Case-Based Reasoning for Construction Hazard Identification: Case Representation and Retrieval

Y. M. Goh[1] and D. K. H. Chua, M.ASCE[2]

Abstract: This paper proposes a case-based reasoning (CBR) approach to construction hazard identification that facilitates systematic feedback of past knowledge in the form of incident cases and hazard identification. This paper focuses on two of the key components of the CBR approach: (1) a detailed knowledge representation scheme, developed based on the modified loss causation model, to codify incident cases and past hazard identification and (2) an intelligent retrieval mechanism that can automatically retrieve relevant past cases. The detailed knowledge representation scheme presented herein is designed to model both incident cases and hazard identification so that both types of knowledge repository can be retrieved simultaneously and adapted for use. The scheme also includes a linguistic structure used to facilitate indexing of cases. The retrieval mechanism is based on the concept of similarity scoring. In this paper, a novel scoring technique based on semantic networks is presented. A case study is presented to demonstrate and validate the proposed approach.

DOI: 10.1061/(ASCE)CO.1943-7862.0000093

CE Database subject headings: Construction management; Risk management; Safety; Information technology (IT); Accidents; Safety; Knowledge-based systems.

Introduction

Learning from experience is a fundamental process that helps individuals and organizations to improve and not repeat past mistakes. However, it is evident from the consistently poor safety record of the construction industry (Reese and Edison 2006; MOM 2004) that the industry has not been able to learn effectively from its safety experience. This could be contributed by a lack of structured feedback process that ensures improvement of current risk assessments based on knowledge of past projects.

One of the key challenges in encouraging feedback of past safety knowledge is the effort involved in retrieving and understanding past incident cases and risk assessments so that useful information can be incorporated into the new risk assessment. To promote the feedback process described above, artificial intelligence (AI) tools such as case-based reasoning (CBR) tools can be used to improve the efficiency and quality of risk assessment. However, application of AI tools in hazard identification and risk assessment is still lacking in the construction industry.

Zhao et al. (2009) presented a learning hazard and operability (HAZOP) expert system based on CBR. It is one of the few recent efforts to apply CBR tools in risk assessment. Remm and Remm (2008) devised a CBR approach to estimate the risk of enterobiasis among nursery school children in Estonia. The approach made used of CBR, genetic algorithm, and numerical simulation to achieve efficient and thorough risk management and planning. CBR approach had also been implemented in the study by Mendes et al. (2003), which aimed to produce offshore well design that takes into account the risk of the design. Balducelli and D'Esposito (2000) developed an AI approach to facilitate risk management and planning of fire emergencies. It can be observed that application of CBR concepts to hazard identification and risk assessment is feasible, but the current AI tools do not appear to be suitable for the construction industry.

Consequently, this paper presents the two main components of a CBR approach to construction hazard identification that would be able to provide the feedback mechanism proposed by Chua and Goh (2004). The two components are: (1) a detailed knowledge representation scheme, developed based on the modified loss causation model (MLCM), to codify incident cases and past hazard identifications and (2) an intelligent retrieval mechanism that can automatically retrieve relevant past cases. This entails the establishment of semantic networks of case attributes to provide the mechanism for determining the similarity between cases. A case study was also presented to illustrate and validate the proposed methodology which has been implemented in the prototype safety knowledge management system (SKMS).

CBR Framework

CBR has its root in psychological theory of human reasoning, which has the intuitive paradigm that humans solve new problems by recalling past experiences (Mount and Liao 2001). The framework for the proposed CBR system for SKMS is shown in Fig. 1, following the usual three key processes of: (1) case representation and indexing; (2) retrieval of cases; and (3) case utilization and adaptation (Kolodner 1993; Chua et al. 2001). Case representation and indexing is the process of codifying the lessons that a case teaches and the context in which the case can teach its lessons. The sources of safety knowledge comprise the incident cases and hazard identifications of past projects, which are natu-

[1]Senior Lecturer, School of Public Health, Curtin Univ. of Technology, GPO Box U1987, Perth, Western Australia 6845, Australia (corresponding author). E-mail: y.goh@curtin.edu.au

[2]Associate Professor, Dept. of Civil Engineering, National Univ. of Singapore, Singapore 119260, Singapore. E-mail: cvedavid@nus.edu.sg

Note. This manuscript was submitted on July 1, 2008; approved on May 9, 2009; published online on May 13, 2009. Discussion period open until April 1, 2010; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Construction Engineering and Management***, Vol. 135, No. 11, November 1, 2009. ©ASCE, ISSN 0733-9364/2009/11-1181–1189/$25.00.

