataset.
- **Handling non-numerical data:** non-numerical data is difficult to deal with when using standard statistical techniques. Therefore, there is a need to transform non-numerical data into a numerical form. This is done by categorising this type of data into positive numerical categories e.g. month = January before categorisation; month = 1 after categorisation.
- **Identifying important variables:** the variables which could not be practicably known at the stage of planning a construction project were removed from the dataset, leaving a smaller dataset of 12 variables. A regression analysis model was developed of this smaller dataset to determine the statistically important variables (shown in table 1). There were 6 explanatory variables which were found to be statistically important and practicable which produced the modelling data.

Prior to the commencement of model development, the data set was separated into two sections: one for development; one for testing and validating the models. The ratio of this split was 80% to 20%, which is fairly typical in a CBR model, resulting in a development data set of 186 cases and a validation data set of 46 cases.

THE CASE-BASED REASONING MODEL

CBR is a cyclical process, which has four steps (Kolodner, 1993). They are:

- **Case representation and storage.** A case usually consists of a description of a problem, in the form of explanatory variables; and a description of the solution to that problem, in the form of dependent variables. In the existing CBR model a case

is described by the 6 explanatory variables described in table 1, with each variable being categorised or 'indexed' to help improve the accuracy of retrieval. For full details of this categorisation the reader should refer to Graham and Smith (2004). To make the model more realistic weightings are added to the explanatory variables. These weightings were calculated using the ID3 algorithm, developed by Quinlan (1986), which builds a decision tree based on the cases in the case base and then uses the generated tree to calculate weights for each variable contained in the tree (Graham and Smith, 2004).

- **Case retrieval.** When a new problem, known as a target case, needs to be solved a search is made of the collection of cases (case base) for the case that best matches the new problem. In the existing model there are two parts to the retrieval of cases: when a target case is presented an initial check is made that that particular problem had not been solved by the model in the past. If it had then the same case would be presented for use as a solution to the problem; if the model had not solved the problem before then a search for a solution is made over the entire case base. A calculation is made of the mathematical distance between each case in the case base and the target case. This is performed in two stages: One, for each case in the case base, the variables of the target case are compared with the variables of the stored case (in the case base) using Equation 1; then, an assessment of the overall similarity between the target case and the stored case is made using Equation 2.
- **Case adaptation.** The retrieved case may not always be readily applied as the solution to the problem; it may need to be adapted in some way. At this stage a case can be adapted – often using some form of mathematical rule to perform the adaptation. In the existing model there is no adaptation applied.
- **Case accumulation.** This step in the process involves the automatic formulation of new cases to be stored in the case base. A new case is formed by using the inputs from the target case (new problem) and the outputs from the retrieved (and possibly adapted) case to form a new case. This mechanism allows the model to learn over time.

$$sim_i(i_{Target}, i_{Stored}) = \begin{cases} 1 & if\ i_{Target} = i_{Stored} \pm 10\% \\ 0 & otherwise \end{cases} \qquad \text{Equation 1}$$

Where: i_{Target} is a variable from the target case; i_{Stored} is a variable from a stored (in the case base) case ; $sim_i(i_{Target}, i_{Stored})$ is a measurement of the similarity between two variables. If the variable of a target case is within a range of +/- 10% of that variable of the stored case, then they are assumed to be similar.

$$sim(Target, Stored) = \frac{1}{p}\sum_{i=1}^{p} w_i \times sim_i(i_{Target}, i_{Stored}) \qquad \text{Equation 2}$$

where weighted sim(Target, Stored) is a measure of the similarity between a stored and target case taking into account the relative importance of the variables; p is the number of variables; w_i is the weighting of a variable, i.

THE REFINED CASE-BASED REASONING MODEL

A CBR model derives its power form its ability to retrieve relevant cases quickly and accurately from its case base (Arditi and Tokdemir, 1999). To recap, this retrieval process is usually performed, as is the case in the above model, by measuring the