J. Constr. Eng. Manage., 2009, 135(11): 1181-1189

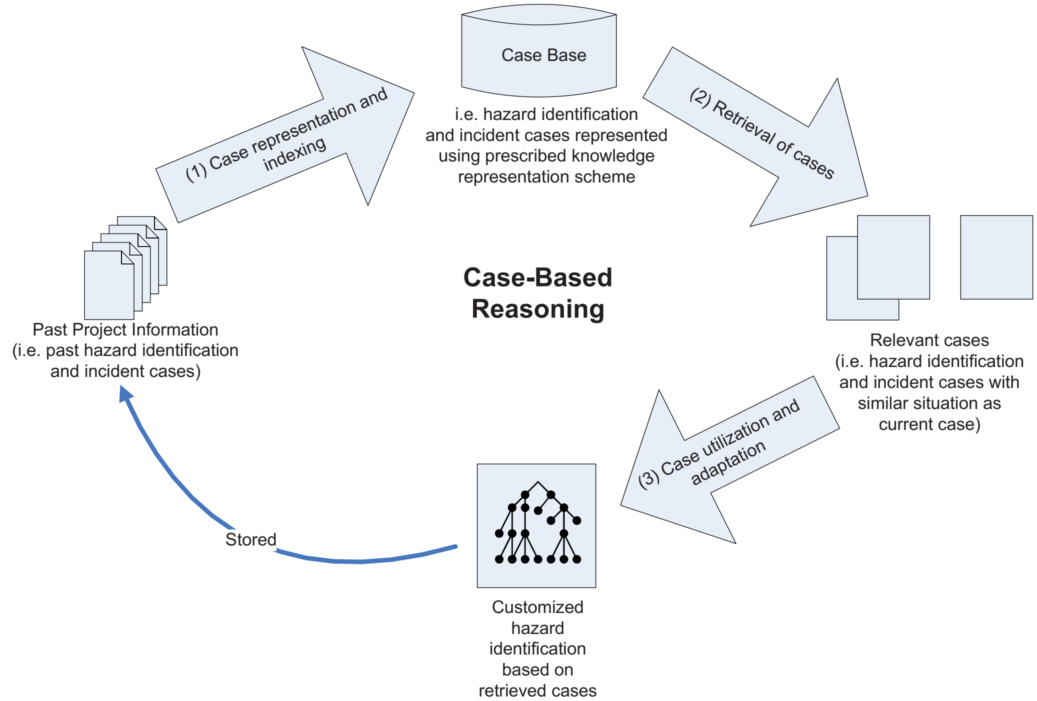

Fig. 1. CBR process

rally episodic and found in the form of cases. Each case contains information on the hazards and events related to a specific work situation. The cases that had been represented in the appropriate knowledge representation scheme will be stored in the case base.

The retrieval of cases is the process of searching and determining relevance or similarity of past cases in the case base with the current situation. CBR utilizes an intelligent and fuzzy methodology so that past cases, such as relevant hazard identification and incident cases in the case base, can be retrieved even if the old situation is not exactly the same as the current situation. This capability is important because most construction work situations are unique and retrieval based on exact matching will often result in relevant knowledge being overlooked. This approach differs significantly from the classical database management system.

Case utilization and adaptation refers to the process of making changes to the retrieved cases to suit the new situation and harnessing the retrieved knowledge to meet the purpose of the users [see Leake et al. (1997) and Kolodner (1993) for discussions on CBR adaptation]. While this is an important aspect of a CBR, the focus of this paper is on the former two elements of the proposed SKMS. Case utilization and adaptation will be discussed in a future paper.

Proposed SKMS Knowledge Representation of Incident Cases and Hazard Identification Trees

One of the basic aims of safety planning is to identify possible incident sequences that can occur (hazard identification). Another key aim is to assess the risk posed by these incident sequences (risk analysis) so that adequate measures can be put in place to reduce the level of risk to a tolerable level (British Standards Institute 2007). As noted earlier, this paper is focused on hazard identification. The MLCM (Chua and Goh 2004) provides the framework for structuring the past incidents. The incident sequence comprises the consequence caused by a contact event which resulted from a breakdown event. It is complemented with the identification of the immediate causes, safety management system failures, and underlying factors. The incident sequence can be inverted to become a hazard identification tree, which is made up of all possible incident sequences identified by the hazard identification team. A set of situational variables describes the context for the incident case or the hazard identification tree, and forms the root node.

In SKMS, a case could be either an incident case or a hazard identification tree. Each case contains two broad types of knowledge: (1) the lessons that it teaches and (2) the context in which it can teach those lessons (Kolodner 1993). The "lessons" that each case teaches can consist of the incident sequences and the risk levels (probability $\times$ severity) posed by them, whereas the context is described by the situational variables. Both types of knowledge have to be represented carefully to ensure appropriate retrieval, adaptation and application of past cases. Since this paper is focused on hazard identification, representation and retrieval of risk levels will not be discussed.

Modeling Approach for the Lessons Learned

In a number of CBR system that handles real-world problems, cases are made up of several interdependent sublessons or subcases. Kolodner (1993) identified two possible approaches in modeling and using these subcases, the snippets approach and the monolithic case approach. Even though the two approaches appear to be different, both agree that each "large" case can be separated into snippets (Thomasson et al. 2006). In the context of this paper, a "large" case would naturally refer to a hazard identification tree or an incident case. Within each "large" case, an incident event (breakdown event, intermediate event, contact event, or consequence) would be a suitable snippet because they form the most basic knowledge blocks of any incident sequence. An incident case contains only one incident sequence, i.e., one breakdown event, one or more intermediate event (if present), one

J. Constr. Eng. Manage., 2009, 135(11): 1181-1189

contact event and their consequences, while a hazard identification tree contains a set of incident sequences forming a tree structure (see Chua and Goh 2004).

The monolithic approach was implemented here because it was recognized that case representation is at best an abstraction of a real-world episode, and it would be more prudent to keep a "large case" intact rather than separating into independent subcases or snippets. In this way, subtle details within a complete case which could be missed if the case has been separated into snippets will be made available to the human user. Since each incident and hazard identification tree contains multiple incident events, the monolithic approach reduces the computational cost of retrieval. This is because the number of cases to be searched and assessed increases tremendously when each snippet (incident event) is treated as an individual case. Besides, it is also more natural for hazard identification teams and incident investigators to view each incident or hazard identification as an episode or scenario. However, it is noted that both approaches are viable and the snippets approach could still be implemented.

Modeling Approach for the Context of Lessons Learned

The modeling of the context of lessons learned is also known as the indexing problem. It is tackled at two levels, first, selection of an appropriate indexing vocabulary, and second, the selection of specific indexes for each case (Kolodner 1993). Indexing vocabulary is a set of possible descriptors that can be used to index all the cases in the case base, while indexes are specific descriptors that designate the situations under which the case is relevant. In this paper the indexing vocabulary corresponds to the situational variables of the MLCM (Chua and Goh 2004).