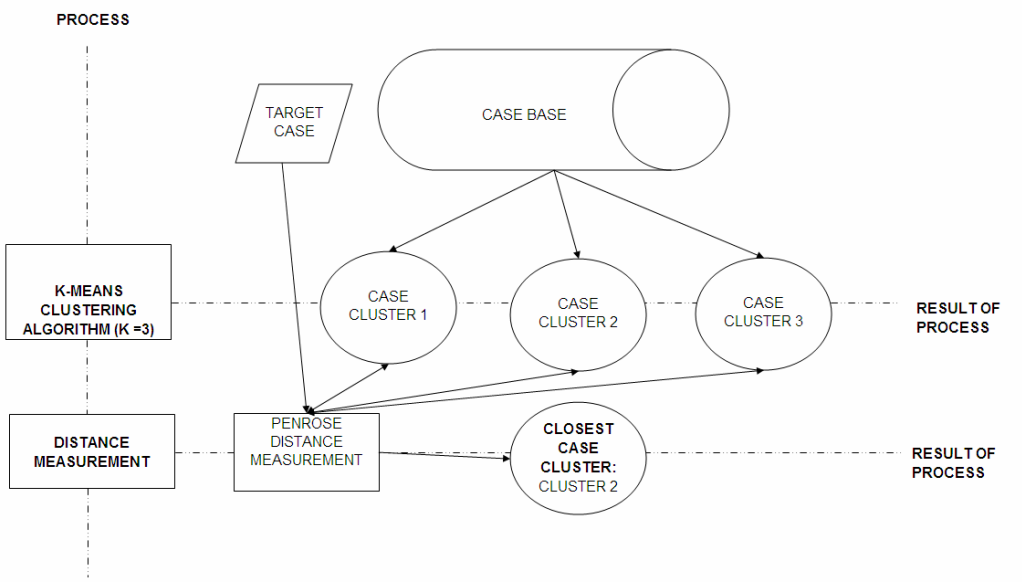

Figure 2: Steps of the 'cluster-based retrieval mechanism to reduce the number of potential solutions

similarity between the target case (problem to be solved) and each of the cases in the case base.

This research has focussed on the refinement of the mechanism for retrieving potential solution(s) in a CBR model in a bid to improve the efficiency, accuracy and reliability of a model. The mechanism has been refined by adding in an additional step into the process of retrieval, which works by: clustering the cases in the case base together using a standard mathematical clustering technique and determining which cluster is most similar to the target case; the cases in the clusters, other than that which is closest to the target case, are discounted from further examination; the cases in the closest cluster to the target case are then measured for similarity to the target case using the techniques outlined for the original CBR model. This is known as the 'cluster-based retrieval' mechanism.

The following is a description of how the 'cluster-based retrieval' mechanism works in the refined CBR model (see figure 2):

- Cluster the cases held in the case-base into 3 clusters using a standard k-cluster algorithm (k = 3 in this instance). For details of this algorithm the reader should refer to Hartigan (1975).
- The Penrose distance measurement, shown in equation 3, is used to measure the distance between the target case and the centroid of each cluster of cases.
- The cases, in the cluster of cases which has the smallest Penrose distance measurement to the target case, are considered further by the model. The other 2 clusters are discounted because the closest cluster is most likely to hold the case which is most similar to the target case. In the example shown in figure 2, cluster 2 was found to be closest to the target case.

$$P_{T\arg et,Clusterk} = \sum_{\text{var}\,iable=1}^{\text{var}\,iable=N} \frac{(\text{var}\,iable_{T\arg et} - \mu_{\text{var}\,iable,Clusterk})^2}{P \times V_{\text{var}\,iable}}$$ Equation 3 (after Penrose, 1953)

Where: $P_{Target,\ Clusterk}$ is the Penrose distance between the target case and the centroid of cluster k; $variable_{Target}$ is the value of Nth variable in the target case; $\mu_{variable,\ Cluster\ k}$ is the centroid value for the Nth variable maximum temperature of all of the cases contained in cluster 1; P is the number of variables; $V_{variable}$ is the covariance matrix of the Nth variable.

- The cases contained in the cluster (selected using the steps outlined above) are used to form a new case base.
- The steps taken in the original CBR model for measuring the similarity between the target case and the case-base cases are now used to determine the most similar case. To recap these are: for each case in the case base, the variables of the target case are compared with the variables of the stored case (in the case base) using equation 1; then, an assessment of the overall similarity between the target case and the stored case is made using equation 2.
- Finally, the most similar case is retrieved as a solution to the target case.

The other steps involved in a CBR model: case representation and storage; case adaptation; and case accumulation are the same in the refined model as in the original CBR model.

It is important to determine if the refinements made to the original CBR model have effected any improvements in the efficiency, accuracy and reliability of the model. To do this a process of model verification and validation was undertaken, which is discussed in the next section.

MODEL VERIFICATION AND VALIDATION

To recap, the CBR models were developed using the development data set of 186 cases (or 80% of the total cases available), leaving 46 cases (or 20% of the total cases available) for use in validation.

Model verification is concerned with determining whether a conceptual model (the model assumptions) has been correctly translated into a computer "program" i.e. debugging the program (Law and Kelton, 2000).

To verify that the model has been developed correctly, a simple test of the model was performed. This involved taking each case held in the case base as a target case. This was done without removing that particular case from the case base. Thus, the case that is being used as a target case should be selected as the most similar case in all 186 instances. This was indeed observed in all cases; and hence the model has been verified as having been developed correctly.