Indexing Vocabulary

In the proposed CBR system, the hazard identification and the overall risk assessment was based on the job hazard analysis (JHA) (also known as job safety analysis) approach. JHA is a common safety planning technique that is focused on a specific job and the analysis begins by separating the job into specific job steps. Each job step is then evaluated for the possible hazards and their risks. Subsequently, relevant risk controls are then selected to eliminate or reduce the risks identified. Since incidents usually occur during a specific job step of an activity, an indexing vocabulary that can describe the situational variables of a job step during JHA can also be used to describe the context of an incident.

The indexing vocabulary describing the situational variables is based on the following linguistic structure:

Action(s) executed on *object(s)-worked-on* using *resource(s)* at *location(s)* with *nearby object(s)* and *nearby action(s)*.

For example, in a lifting operation using a crawler crane, the situational variables could be expressed as: *lifting operation (action)* executed on *precast segment (object-worked-on)* using *crawler crane (resource)* at *unspecified (location)* with *nearby access scaffold (nearby object)* and *general work (nearby action)*.

Each of the italic terms in the above linguistic structure corresponds to potential sources of harm or hazards that can contribute to the occurrence of incident sequences in a job step. "Harm" is usually due to an uncontrolled source of energy or substance being released (Chua and Goh 2004). This release of energy or substance could originate from: (1) human actions applied during some course of work ("action" or "nearby action"); (2) any object or substance that was used to facilitate work, acted upon or spatially close to the human action ("object-worked-on," "resource," or "nearby object"); or (3) the environment or location in which the job was being executed in ("location"), and collectively establish the situational variables or the context of the case.

Indexes

Not all six variables in the index have equal importance in relation to the lessons of the case. For example, in the above lifting example, a cushioning timber for the precast segment got stuck onto the precast segment, and fell off while the precast segment was being lifted, and struck the rigger. The situational variables, *location, nearby objects*, and *nearby action*, are not contributory to the incident but only provide contextual information and a richer picture of the case, thus should be given very low weights. Moreover, the direct influence of the remaining situational variables on the incident varies. The breakdown event, "Cushioning timber fall from height," could only happen when the *action*, lifting, is executed and the cushioning timber is present due to the *object-worked-on*, precast segment. These are the necessary situational variables. Although crane is a resource used in the lifting, it is not directly related to the breakdown event, and thus could carry less weight.

A further point is that there can be more than one value for each type of situational variable. For example, an action that uses two resources or handles two objects-worked-on at the same time will require two resources or object-worked-on in the index.

Case Retrieval

Overview

Retrieval of past cases is one of the most important processes of any CBR system. Its quality directly affects the relevance of retrieved cases and hence the overall quality of the reminding capability of a CBR system. Two main types of retrieval approaches are usually employed: indexing and similarity scoring (or distance-based) (Liao et. al. 1998).

The former approach organizes cases using an indexing structure that is derived through various machine learning methods such as decision tree, neural network, and clustering algorithms. During the retrieval, the system will then traverse the indexing structure and search for the stored cases that match the input case's indexes. However, the need for a relatively large and varied case base, a clear outcome variable, and the need to frequently retrain or redevelop the indexing structure makes the indexing approach unsuitable for the proposed hazard identification approach.

Similarity scoring approaches compute a quantitative distance or similarity score between the input case and each stored case during retrieval. The higher the similarity score the more relevant is the stored case to the input. Subsequently, the top K number of cases (with greater relevance) will be retrieved for utilization or further adaptation. The key advantage of using similarity score is the flexibility of the approach. Furthermore, the approach can even be applied on relatively small case bases.

Similarity scores are determined at two levels: local and global similarity (Empolis Knowledge Management GmbH 2001). Local similarity refers to the similarity between the values of a particular attribute (or situational variable) of two cases. On the other hand, global similarity refers to the similarity between two cases. Local similarity is usually determined using a similarity function,

J. Constr. Eng. Manage., 2009, 135(11): 1181-1189

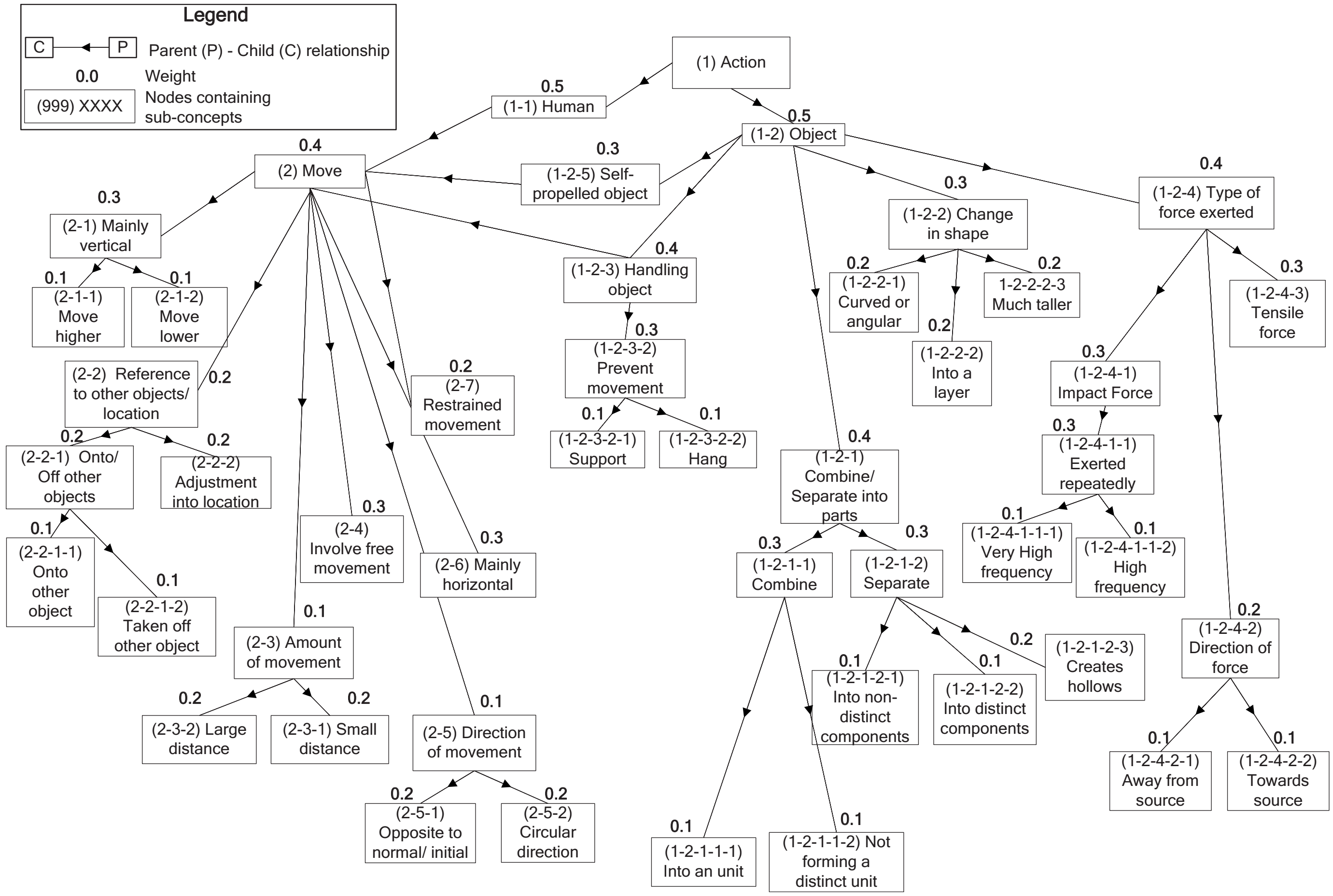

Fig. 2. Semantic network for situational variable "Action"

and global similarity is determined through a weighting function that places different importance on local similarity (LSSs) scores of different attributes.

Proposed Similarity Scoring Approach

Various functions have been adopted for deriving the similarity score. Some examples include taxonomy tree approach (Empolis Knowledge Management GmbH 2001), use of tables or rules (Karim and Adeli 2003; Luu et al. 2003), and the evaluation of subattributes or subconcepts (Domeshek 1991; Kolodner 1993). Another possible approach is to adopt a mix of approaches for different types of attributes, e.g., Zhao et al. (2009) used different five different similarity scoring functions.

An approach that is based on the subconcept approach has been proposed herein. The similarity scoring approach, which will be described in detail later, makes use of a semantic network (Russell and Norvig 1995) to calculate the similarity score between two values of a situational variable. This is in contrast with the taxonomy tree approach as used by Kolodner (1993). A semantic network was constructed for each situational variable. Each semantic network contains subconcepts that describe all the possible values of the situational variable. The basic idea of the proposed similarity scoring approach is that the higher the proportion of common subconcepts two values have, in relation to all the relevant subconcepts of the two values, the more similar the two values are. The similarity score is computed based on an equation that will be explained later. An example of a semantic network for the situational variable "Action" is shown in Fig. 2.

With reference to Fig. 2, a semantic network is made up of nodes (illustrated as boxes with words). A semantic network can be seen to be an extension of a taxonomy tree, where the key difference is that nodes in a semantic network can have more than one parent. For example, the node "(2) Move" of Fig. 2 has two parent nodes, "(1-1) Human" and "(1-2-5) Self-propelled object." The semantic network allows for greater flexibility in classifying the concepts or subconcepts associated with the situational variable, and would not have the problem of a taxonomy tree where each value can only have one parent node.

The semantic networks had been constructed based on the procedure depicted in the flowchart of Fig. 3. A case base was first constructed based on more than 100 incident cases taken from the Safety Information System (SITS) of Singapore's Land Transport Authority (LTA) and 10 hazard identification trees obtained from the Mine Safety and Health Administration (2004), LTA and various construction contractors. The incident and hazard identification tree case base was then queried and all the possible values for each situational variable were collated. Each of the values was then given a working definition that highlights the key characteristics of the value that distinguished it from the other possible values, and at the same time emphasized the usual types of hazardous object, energy or harmful substance associated with the value. For example, the value, "Lift" of the situation variable

J. Constr. Eng. Manage., 2009, 135(11): 1181-1189

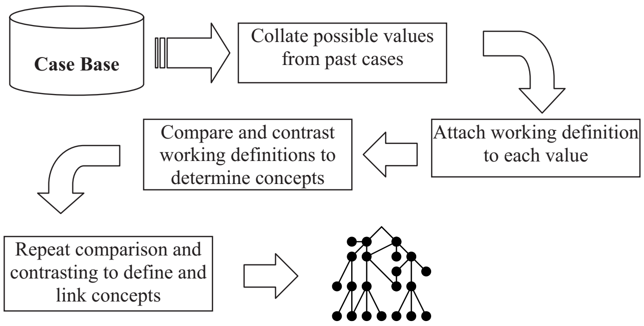

Fig. 3. Flowchart for construction of taxonomy tree

"Action" was defined as, "to move an object from a lower to a higher position, hence accumulating gravitational potential energy and producing kinetic energy during the movement." This definition highlighted the nature of the action, and at the same time made reference to the type of energy or harmful substance that was produced or accumulated during the action.

Based on the working definitions, the values were then compared and contrasted to identify similar and contrasting subconcepts that the values represented. These subconcepts were then repeatedly evaluated to identify more specific concepts. Related subconcepts were then linked together by an arrow, where the arrow points from the more general subconcept to the more specific subconcept. As mentioned earlier, in a semantic network a lower level subconcept can be linked to more than one parent. However, such multiple linkages can cause the semantic network to be complex and incomprehensible, and was avoided whenever possible.

The semantic networks developed represent all the possible subconcepts for each situational variable. Each value can then be represented by a list of subconcepts picked off from the corresponding semantic networks. For example, Fig. 4 shows the list of subconcepts for the values "Hack," "Excavate," and "Extract" of the situational variable "Action." Any new subconcepts can be easily added into the network following the same procedure in outlined in Fig. 3.