Model validation is concerned with determining whether a model (as opposed to a computer program) is an accurate representation of the system under examination, for the particular objectives of a study (Law and Kelton, 2000). The particular objectives of this research were to accurately estimate how a construction activity will be performed, under given project conditions. It is therefore important to determine whether or not the refined CBR model can accurately estimate the productivity of a concrete pour, given the values of the explanatory variables held in each of the 46 validation cases.

It is noteworthy that the validation data was not included in the development of the model and is hence new to the model. For model validation, each of the 46 validation cases was presented, as a target case, to the CBR model and the estimate of

productivity suggested by the model (taken from the most similar case held in the case base) for each validation case, was compared directly with the actual productivity observed, on the construction site, for that particular case.

It was found that the refined CBR model provided, on average, estimates of the productivity of a concrete pour, to within 7% accuracy in 95% (44 out of 46) of the validation cases. Importantly, it took the refined CBR model, on average, 120 seconds to find a solution.

COMPARISON OF ORIGINAL AND REFINED MODELS

The original CBR model was found, by Graham and Smith (2004), to be capable of estimating the productivity of concrete pours to within 10% accuracy with 90% reliability. It took, on average, 600 seconds for the model to produce an estimate of productivity.

To ensure that the refined model and the original model could be fairly compared, the original model was redeveloped using the same data sets as those used in the refined model. The same validation approach, as that outlined above, was applied to the original CBR model and it was found that the model could produce estimates of the productivity of concrete pours to within 12% accuracy with 95% reliability. The time to produce an estimate remained, on average, at 600 seconds.

A comparison of the performance of the original and refined CBR models demonstrates that an improvement has been made in the accuracy, reliability and efficiency of the model. The accuracy was improved by 5% by refining the CBR model without compromising the reliability of the model. The refined model was made 80% more efficient, than the original, by utilising the 'cluster-based retrieval' mechanism.

CONCLUSIONS

Case-based reasoning has been shown in previous research to be capable of providing estimates of the productivity of stochastic construction processes to a good degree of accuracy (10% error) with a good amount of reliability (90%). However, the model which produced these estimates took, on average 600 seconds, to produce a solution. This time is a major drawback for the CBR modelling paradigm's application to construction problems; hence this research has sought to find a method of improving the efficiency of the CBR model, without compromising on the quality of the output.

To this end, a new method of performing a critical function of a CBR model, the retrieval of potential solutions, was devised and explored. This method centred on an early removal of potential solutions which are unlikely to produce a good estimate. This permits the CBR model to focus its search, for a solution, among: a vastly reduced set of potential solutions – affecting a large increase (80%) in the efficiency of the model; a set of cases which are potentially good solutions to the problem – affecting an increase in the accuracy of the model – the model's likelihood of producing an erroneous solutions are reduced by 5% as there are less solutions of this type in the set of potential solutions.

In summary, the refinement of the CBR model has led to improvements in the capability of the paradigm in its application to the modelling of stochastic construction processes. The challenge is now for the construction industry to take up this technology to take advantage of its ability to grasp the complexity of the processes

that make up a construction project. This ability to accurately model projects at a process level can be assimilated into a project planning tool to permit a construction practitioner to improve the accuracy of a project plan through having a better grasp of the construction project at a process level. The result should be improvements in construction projects in terms of delivering on time and within budget.

REFERENCES

AbouRizk, S., Knowles, P. and Hermann, U. R. (2001) Estimating Labor Production Rates for Industrial Construction Activities. *Journal of Construction Engineering and Management,* **127**(6), 502-511

Arditi, D. and Tokdemir, O. B. (1999) Comparison for Case-Based Reasoning and Artificial Neural Networks. *Journal of Computing in Civil Engineering*, **13**(3), 162-169

Graham, D. and Smith, S.D. (2004) Estimating the Productivity of Cyclic Construction Operations using Case-Based Reasoning. *Advanced Engineering Informatics,* **18**(1), 17-28

Kolodner, J. (1993) *Case-Based Reasoning.* Morgan Kaufmann, California, USA.

Law, A.M. and Kelton, W.D. (2000) *Simulation Modeling and Analysis*. 3rd Edition. McGraw-Hill, London, UK.

Little, R.J.A. and Rubin, D.B. (1987) *Statistical Analysis with Missing Data*. Wiley, New York, USA

Penrose, L.W. (1953) Distance, Size and Shape. *Annals of Eugenics*, **18**, 337-343

Quinlan, J.R. (1986) Introduction to Decision Trees. *Machine Learning*, **1**(1), 81-106

Rowings, J. E. and Sonmez, R. (1996) Labor Productivity Modeling With Neural Networks, *In: AACE International Transactions, PROD.01*, Morgantown, WV, USA, 1-3

Santos-Pereira, C.M. and Pires, A.N. (2002) Detection of Outliers in Multivariate Data: a Method Based on Clustering and Robust Estimators, In: *Proceedings in Computational Statistics, 15th Symposium*, Berlin, Germany