In order to differentiate the subconcepts in terms of specificity and importance, a weight was assigned to each of the nodes (see Fig. 4). These weights were assigned based on the guiding principle that higher nodes, or nodes that are closer to the root, are more influential on the categorization of values, and thus given a higher weight. Subconcepts that are more directly related to potential hazards are also given higher weights. Thus these weights can be viewed as the incremental similarity due to a match on the subconcept represented by the node.

The LSSs between values V_1 and V_2 can be calculated based on the following equation:

$$\mathrm{LSS}(V_1,V_2)=\sum w_{ci}\Big/\left(\sum w_{ci}+\sum w_{dj}\right) \quad (1)$$

where $i=1,2,\ldots common$, $j=1,2,\ldots different$, *common* is the number of common subconcepts to both V_1 or V_2, *different* the number of subconcepts that belongs only to either V_1 or V_2; w_{ci}=weight of the common subconcept i; and w_{dj} the weight of

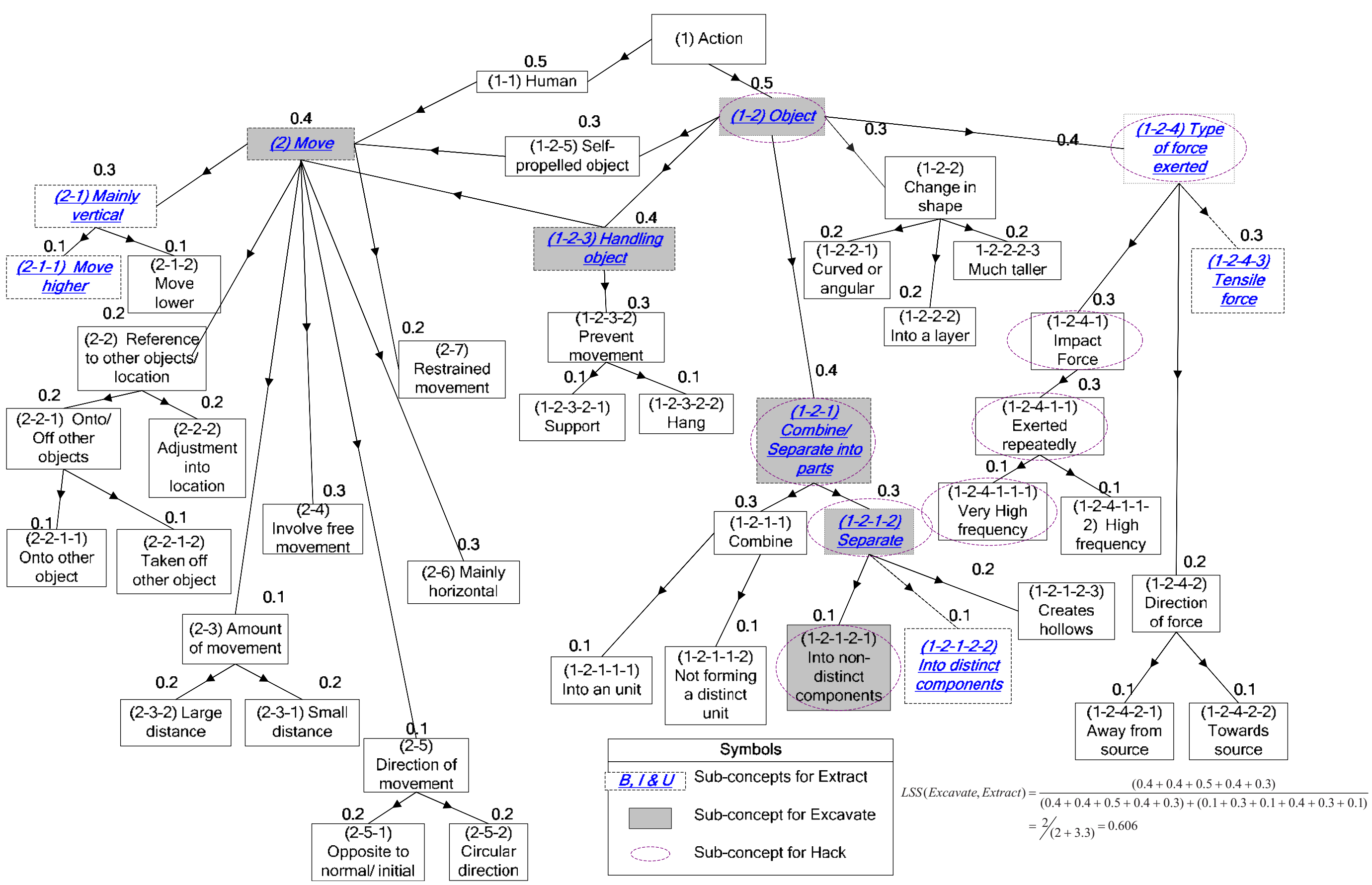

Fig. 4. Subconcepts for the values "Hack," "Extract," and "Excavate" under situational variable "Action"

J. Constr. Eng. Manage., 2009, 135(11): 1181-1189

the subconcept j that belongs only to either V_1 or V_2. Essentially, LSS measures the proportion of weights represented by the common subconcepts to the weights of the union of the subconcepts, so that the higher the proportion, the greater will be the similarity of V_1 and V_2. A value of 0 will mean that the subconcepts are clearly distinct, while a value of 1 indicates that the subconcepts are exactly identical.

The application of Eq. (1) is demonstrated in Fig. 4, wherein the LSS for "Excavate" and "Extract" were determined to be 0.606 and shown in Fig. 4. Likewise, the LSS for "Excavate" and "Hack" can be computed to be 0.406.

The comparison in the example shows that "Excavate" is more similar to "Extract" than to "Hack." This appears logical because both excavation and extraction usually involves moving objects for the purpose of separating them from other objects. Although hacking also has the purpose of separating objects, it employs a (very) high frequency impact force in the process. Consequently, the hazards that hacking pose is generally different from those posed by excavation and extraction. Following this comparison, the similarities between other pairs of activities are computed and verified to be reasonable.

GSS

The global similarity score (GSS) is the overall similarity score between two cases. It is computed based on a weighted sum of the LSS of all the attribute-value pairs of cases C_1 and C_2 being compared, so that

$$\mathrm{GSS}(C_1,C_2) = \sum (w_i \times \mathrm{LSS}_i)/\left(\sum w_i\right) \quad i=1,2,\ldots,n \qquad (2)$$

where w_i=corresponding weight of attribute or situational variable i, which as explained earlier is to indicate the relative importance of the attribute to the incident. LSS_i denotes the LSS of attribute i of cases C_1 and C_2 obtained from Eq. (1), and n the number of attributes (or six in the present SKMS).

The importance rating for each of the situational variables is based on a Likert 5-point-scale, where "1" denotes that the attribute-value pair is of low importance to the case, such as situational variables that provide only contextual information, while a "5," on the other end of the scale, denotes that the case's occurrence is highly dependent on that situational variable.

Incident cases with GSS greater than the user-specified GSS threshold, $\mathrm{SS_{GT}}$, will be retrieved. $\mathrm{SS_{GT}}$ can be increased to achieve higher relevance of retrieved cases or lowered to increase the number of retrieved cases for utilization. At the time of writing, available CBR shells were not able to facilitate the proposed approach. The prototype was developed using Visual Basic for Applications to create the necessary functions for LSS and GSS calculations based on the semantic networks developed.

Case Study

Case Base of Safety Knowledge

The case study described herein is to validate the key concepts presented, especially to demonstrate how a set of past safety knowledge, i.e., a most similar hazard identification and relevant incident cases, can be retrieved based on the proposed knowledge structure and similarity scoring functions. The case base used in the case study comprises two types of cases, namely incident cases and hazard identification trees. The incident cases were obtained from the SITS of the LTA of Singapore. To ensure the integrity of the cases, the details of the incident cases were verified through interviews with relevant site personnel and review of appropriate site documents.

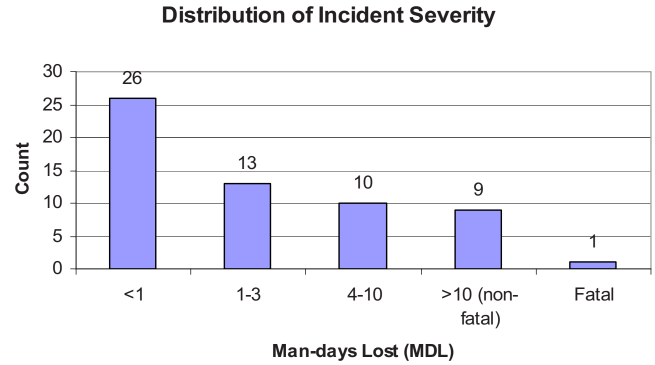

Fig. 5. Distribution of incident severity in terms of MDL

The incident cases belong to a rapid rail project involving the construction of precast viaducts and above ground train stations. The contract had a total of 59 reported incidents after 1,444,8300 mh of work (project duration of 3 years and 9 months). Fig. 5 shows the distribution of the incident cases in terms of severity measured by number of man-days lost (MDL). Most of the cases (66.1%) were less than 3 MDL and were not required to be reported to the Ministry of Manpower (MOM) in Singapore, but were recorded in SITS. There were 20 cases (33.9%) which are required by law to be reported to the MOM. Among these there was a case with fatality, where one worker was killed. The incident cases occurred in a various types of activity, such as soil boring, hoisting, concreting and manual handling work. This variation forms a rich source of knowledge.

The case base also contains 10 hazard identification trees as shown in Table 1. These hazard identification trees are developed from safety documents obtained from numerous sources, including the main contractors' safety management systems, tender documents submitted to LTA, and training materials of Mine Safety and Health Administration (2004). Two experienced construction safety practitioners with at least 8 years in construction safety each were asked to verify the content of the cases and assign appropriate weights to assist in the assessment of the similarities between cases.

Table 1. Case Titles of Hazard Identification Trees in Case Base

Number	Case title
1	Gas-cutting in confined space (tank)
2	Rigging up precast element
3	Lift precast wall using crawler crane
4	Arc welding of suspended pipes in trench
5	Concreting work using bucket
6	Lowering pipe into trench using excavator
7	Loading truck with soil using excavator
8	Concrete breaking
9	Frame scaffold erection
10	Gas-cutting of H-pile

J. Constr. Eng. Manage., 2009, 135(11): 1181-1189

Table 2. GSSs of All Hazard Identification Trees in the Case Base

Number	Case title	GSS
6	Lowering pipe into trench using excavator	0.61
3	Lift precast wall using crawler crane	0.59
7	Loading truck with soil using excavator	0.51
5	Concreting work using bucket	0.39
2	Rigging up precast element	0.27
4	Arc welding of suspended pipes in trench	0.24
8	Concrete breaking	0.15
9	Frame scaffold erection	0.10
10	Gas-cutting of H-pile	0.10
1	Gas-cutting in confined space (tank)	0.05

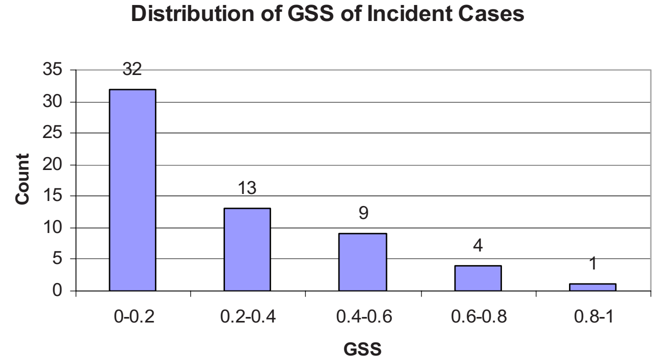

Fig. 6. Distribution of GSS of incident cases

Case Retrieval

The job scenario that was being assessed in the case study was a material delivery activity where a lorry crane was used to unload a bundle of timber strips. This is a common activity for most construction worksites. The situational variables was based on the following description of the job step, "*unloading* of *timber strip (bundle)* using *chain sling* and *lorry crane* at *site entrance* with nearby *plants/vehicles* (and *no nearby actions*)."

Table 2 shows the GSS between the input case and each of the 10 hazard identification trees stored in the case base. The GSS were calculated with Eq. (2), which requires the LSS of different situational variables to be determined using Eq. (1). The calculation of LSS had been described in the earlier sections and Fig. 4. Intuitively from the descriptions of the case, the hazard identification trees with GSS $>$ 0.5 is more similar to the input case as compared to hazard identification trees with GSS $<$ 0.5. Table 3 shows a detailed breakdown of the LSS between the input case and the most similar (i.e., retrieved) hazard identification tree (i.e., lowering pipe into trench using excavator). The calculation of the GSS for the most similar case (Case Number 6 in Table 2) had been included at the bottom of Table 3. Upon inspection, it can be observed that the LSS for some of the situational variables were very similar, i.e., action, nearby object and one of the resources used (Excavator and Lorry Crane) had LSS of 0.73, 1.00, and 0.80, respectively. The high GSS was contributed by the high weights assigned by the hazard identification team for the situational variables with high LSS ("5" denoting "very important," and "1" denoting "very unimportant" at the other end of the scale, with 3 denoting "neutral").

When the input case has more than one value for a particular situational variable such as "Chain sling" and "Lorry crane" for "Resource" in the case study, the highest LSS value will be used as representative. Thus, when the LSS for the index "Resource = Excavator" of the stored case was computed, it was compared with each of the two Resource indexes of the input case, "Chain Sling" and "Lorry crane" in turn with the corresponding LSS 0.09 and 0.80, respectively. Therefore, the index with higher LSS, i.e., "Lorry crane," matched against the stored case's "Excavator." Similarly, for the other Resource index, "Lifting gear" in the stored case, "Chain sling" with an LSS of 0.6 is used in the similarity matching.

Fig. 6 shows the distribution of the GSS of all the incident cases in the case base. The majority of the stored cases, or 32 (54%) cases, had GSS between 0 and 0.2. Another 22 cases (37%) had GSS between 0.2 and 0.6. Only 5 cases or 8.5% had GSS $\geq$ 0.6. Table 4 shows the GSS and LSS of the incident cases with GSS $\geq$ 0.6. The GSS were calculated using Eq. (2). For instance, GSS of incident case number 230 has a value of 0.78; it was determined by dividing the sum of the weighted LSS (8.53) by the sum of the weights (11). Note that the cases without a complete set of situational variables can still be retrieved. Case 1153 and Case 141 occurred during lifting work using mobile plants, Case 230 occurred during unloading of material, and Case 160 and Case 292 were related to mobile plants that were used during construction activities. These cases were then returned to the case base for adaptation which is beyond the scope of the present paper.

Discussions

Generally, the case study had demonstrated that the proposed knowledge representation scheme and similarity functions are able to facilitate retrieval of relevant safety knowledge. The proposed retrieval mechanism utilizes the knowledge stored in semantic networks and importance weights captured in past (or stored) cases to measure their similarity to the input case. The

Table 3. LSSs of Retrieved Hazard Identification Tree

Situational variables	Value: Stored case	Value: Input case	LSS	Weights	Weighted LSS
Action	Lower	Unload	0.73	5	3.64
Location	Trench	Site entrance	0.00	2	0.00
Nearby object	Plants/vehicles	Plants/vehicles	1.00	3	3.00
Object-worked-on	Pipe	Timber strip (bundle)	0.11	2	0.21
Resource	Lifting gear	Chain sling	0.60	2	1.20
Resource	Excavator	Lorry crane	0.80	3	2.40
			Total:	17	10.45
			GSS	=10.45/17	=0.61

J. Constr. Eng. Manage., 2009, 135(11): 1181-1189

Table 4. Local Similarity Scores of Retrieved Incident Cases

Number	GSS	Case title	Situational variables	Values: Stored case	Values: Input case	LSS	Weights	Weighted LSS
160	0.86	Worker injured by forklift	Resource	Forklift	Lorry crane	0.86	5	4.32
230	0.78	Finger trapped between I-beams during unloading	Action	Unload	Unload	1.00	5	5.00
			Object-worked-on	I-beams	Timber strip (bundle)	0.18	3	0.53
			Resource	Lorry crane	Lorry crane	1.00	3	3.00
292	0.77	Worker injured while coming down from lorry	Resource	Lorry	Lorry crane	0.77	5	3.86
1,153	0.74	Parapet wall fall from height while lifting	Action	Lift	Unload	0.73	4	2.91
			Object-worked-on	Precast parapet wall	Timber strip (bundle)	0.07	1	0.07
			Resource	Crane	Lorry crane	0.64	3	1.91
			Resource	Chain sling	Chain sling	1.00	4	4.00
141	0.67	Worker injured by I-beam during lifting	Action	Lift	Unload	0.73	4	2.91
			Object-worked-on	I-beam	Timber strip (bundle)	0.20	2	0.40
			Resource	Mobile crane	Lorry crane	0.91	3	2.73

LSS and GSS provide specific information on the rationale why a stored case has been retrieved as relevant. The relevance of the case can be assessed and adjustments may be made, if necessary.

The similarity scoring approach allows similar (but not exactly matching) cases to be retrieved and used. It is evident from Table 4 that all the retrieved incident cases do not have exactly matching indexes as the input case ($GSS < 1$). If an exact matching approach is used as in traditional databases, these similar cases would not have been retrieved and important information on hazards and risk associated with these cases would be lost to the new case. The threshold similarity score can be adjusted to achieve a balance between relevance and number of retrieved cases. Less number of cases will be retrieved when the threshold is increased although the relevance can be increased.

It is noted that the number of cases does not affect the essence of the case study because the key purpose is to validate the feasibility of the proposed concepts and methodologies. Furthermore, all CBR system are learning systems that accumulate knowledge as more experiences are being accumulated and are capable of fully using available knowledge.

Even though the proposed CBR system can potentially improve the efficiency and quality of hazard identification, the system can only facilitate the hazard identification process. The users will have to assess the usefulness of the cases that are retrieved, guided by the similarity scores calculated by the system. Furthermore, the approach can only be implemented in organizations with a well-established system to capture incident cases and hazard identification. Without the cases, the proposed CBR approach cannot work. As in any other data repository, the quality of the case base has to be maintained through periodic data validation and data audit. Despite the fact that organizations implementing the approach will incur the initial cost of developing semantic networks and the on-going cost of maintaining the case base, it is believed that the costs will be similar to that of a comprehensive accident database.

Conclusions

This paper discusses two of the key components of a CBR approach to construction hazard identification: the knowledge representation scheme and case retrieval mechanism. The knowledge representation was structured based on two types of knowledge, the lessons learned and the context of the lessons learned or indexes. The lessons learned were modeled as a "large case" with the snippets of the case being linked to the indexes to allow for adaptation after retrieval. The indexes correspond to the situational variables described in the MLCM described in Chua and Goh (2004). The proposed knowledge representation scheme was implemented in a relational database.

Case retrieval consists of two main parts, the determination of the LSS and the computation of the GSS based on the LSS. The LSS is calculated using a weighted subattributes approach that is dependent on a series of semantic networks. The LSS are then combined through a weighting function to compute the GSS. Unlike the convention of having end users assign the weights for the different attributes for the input case, these weights are assigned by the investigators or hazard identification team during input of the stored case.

The scope of the proposed SKMS is presently limited to the identification of possible incident sequences, nevertheless, the same principles can be employed to extend the CBR system to include the risk analysis and risk control elements as well.

References

Balducelli, C., and D'Esposito, C. (2000). "Genetic agents in an EDSS system to optimize resources management and risk object evacuation." *Safety Sci.*, 35(1–3), 59–73.

British Standards Institute. (2007). *OHSAS 18001—Occupational health and safety management systems—Requirements*, British Standards Institute, London.

Chua, D. K. H., and Goh, Y. M. (2004). "An incident causation model for improving feedback of safety knowledge." *J. Constr. Eng. Manage.*, 130(4), 542–551.

Chua, D. K. H., Li, D. Z., and Chan, W. T. (2001). "Case-based reasoning approach in bid decision making." *J. Constr. Eng. Manage.*, 127(1), 35–45.

Domeshek, E. (1991). "Indexing stories as social advice." *Proc., AAAI-91*, AAAI Press/MIT Press, Cambridge, Mass.

J. Constr. Eng. Manage., 2009, 135(11): 1181-1189

Empolis Knowledge Management GmbH. (2001). *cbr:works 4*, Compendium, ⟨http://www.empolis.com⟩.

Karim, A., and Adeli, H. (2003). "CBR model for freeway work zone traffic management." *J. Transp. Eng.*, 129(2), 134–145.

Kolodner, J. (1993). *Case-based reasoning*, Morgan Kaufmann Publisher, San Mateo, Calif.

Leake, D. B., Kinley, A., and Wilson, D. (1997). "Case-based CBR: Capturing and reusing reasoning about case adaptation." *Int. J. Expert Systems*, 10(2), 197–213.

Liao, T. W., Zhang, Z. M., and Mount, C. R. (1998). "Similarity measures for retrieval in case-based reasoning systems." *Appl. Artif. Intell.*, 12(4), 267–288.

Luu, D. T., Ng, S. T., and Chena, S. E. (2003). "A case-based procurement advisory system for construction." *Adv. Eng. Soft.*, 34, 429–438.

Mendes, J. R. P., Morooka, C. K., and Guilherme, I. R. (2003). "Case-based reasoning in offshore well design." *J. Pet. Sci. Eng.*, 40(1–2), 47–60.

Mine Safety and Health Administration. (2004). U.S. Dept. of Labor, Mine Safety and Health Administration official website, ⟨http://www.msha.gov⟩ (May 2004).

Ministry of Manpower (MOM). (2004). "Opening address by Ms. Yong Ying-I, Permanent Secretary, Ministry of Manpower at the CEO workshop on the Occupational Safety and Health for the Construction Industry." Ministry of Manpower official website, ⟨http://www.mom.gov⟩ (Sept. 2007).

Mount, C., and Liao, T. W. (2001). "Prototype of an intelligent failure analysis system." *ICCBR 2001, LNAI 2080*, D. W. Aha and I. Watson, eds., Springer, Heidelberg, Berlin.

Reese, C. D., and Edison, J. V. (2006). *Handbook of OSHA construction safety and health*, Taylor and Francis, Bristol, Pa.

Remm, M., and Remm, K. (2008). "Case-based estimation of the risk of enterobiasis." *Artif. Intell. Med.*, 43(3), 167–177.

Russell, S., and Norvig, P. (1995). *Artificial intelligence: A modern approach*, Prentice-Hall, Englewood Cliffs, N.J.

Thomasson, B. J., Ratcliffe, M. B., and Thomas, L. A. (2006). "Improving the tutoring of software design using case-based reasoning." *Adv. Eng. Inf.*, 20(4), 351–362.

Zhao, J. S., Cui, L., Zhao, L. H., Qiu, T., and Chen, B. Z. (2009). "Learning HAZOP expert system by case-based reasoning and ontology." *Comput. Chem. Eng.*, 33(1), 371–378.

J. Constr. Eng. Manage., 2009, 135(11): 1181-1189